

Genetic diversity analysis of the *Amaranthus* genus using genomic tools

BY

ERIKA VILJOEN

Submitted in partial fulfillment of the requirements for the degree

PHILOSOPHIAE DOCTOR

Department of Plant and Soil Sciences

In the Faculty of Natural and Agricultural Sciences

University of Pretoria

Pretoria

June 2018

SUPERVISOR: PROF. D. K. BERGER

Declaration

I, Erika Viljoen, declare that the thesis/dissertation, which I hereby submit for the degree *Philosophiae Doctor* at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Signature:

Date:

Table of Contents

Summary.....	7
Opsomming	9
Acknowledgements	10
List of Tables.....	11
List of Figures.....	13
Abbreviations	16
Preface	21
Hypothesis	24
Research Questions	24
Objectives.....	25
Structure of Dissertation.....	26
Publications	27
References	29
CHAPTER 1.....	30
LITERATURE REVIEW	30
Introduction	31
Genus <i>Amaranthus</i> (L.)	33
Weedy Amaranth.....	35
<i>Amaranthus viridis</i> (2n=34)	36
<i>Amaranthus spinosus</i> (2n=34).....	37
Grain Amaranths	38
<i>Amaranthus caudatus</i> (2n=32)	39
<i>Amaranthus cruentus</i> (2n=32).....	40
<i>Amaranthus hypochondriacus</i> (2n=32)	41
Vegetable amaranths	42
<i>Amaranthus dubius</i> (2n=64).....	43
<i>Amaranthus blitum</i> (synonym <i>Amaranthus lividus</i>) (2n=34).....	44
<i>Amaranthus tricolor</i> (2n=34).....	45
<i>Amaranthus graecizans</i> (2n=34)	46
<i>Amaranthus hybridus</i> (2n=32/34).....	47
Economic Importance.....	48
Tolerance to climatic conditions.....	48
Phytoremediation.....	49
Role in biofuels.....	50
Nutritive value.....	51

Antinutritional factors	56
<i>Amaranthus</i> Breeding	57
References	60
CHAPTER 2.....	68
REVIEW ON CROP GENETICS AND IMPROVEMENT WITH EMPHASIS ON ORPHAN CROPS FOR SUSTAINABLE GLOBAL FOOD SECURITY	68
Abstract	69
Introduction	70
Role of traditional crops and their wild relatives.....	71
Important factors in orphan crops research	74
Traditional plant breeding	75
Genomics-assisted plant breeding	76
Transgenic Technologies.....	76
Genomic Resources and Trait Mapping	80
Molecular markers.....	80
Next generation sequencing.....	81
Trait Mapping.....	82
Molecular breeding in orphan crops.....	84
<i>Amaranthus</i>	84
Pigeon pea	86
Tef.....	87
Sweet Potato	88
Conclusion.....	89
References	90
CHAPTER 3.....	101
A DRAFT GENOME AND TRANSCRIPTOME ASSEMBLY OF THE ECONOMICALLY IMPORTANT ORPHAN CROP <i>AMARANTHUS TRICOLOR</i>.....	101
Abstract	102
Introduction	103
Materials and Methods	107
Plant growth.....	107
Extraction of DNA	108
DNA Library Preparation and Sequencing.....	108
Extraction of RNA.....	109
RNA Library Preparation and Sequencing.....	110
Read Quality Control.....	111
<i>De novo</i> Genome Assembly	111
<i>De novo</i> Transcriptome Assembly	112

Genome Annotation.....	112
Results and Discussion.....	114
<i>De novo</i> Genome Assembly	114
<i>De novo</i> Transcriptome Assembly	118
Genome Annotation.....	120
Conclusion.....	128
References	129
CHAPTER 4.....	133
SPECIES RELATIONSHIPS WITHIN THE AMARANTHUS GENUS REVEALED BY ILLUMINA MISEQ SEQUENCING OF MICROSATELLITE MARKERS	133
Abstract	134
Introduction	135
Materials and Methods	139
<i>Amaranthus</i> accessions and growth.....	139
Identify microsatellite loci from draft genome assembly	141
Marker validation	141
Marker screening via ABI capillary gel electrophoresis technology.....	142
Marker screening via MiSeq NGS technology	143
Genetic diversity of ABI capillary electrophoresis compared to MiSeq NGS technologies	145
Species relationships of unknown SAG genotypes revealed by MiSeq analysis of SSR loci	146
Results	147
Identification of microsatellite loci from a draft genome assembly of <i>Amaranthus tricolor</i>	147
Microsatellite validation	148
Marker screening via ABI capillary gel electrophoresis technology.....	151
Marker screening via NGS Illumina MiSeq technology	154
Genetic diversity of <i>Amaranthus</i> as revealed by ABI capillary electrophoresis compared to MiSeq NGS technologies.....	159
Genetic diversity of the genus <i>Amaranthus</i> revealed by MiSeq analysis of polymorphic SSR loci	161
Discussion.....	169
Conclusion.....	173
References	175
Supplementary Data	179
CHAPTER 5.....	186
APPLICATION OF CHLOROPLAST PHYLOGENOMICS TO RESOLVE SPECIES RELATIONSHIPS WITHIN THE PLANT GENUS AMARANTHUS	186
Abstract	187
Introduction	188
Methods and Materials	192

Growth and maintenance of germplasm accessions	192
Genomic DNA isolation	192
Illumina library preparation and sequencing	194
Assembly of the <i>Amaranthus tricolor</i> chloroplast genome sequence	194
<i>Amaranthus tricolor</i> chloroplast genome annotation	195
Mapping of the additional <i>Amaranthus</i> germplasm sets	195
Angiosperm whole chloroplast genome phylogeny	196
<i>Amaranthus</i> barcoding and phylogenetic analysis	196
<i>Amaranthus</i> whole chloroplast genome phylogenetic analysis	197
Identification of an alternative barcoding region for the <i>Amaranthus</i> genus	198
Results	200
Assembly and annotation of <i>A. tricolor</i> chloroplast genome sequence.....	200
Angiosperm whole plastid genome phylogeny.....	203
<i>Amaranthus</i> phylogeny using barcoding genes	205
<i>Amaranthus</i> whole chloroplast genome phylogenetic analysis	207
Investigation of genes to complement existing barcoding genes for phylogenetic analysis of <i>Amaranthus</i> species.....	214
Discussion.....	216
The <i>Amaranthus tricolor</i> chloroplast assembly.....	216
Resolving the position of <i>A. tricolor</i> within the Angiosperm phylogeny.....	217
Whole chloroplast analysis provides better phylogenetic resolution within <i>Amaranthus</i> than DNA barcoding	218
Implication of the low phylogenetic resolution to <i>Amaranthus</i> phylogeny.....	220
Conclusion.....	224
References	228
Supplementary Data	234
CHAPTER 6.....	243
CONCLUSIONS AND FUTURE WORK.....	243
References	251
APPENDIX A	253
MAND Data Script.....	253

Summary

The benefits of traditional leafy vegetables and global concerns about biodiversity loss have led researchers to renew their interest in nutritious underutilized crops. *Amaranthus* species are highly promising nutritious leafy vegetable food sources in Africa and have appreciable levels of essential micronutrients. Uninformative morphological characters and poorly resolved marker dendrograms have led to interest in alternative forms of molecular identifications in the *Amaranthus* genus. A draft genome sequence was assembled for a leafy amaranth accession, *A. tricolor*. Approximately 1.5 billion sequence reads were generated using the Illumina HiScan and MiSeq high-throughput sequencing platforms, composed of both paired end and long mate pair libraries. The *de novo* assembled genome had 44x coverage of the estimated 402 Mb genome size, had an N50 scaffold size of 51 Kb and a 54% repetitive sequence fraction. Additionally, the RNA-Seq transcriptomic analysis of leaf, shoot and root material was used to assemble a draft transcriptome (125 350 contigs, contig N50 of 1 036 bp) and subsequently used as evidence to infer gene models on the genomic scaffolds. Inadequate phylogenies, which revealed poor inter and intra species resolution, were obtained for a diverse set of *Amaranthus* species (14 unknown and 45 previously identified accessions) using chloroplast barcoding genes (*MatK*, *rbcL* and ITS). However, using the sequence data from the draft genome assembly, it was possible to extract, assemble and annotate a full-length chloroplast sequence for *A. tricolor* with approximately 200x coverage. Furthermore, shallow sequencing of the remaining *Amaranthus* species enabled the reference guided assembly of all the chloroplasts. The phylogenomic analysis of the *Amaranthus* species revealed classifications of unknown accessions and shed light on the possible domestication events of modern day cultivated grain amaranth. *Amaranthus tricolor* grouped together with additional leafy amaranth, *A. blitum*, *A. viridis* and *A. graecizans*, representing the *Amaranthus* subgenus *Albersia*. The genome also revealed 671 microsatellites with the potential as SSR markers and further validation revealed six robust SSR markers for inter- and intraspecies genotyping. Allele scores obtained from both gel based capillary electrophoresis (ABI) and next generation sequencing (MiSeq) technologies were highly comparable. Downstream analysis of SSR markers screened with NGS revealed informative genetic relationships between known and unknown *Amaranthus* accessions. The draft genome sequence proved to be extremely useful in generating molecular tools for this important plant genus.

In combination, these genomic tools can be further utilized in future to assist in selecting and exploiting genotypes for breeding programs and contribute to global food security.

Opsomming

Globale bekommernisse oor biodiversiteitsverlies en verborge voordele van blaargroentes het navorsers gelei om hul belangstelling in die ontwikkeling en instandhouding van voedsame, onderbenutte gewasse te vernuwe. *Amaranthus*-spesies is hoogs belowende blaargroente-tipes in veral Afrika, met aansienlike hoë vlakke van noodsaaklike mikronutriënte. 'n Konsepgenoom-volgorde is saamgestel vir 'n blaaragtige amarant spesie, *Amaranthus tricolor*. Ongeveer 1.5 biljoen volgorde fragmente is gegenereer deur die Illumina HiScan- en MiSeq-platforms, deur kortmaat-paar en langmaat-paar volgordes te gebruik. Die *de novo* saamgestelde genoom het 44x dekking van die geraamde 402 Mb genoomgrootte (N50 van 51 Kb) geopenbaar en besit 'n 54% herhalende sekwensie-fraksie. Daarbenewens is die RNA-Seq transkriptomiese analise van blaar-, skiet- en wortelmateriaal gebruik om 'n transkripsoom (125 350 fragmente, N50 van 1 036 bp) op te stel, en daarna gebruik as bewys om geen-modelle op die genomiese fragmente af te lei. Onvoldoende filogenieë is verkry vir *Amaranthus* spesies met chloroplast “barcoding” gene (*MatK*, *rbcL* en *ITS*), wat swak inter- en intraspesie-resolusie geopenbaar het. Met behulp van die volgorde data uit die konsep genoom, was dit egter moontlik om 'n vol-lengte chloroplast volgorde vir *A. tricolor* te onttrek (met ongeveer 200x volgorde dekking) en te annoteer. Deur die *A. tricolor* volgorde as verwysing te gebruik, kon die res van die spesies se chloroplaste saamgestel word deur middel van lae vlak volgorde bepaling. Filogenomiese analise van die *Amaranthus*-spesies het die klassifikasies van onbekende spesies moontlik gemaak, en verder lig gewerp op die moontlike domestikasie van huidige gekweekte amarant. *Amaranthus tricolor* groepeer met ander blaar amarant, *A. blitum*, *A. viridis* en *A. graecizans*, wat die subgenus *Albersia* verteenwoordig. Daarbenewens het die genoom 671 potensiële SSR lokusse onthul, en verdere validering van die merker sub-stel het 6 robuuste SSR merkers vir inter- en intraspesie genotipering beskikbaar gestel. Resultate vir alleel inligting verkry uit gel-gebaseerde kapillêre elektroforese (ABI) en volgende generasie volgorde bepaling (NGS) tegnologieë was hoogs vergelykbaar. Die NGS analise van SSR-merkers, het informatiewe genetiese verwantskappe tussen *Amaranthus*-toetredings ge-openbaar. Die genoom bepaling vir *A. tricolor* het gelei na die suksesvolle ontwerp van molekulêre gereedskap vir hieropvolgende molekulêre analises in die *Amaranthus* genus. Hierdie genomiese/ molekulêre instrumente kan toekomstig gebruik word om hulp te verleen met die kies en implimentering van genotipes vir voornemende teelprogramme en moontlik uiteindelik bydra tot wêreldwye voedselsekuriteit.

Acknowledgements

“Die Here is my Herder, niks sal my ontbreek nie”

Psalm 23:1

I would like to sincerely thank my mentor, advisor and supervisor, Prof Dave Berger, for his patience, motivation and continued support throughout my scientific career. Not only has Prof Berger fueled my appreciation of scientific research, he has also taught me the value of critical thinking and determination. The last 12 years have been an awesome learning curve, and I thank you!

Besides my advisor, I would like to thank Dr Damaris Odeny and Dr Martin Coetzee for their enduring patience and assistance with the writing of the thesis and research publication. Their shared knowledge and expertise enabled me to produce research of high quality, and I will be sure to carry your insightful comments with me in future research endeavors.

This work was funded by the Agricultural Research Council in conjunction with the Department of Science and Technology, as well as the National Research Foundation. I thank the ARC for giving me the opportunity to work on my PhD at the Biotechnology Platform, where the latest technology and expertise were always available. Thank you to Dr Farai Muchadeyi and the Core Facility members for their assistance with various aspects of my project, from conceptualization to experimental protocols and data analysis. A special thank you to Dr Charles Hefer for sharing his immense knowledge on genomics and bioinformatics, without whom I would not have been able to interpret my results. In addition, I would like to thank Mr. Willem Jansen van Rensburg and his team at ARC-VOPI for supplying the germplasm of *Amaranthus* species, as well as their service in plant growth and maintenance.

Lastly, I would like to express my eternal gratitude and appreciation to my husband, Chris Viljoen. Thank you for the enduring support, without which I would certainly have given up a long time ago. Thank you for carrying me for the last 16 years of studying, you have been my rock through the dark times. Similarly, thank you to my other biggest “fans”, my mom Sophia, my dad Gerhard, my sister Monica and my oupa and oma, Jan and Annetjie, for always believing in me, and never once faltering in your support. To my other mom (Ingrid) and dad (Frik), words cannot express how I have appreciated your kindness and understanding during the course of my studies. And finally, a big thank you to the rest of my family and friends, for keeping me afloat, motivated and inspired to complete my PhD. To my Golden Girls, Minique de Castro, Stephanie Cornelissen and Annie Gous, I love you awesome nerds!

“It’s not what you know in life, it is who you know”

List of Tables

Table 1.1: Comparison of amaranth with recommended daily allowance

Table 1.2: Nutritional properties of selected *Amaranthus* species per 100g edible portion

Table 2.1: Globally prioritized orphan crop species for research and development

Table 2.2: Examples of transgenic orphan/neglected crop species

Table 3.1: Read quality control (QC) for genomic and transcriptomic data

Table 3.2: Summary of whole genome sequencing assemblies of *A. tricolor* compared to previously assembled *A. hypochondriacus* and *Beta vulgaris*

Table 3.3: Summary of repetitive regions within the *A. tricolor* genomic scaffolds

Table 3.4: Summary of *de novo* transcriptome sequencing assemblies of *A. tricolor* compared to previously assembled *A. hypochondriacus*

Table 3.5: Genome annotation statistics of *A. tricolor*

Table 4.1: *Amaranthus* germplasm sets

Table 4.2: Species groupings of GRIN and SAG accessions

Table 4.3: Primer and repeat properties for 23 microsatellite loci identified from *A. tricolor* scaffolds

Table 4.4: Multiplexed PCR layout for each accession

Table 4.5: Genotyped identified using ABI capillary gel electrophoresis technology and Genemapper v4 software

Table 4.6: Genotyping calls using Illumina NGS MiSeq technology and MAND script

Table 4.7: Genetic diversity, PIC values and fixation indices for ABI vs. MiSeq data

Table 4.8: Genetic diversity, PIC values and fixation indices for *Amaranthus* germplasm set based on MiSeq SSR data after the addition of SAG genotypes

Supplementary Table S4.1: *Amaranthus tricolor* genome assembly statistics

Supplementary Table S4.2: SSR Discovery on *A. tricolor* scaffolds using the MSATCOMMANDER 1.8.0 software package

Supplementary Table S4.3: Estimated cost analysis of marker screening using ABI vs NGS technology

Table 5.1: *Amaranthus* germplasm sets

Table 5.2: Additional barcoding region statistics

Table 5.3: Classification of gene regions identified on chloroplast genome

Table 5.4: Barcoding nucleotide site analysis

Table 5.5: Whole chloroplast GRIN nucleotide site analysis

Table 5.6: Whole chloroplast GRIN and SAG nucleotide site analysis

Supplementary Table S5.1: Chloroplast gap closing primer statistics

Supplementary Table S5.2: Whole chloroplast mapping statistics

Supplementary Table S5.3: Whole angiosperm chloroplast sequences obtained from GenBank (NCBI)

Supplementary Table S5.4: Genic information for partitioned phylogenomics analysis

List of Figures

Figure 1.1: Global distribution of *Amaranthus* species

Figure 1.2: *Amaranthus caudatus*

Figure 1.3: *Amaranthus cruentus*

Figure 1.4: *Amaranthus hypochondriacus*

Figure 1.5: *Amaranthus dubius*

Figure 1.6: *Amaranthus blitum*

Figure 1.7: *Amaranthus tricolor*

Figure 1.8: *Amaranthus graecizans*

Figure 2.1: Interacting strategies between crop genotypes, environmental influences and application of efficient management strategies

Figure 2.2: Simplified diagram of genomics/marker assisted selection steps

Figure 3.1: Mature *Amaranthus tricolor* plants

Figure 3.2: High molecular weight genomic DNA extraction from leaves of a 30-40-day old *A. tricolor* plant

Figure 3.3: Total RNA extracted from 30-40-day old *A. tricolor* leaves

Figure 3.4: *Amaranthus tricolor* scaffold size frequencies

Figure 3.5: BUSCO analysis of the genome assembly

Figure 3.6: The number of trinity genes/transcripts (A) and contig statistics (B) of unfiltered trinity statistics compared to filtering of FPKM >1 and FPKM >2

Figure 3.7: Percentage of scaffold annotations with correlating AED scores

Figure 3.8: BUSCO analysis of *A. tricolor* protein predictions

Figure 3.9: Top % species nucleotide BLAST hits against the SwissProt database

Figure 3.10: Representation of the top 25 protein annotations found within the *A. tricolor* genome

Figure 3.11: Gene ontology terms could be assigned to 55.4% of the identified proteins from the *A. tricolor* genome.

Figure 3.12: Most dominant GO terms identified

Figure 4.1: Steps followed by the Microsatellite Analysis from NGS Data (MAND) script

Figure 4.2: Top nucleotide BLAST hit results for individual microsatellite sequences

Figure 4.3: Sanger sequencing of ATR32 in *A. graecizans* (GRIN15), *A. tricolor* (SAG29), *A. caudatus* (GRIN3) and *A. cruentus* (GRIN8) reveals the inter-species polymorphic nature of the marker

Figure 4.4: Multiplexed amplification of two marker sets on 10 diverse accessions (GRIN1-GRIN10) of *Amaranthus*

Figure 4.5: Number of raw reads (blue) and number of filtered reads (red) generated for each *Amaranthus* (GRIN) accession on one lane of the Illumina MiSeq platform

Figure 4.6: Representation of MAND script filtering steps for two marker loci

Figure 4.7: Allele frequencies observed during (A) ABI capillary electrophoresis and (B) NGS MiSeq analysis

Figure 4.8: Principle component analysis of 59 *Amaranthus* germplasm accessions

Figure 4.9: Dendrogram analysis of 59 *Amaranthus* germplasm accessions based on MiSeq SSR data

Figure 4.10: Loci unique to a species group (A) and the percentage contribution of each marker to the private allele set (B) across the entire germplasm set (GRIN and SAG)

Supplementary Figure S4.1: Illumina MiSeq library preparation for *Amaranthus* accessions

Supplementary Figure S4.2: Sanger sequencing of SSR markers

Supplementary Figure S4.3: Sequence alignments of Sanger sequenced ABI fragments and MiSeq sequenced fragments

Figure 5.1: Circular gene map for *A. tricolor* chloroplast genome

Figure 5.2: Classification of species within the angiosperm plant group based on maximum likelihood phylogeny of whole chloroplast genomes

Figure 5.3: *Amaranthus* phylogeny based on DNA barcoding of chloroplast *matK*, *rbcL* and nuclear ITS gene regions

Figure 5.4: Phylogeny of taxonomically described *Amaranthus* accessions

Figure 5.5: Phylogenetic distribution of unknown *Amaranthus* accessions in relation to previously described *Amaranthus* species

Supplementary Figure S5.1: Mapping of raw *A. tricolor* Illumina sequencing reads to the *Beta vulgaris* chloroplast

Supplementary Figure S5.2: *Amaranthus* phylogeny based on DNA barcoding of chloroplast *matK*, *rbcL* and nuclear ITS gene regions using Bayesian analysis

Supplementary Figure S5.3: Amaranth phylogeny based on Bayesian analysis of whole chloroplast genome GRIN sequences

Supplementary Figure S5.4: Bar graph indicating the percentage disagreement between a new barcoding phylogeny vs. the whole chloroplast phylogeny using the TOPD/FTMS software

Supplementary Figure S5.5: *Amaranthus* phylogeny based on DNA barcoding of chloroplast *matK*, *rbcL* and nuclear ITS gene regions on the entire germplasm set

Abbreviations

°C	Degrees Celsius
16S MLSP	Metagenomic Sequencing Library Preparation
ABI	Applied Biosystems Genetic Analyser
AED	Annotation Edit Distance
AFLP	Amplified Fragment Length Polymorphism
AK	Aspartate Kinase
ALS	Acetolactate Synthase
AOCC	African Orphan Crops Initiative
ARC-BTP	Agricultural Research Council Biotechnology Platform
ARC-VOPI	Agricultural Research Council Vegetable and Ornamental Plant Institute
ATP	Adenosine TriPhosphate
ATR	<i>A. tricolor</i> SSR Marker
AU-NEPAD	African Union's New Partnership for African Development
AVRDD	World Vegetable Centre
BLAST	Basic Local Alignment Sequence Tool
Bp	Basepairs
BS	Bootstrap
BUSCO	Benchmarking Universal Single-Copy Orthologue
CDS	Coding Sequences
CEG	Core Eukaryotic Genes
CEGMA	Core Eukaryotic Genes Mapping Approach
cm	Centimetre
CNV	Copy Number Variation
COI	Cytochrome Oxidase I

CRISPR-cas	Clustered Regularly Interspersed Short Palindromic Repeat Associated Proteins
DendroUPGMA	Dendrogram of Unweighted Pair Group Method with Arithmetic Mean
DHDPS	Dihydrodipicolinate Synthase
DNA	Dioxyribonucleic Acid
dNTP	Deoxyribonucleotide Triphosphate
DOGMA	Dual Organellar Genome Annotator
dsDNA	Double Stranded DNA
EB	Elution Buffer
EDR2	Enhanced Disease Resistance 2
eQTL	Expressed Quantitative Trait Loci
eSNPs	Expressed Single Nucleotide Polymorphism
ESS	Effective sample size
EST	Expressed Sequence Tags
FAO	Food and Agriculture Organization
FI	Fixation Index
FLG	Fragment Length Groups
FPKM	Fragments Per Kilobase of Transcript per Million
g	Gram
GAB	Genomics Assisted Breeding
GC	Guanine-Cytosine
GDP	Gross Domestic Product
GM	Genetically Modified
GO	Gene Ontology
GTR+G	General Time Reversible model plus Gamma
GWAS	Genome Wide Association Studies

HIV	Human Immunodeficiency Virus
HKY	Hasegawa-Kishino-Yano
HMM	Hidden Markov Model
iBOL	International Barcoding of Life Initiative
ICRAF	World Agroforestry Centre
IDT	Integrated DNA Technologies
INDEL	Insertion-Deletion
IR _A /IR _B	Inverted Repeat region A and B
Kb	Kilobases
KEGG	Kyoto Encyclopaedia of Genes and Genomes
kg/m ²	Kilogram per square meter
L	Linneaus
LINE	Long Interspersed Nuclear Elements
LSC	Large Single Copy region
LTR	Long Terminal Repeats
m	Meter
MAND	Microsatellite Analysis using NGS Data
MAS	Marker Assisted Breeding
Mb	Megabases
MCMC	Markov Chain Monte Carlo
MEGA	Molecular Evolutionary Genetics Analysis
min/s	Minutes per Second
MJ/kg	MegaJoule per kilogram
mm	Millimeter
mM	Millimolar

MP1	Mixed group 1
MPI	Message Passing Interface
MUSCLE	Multiple Sequence Comparison by Log Expectation
Mya	Million years ago
NCBI	National Centre for Biotechnology Information
NCRPIS	North Central Regional Plant Introduction Station
NGS	Next Generation Sequencing
Nl/kg VS	Normal litre biomethane per kilogram Volatile Solids
NST1	NAC Secondary wall thickening Promoting Factor 1
ORF	Open Reading Frame
PBC	Plant Breeding Academy
PCoA	Principle Component Analysis
PCR	Polymerase Chain Reaction
PIC	Polymorphism Information Content
PP	Posterior Probability
PPR	PentatrικοPeptide Repeats
QTL	Quantitative Trait Loci
Q-value	Quality Value
RAPD	Random Amplified Polymorphic DNA
RAxML	Randomized Axelerated Maximum Likelihood
RDA	Recommended Daily Allowance
RFLP	Random Fragment Length Polymorphism
RIN	RNA Integrity number
RNS	Ribonucleic Acid
RPW8	Powdery Mildew Resistance Protein

RSEM	RNA-Seq by Expectation-Maximization
SD	Split Difference
SINE	Short Interspersed Nuclear Elements
SNAP	Semi-HMM-based Nucleic Acid Parser
SNP	Single Nucleotide Polymorphism
SRA	Short Read Archive
SSA	Sub-Saharan Africa
SSC	Small Single Copy region
SSR	Simple Sequence Repeats
t/ha	Tons per hectare
TALEN	Transcription Activator-like Effector Nucleases
TOPD/FTMS	Topological Distance/From Multiple to Single
TRF	Tandem Repeat Finder
USDA-ARS	United States Department of Agriculture – Agricultural Research Service
WGD	Whole Genome Duplication
WHO	World Health Organization
WWF	World Wildlife Fund
ZFN	Zinc Finger Nuclease
μM	Micromoles

Preface

The agricultural revolution started roughly 20 000 years ago with the naturalization and willful propagation of certain food species, which has altered human consumption patterns significantly. Consequently, global food security became dependent on a handful of commercialized food crops, leading to a reduced dietary diversity. Today, only four species provide 60% of the global proteins and calorie requirement; rice, wheat, maize and potato (FAO 2015). The narrowed number of crops accessible for consumption has led to high health risks associated with what has recently been coined as “the triple burden of human malnutrition”, broadly composed of (i) malnutrition (lack of macronutrient intake such as calories and proteins); (2) micronutrient deficiencies (lack of micronutrients such as vitamins and minerals) and (3) overweight and obesity (excessive dietary intake) (Gomez *et al.*, 2013). The Green Revolution, implemented globally between 1930 and 1960, aimed to develop and distribute high-yielding or modern varieties of the most common food crops to boost the daily calorific and protein intake of consumers, and consequently target the effects of undernutrition. However, by doing so, two facets of malnutrition were neglected and gave rise to an increase in incidences of micronutrient deficiencies and/or overweight and obese individuals (Gomez *et al.*, 2013). Furthermore, the effects of global warming and the surging human population growth will be felt acutely in the agricultural sector, with drastic changes in weather patterns and less available arable land for farming practices having large impacts on the state of food security.

Several nutritional challenges associated with human malnutrition can be addressed by exploiting natural diversity present in native/alternative plant species (van Rensburg *et al.* 2007). Traditional vegetables thrive in their native habitats and are often highly adaptable to changing environmental patterns due to their natural genetic diversity. Advantages of farming an array of traditional vegetables include a highly diversified diet, retaining cultural heritage, creating employment opportunities and lessening farming demands as most traditional vegetables require less maintenance and have reduced associated costs compared to their exotic counterparts (Tadele and Assefa 2012; Dube *et al.* 2016). At least two of the three aspects of human malnutrition can be addressed by focusing on farming traditional vegetables, since most of them contain enough macronutrients to fight undernourishment, and also sufficient amounts of micronutrients to combat the effects of vitamin and mineral deficiencies (Alemayehu *et al.* 2015). Since most traditional crops are grown in their natural habitats, they already have inherent adaptations to the local ecosystems (pertaining to

climatic or soil conditions, pests and diseases). In contrast, the farming of exotic vegetables may require additional input costs in water/nutrient application or pest and disease control (Mlakar *et al.* 2010; Oh *et al.* 2011; Costinot *et al.* 2016). The farming of exotic vegetables can have a large impact on a country's economy, since they become more expensive to produce. Global concerns for biodiversity loss and hidden benefits of underutilized plants have led researchers to renew their interest in developing and maintaining traditional crops. Research shows that orphan/neglected crops have the potential to contribute to nutrition, income generation, food security and general health of the public (Vorster and Van Rensburg 2005).

The African Union's New Partnership for Africa's Development (AU-NEPAD agency), Mars Incorporated, World Agroforestry Centre (ICRAF), University of California Davis, the World Wildlife Fund (WWF) and the Plant Breeding Academy (PBC) (among others) have partnered to form the African Orphan Crops Consortium (www.africanorphancrops.org) to identify, sequence and assemble the genomes of important traditional African food crops. Their aim is to make these highly nutritious African crops available to rural and urban consumers by promoting modern breeding practices. The consortium will develop several molecular tools (whole genome sequences, transcriptome sequencing and gene discovery) to investigate and improve the nutritional content and production capabilities of traditional vegetables. Among the 101 traditional vegetables chosen for this purpose, three *Amaranthus* species are also present; *A. blitum*, *A. cruentus* and *A. tricolor* (leafy amaranth accessions). The Agricultural Research Council (Biotechnology Platform) is actively involved with the next generation sequencing of a subset of these African crop species, including the *A. tricolor* accession discussed in this thesis.

Several *Amaranthus* species are commercially grown as traditional, leafy vegetables across the world (North, Central and South America, Eastern Asia and Africa) and form a small part of the commercial trade in leafy and grain vegetables in several provinces of South Africa (Limpopo, North West, Mpumalanga and Kwa-Zulu Natal) (van Rensburg *et al.* 2007). The genus consists of approximately 60 species, composed mostly of cosmopolitan weeds, but also a few cultivated species utilized as grain or leafy vegetables and garden ornamentals (Mlakar *et al.* 2010). Amaranth species utilized for their grain (*A. caudatus*, *A. cruentus*, *A. hypochondriacus*) are known to be high in protein and essential amino acids (such as lysine and methionine), while species utilized as leafy vegetables (*A. blitum*, *A. viridis*, *A. tricolor*, *A. cruentus*, *A. graecizans*, *A. hybridus*) contain considerable amounts of vitamins (especially β -carotene) and minerals (O'Brien and Price

1983; Steyn *et al.* 2001). Coupled with the natural resistance to pests and diseases, adaptability to changing weather patterns and high nutritional values, amaranth is an ideal choice for sustainable food security across the world.

A need has arisen for the development of comprehensive genomic tools to harness and exploit beneficial traits in traditional leafy vegetable crops such as *Amaranthus*. In particular, *A. tricolor* is a popular choice in several African countries to grow as a leafy vegetable (Grubben 1980; Jansen *et al.* 2004; Ribeiro *et al.* 2017). The natural resistance to several pests and diseases, palatable texture and high nutritional content makes it an ideal candidate for further research. It has also been listed as one of the most important African crops for commercial development by the AOCC, which will therefore benefit from the data generated during the present study.

The depth of information that can be obtained from next generation sequencing has reached unparalleled heights. The identification hundreds of useful genes or informative genetic markers can be done in a single experiment, which in turn could initiate marker assisted breeding programs and superior lines can be developed within a fraction of the usual time. In this thesis, a draft genome and transcriptome for the traditional leafy vegetable, *A. tricolor*, was sequenced and assembled to facilitate the development of molecular tools to investigate a geographically diverse *Amaranthus* germplasm collection with comparisons to global *Amaranthus* species. Genome sequencing with supporting transcriptome data and marker analysis will be the first step towards a more long-term varietal improvement of this crop in South Africa. The results of this study will provide important and essential resources for the enhanced breeding of superior *Amaranthus* genotypes.

Hypotheses

1. The sequencing, assembly and annotation of a draft genome sequence for *A. tricolor* will form the foundation for downstream molecular tool development, which could enable researchers to exploit targeted agricultural traits for crop improvement.
2. Whole chloroplast phylogenomics will reveal/confirm the placement of *A. tricolor* within the expected *Albersia* subgenus, as well as improved phylogenetic relatedness to other *Amaranthus* species, compared to the commonly used genetic barcoding technique.
3. Whole chloroplast genomics will further enable the identification of unknown *Amaranthus* accessions through phylogenomic grouping with previously described species.
4. The draft genome sequence will reveal a high number of microsatellite repeat loci that are potentially polymorphic across diverse species of the genus *Amaranthus*.
5. The Illumina next generation sequencing platform can be used to quickly and accurately screen highly multiplexed microsatellite markers across diverse *Amaranthus* species, to assess their genetic diversity.
6. Dendrogram analysis using the microsatellite markers will reveal the genetic relatedness between *A. tricolor* and closely related *Amaranthus* species, while also enabling the identification of unknown accessions.

Research Questions

1. Can a draft genome for *Amaranthus tricolor* be assembled and annotated using genomic and transcriptomic paired end and long mate pair sequencing data?
2. Can complete chloroplast phylogenomics improve our understanding of the phylogenetic relatedness between different *Amaranthus* species?

3. What is the level of diversity within the South African amaranth germplasm set in relation to global representative accessions, and can inferences be made to the identities of the unknown accessions using microsatellite marker analysis?

Objectives

1. To assemble and annotate a draft genome sequence for *Amaranthus tricolor*, produced with genomic and transcriptomic next generation sequencing Illumina technologies.
2. To extract sequencing data from the genomic assembly and construct a whole chloroplast genome for *Amaranthus tricolor* and subsequently map and assemble chloroplast genomes for the entire germplasm set.
3. To perform whole chloroplast phylogenomics to reveal species relationships between members of the *Amaranthus* genus.
4. To investigate clade resolution between *Amaranthus* accessions using traditional molecular barcoding techniques.
5. To identify and test a set of polymorphic microsatellite markers from the draft genome sequence of *A. tricolor*, which can be used to investigate inter- and intra-species relationships between diverse accessions of the *Amaranthus* genus.
6. To evaluate next generation sequencing for the typing of polymorphic microsatellites as an alternative to conventional capillary electrophoresis or Sanger sequencing techniques.
7. To identify unknown *Amaranthus* accessions within the South African germplasm set through species relationships with global representative accessions, using marker dendrogram and phylogenomic analysis.

Structure of Dissertation

Chapter 1: An overview of published literature pertaining to *Amaranthus* taxonomy, nutritional content, economical uses and breeding.

Chapter 2: A review on genome-aided molecular breeding and transgenic technologies on orphan crops. The review also touches on the disciplines which have an impact on the management of orphan crops, and elaborates on cutting edge technologies that can be employed to develop genomic tools for crop development for future food security.

Chapter 3: This chapter focusses on the sequencing and assembly of a draft genome sequence for *Amaranthus tricolor*, an important traditional leafy vegetable. Combined with transcriptomic data of different tissue types, it was possible to partially annotate the genome for potential downstream gene discovery.

Chapter 4: A set of microsatellite markers were identified from the draft genome assembly and screened across the *Amaranthus* germplasm set using capillary electrophoresis and confirmed with Sanger sequencing. The polymorphic markers were also screened using the next generation sequencing Illumina platform and analyzed with a novel bioinformatic script.

Chapter 5: Using the draft genome sequencing reads, it was possible to extract, assemble and annotate a complete chloroplast genome for *A. tricolor*, by using the whole chloroplast sequence of a close relative (*Beta vulgaris*) as a reference. Consequently, the full-length chloroplast genomes for the remaining accessions within the *Amaranthus* germplasm set were sequenced and assembled by using the newly assembled *A. tricolor* chloroplast as a reference sequence. A whole chloroplast phylogeny was constructed to investigate the genetic relatedness of the different *Amaranthus* species.

Chapter 6: This chapter includes a general discussion of the results obtained during this study, the shortcomings and recommendations for future research within the *Amaranthus* genus.

Publications

Published:

Viljoen E, Odeny DA, Coetzee MPA, Berger DK, Rees DJG. (2018) Application of Chloroplast Phylogenomics to Resolve Species Relationships Within the Plant Genus *Amaranthus*. *Journal of Molecular Evolution* 86:216-239

In Preparation:

Viljoen E., Rees D.J.G., Berger D.K. Review on crop genetics and improvement with emphasis on orphan crops for sustainable food security.

Viljoen E., Rees D.J.G., Berger D.K. Species relationships within the *Amaranthus* genus revealed by Illumina MiSeq sequencing of microsatellite markers.

Conferences and Workshops:

International:

6th International Conference of the European Amaranth Association: Neglected and Under-Utilized species research in the 21st Century; Nitra, Slovakia

Oral: Viljoen E., Odeny D.O., Rees D.J.G. & Berger D.K. (2012) Progress in sequencing *Amaranthus tricolor* and chloroplast genome phylogenetics.

Plant and Animal Genomes Conference XIII; San Diego, USA

Poster: Viljoen E., Rees D.J.G. & Berger D.K. (2015) Next Generation Sequencing, Assembly and Annotation of Chloroplasts for Phylogenetic Analysis of Amaranths.

Local:

9th Southern African Plant Breeding Symposium (SAPBA)

Oral: Viljoen E., Odeny D.O., Rees D.J.G. & Berger D.K. (2012) Development of genomic tools for nutritional assessment of *Amaranthus* species in South Africa.

UP Plant Science Department Postgraduate Symposium 2012

Oral: Viljoen E., Odeny D.O., Rees D.J.G. & Berger D.K. (2012) Progress in sequencing *Amaranthus tricolor* and chloroplast genome phylogenetics. 1st Prize for Best PhD Presentation

10th Southern African Plant Breeding Symposium (SAPBA)

Oral: Viljoen E., Rees D.J.G. & Berger D.K. (2014) Next generation sequencing, assembly and annotation of chloroplasts for phylogenetic analysis of amaranths.

Professional Development Program (PDP) Conference hosted by the Agricultural Research Council

Oral: Viljoen E., Rees D.J.G. & Berger D.K. (2014) Sequencing the *Amaranthus tricolor* genome and amaranth chloroplast phylogenetics. 2nd Prize for Best Crop Science PhD

2nd Joint Conference of the South African Society for Bioinformatics (SASBi) and the South African Genetics Society (SAGS)

Oral: Viljoen E., Rees D.J.G. & Berger D.K. (2014) Next generation sequencing, assembly and annotation of chloroplasts for phylogenetic analysis of amaranths.

National Three Minute Thesis (3MT) Competition hosted by the Post Graduate School at the Free State University

Oral: Viljoen E., Rees D.J.G. & Berger D.K. (2015) Are Orphan Crops the Future? 2nd Prize for Best PhD Presentation

References

- 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
- Costinot A, Donaldson D, Smith C (2016) Evolving comparative advantage and the impact of climate change in agricultural markets: Evidence from 1.7 million fields around the world. *Journal of Political Economy* 124:205-248
- Dube T, Moyo P, Ncube M, Nyathi D (2016) The impact of climate change on agro-ecological based livelihoods in Africa: A review. *Journal of Sustainable Development* 9:256-267
- FAO I (2015) WFP. 2015. The State of Food Insecurity in the World
- Gómez MI, Barret CB, Raney T, Pinstup-Andersen P, Meerman J, Croppenstedt A, Carisma B, Thompson B (2013) Post-green revolution food systems and triple burden of malnutrition. *Food Policy* 42:129-138
- Grubben G (1980) Germplasm of vegetable and grain amaranths. *Plant Genetic Resources Newsletter*:33-35
- Jansen P, Grubben G, Denton O, Messiaen C, Schippers R, Lemmens R, Oyen L (2004) *Amaranthus hypochondriacus* L. *Plant Resources of Tropical Africa* 2:78-80
- 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
- O'Brien GK, Price M (1983) Amaranth: Grain and vegetable types. Educational Concerns for Hunger Organization (ECHO) Technical Note. ECHO, North Fort Myers, FL
- Oh JE, Kwon Y, Kim JH, Noh H, Hong S-W, Lee H (2011) A dual role for MYB60 in stomatal regulation and root growth of *Arabidopsis thaliana* under drought stress. *Plant Molecular Biology* 77:91-103
- Ribeiro JE, Pieterse PJ, Famba SI (2017) Vegetative growth of *Amaranthus hybridus* and *Amaranthus tricolor* under different watering regimes in different seasons in southern Mozambique. *South African Journal of Plant and Soil* 34:201-210
- Steyn N, Olivier J, Winter P, Burger S, Nesamvuni C (2001) A survey of wild, green, leafy vegetables and their potential in combating micronutrient deficiencies in rural populations: research in action. *South African Journal of Science* 97:276-278
- Tadele Z, Assefa K (2012) Increasing food production in Africa by boosting the productivity of understudied crops. *Agronomy* 2:240-283
- Van Rensburg WJ et al. (2007) African leafy vegetables in South Africa. *Water SA* 33:317-326
- Vorster H, Van Rensburg W. (2005) Traditional vegetables as a source of food in South Africa: Some experiences. In: African Crop Science Society Conference Proceedings pp 669-671

CHAPTER 1

LITERATURE REVIEW

Introduction

Across the world, the globalization of agriculture has led to the technical and agricultural development of only a few commercial plant species, which are responsible for meeting the global demand for foodstuff. Of the possible 10 000 plant species in the world that are edible, 150 are commercialized at a significant global scale. Only 12 of these are responsible for providing 80% of human dietary energy from plants and over 60% of the worldwide requirement for proteins and calories. The four most important crop species globally are rice, wheat, maize and potato (Mlakar *et al.* 2010). The world population has been growing steadily since the 1400s, and is projected to increase to approximately 9.7 billion people (from 7.5 billion in 2017) by 2050, but it can also be as high as 10 billion (United Nations, 2004; (Oh *et al.* 2011). The ever-intensifying effects of global warming may increase the incidence in extreme weather disasters in certain regions, such as droughts, floods and extreme temperatures (Lesk *et al.* 2016). It has been estimated that the impact of climate change on global agricultural markets can lead to 0.26% reduction in global Gross Domestic Product (GDP), which corresponds to approximately one-sixth of the total crop value (Mlakar *et al.* 2010; Oh *et al.* 2011; Costinot *et al.* 2016). To cope with the estimated 22% increase in world population within the next 30 years, the agricultural sector will have to increase its production by up to 70% to compensate for the increase in food demand (Ebert 2014).

Sub-Saharan Africa (SSA) is home to approximately 16% of the world's population, where it is estimated that one in three people are chronically hungry (Odhav *et al.* 2007; Alemayehu *et al.* 2015). Poverty, low employment rate, natural disasters and the soaring incidence of the human immunodeficiency virus (HIV) are some of the factors that contribute to the livelihood of people in Africa (UNAIDS, 2007). The current changes in climate are already affecting crop productivity in Africa, and will result in a decline in production systems, infrastructure quality and market availability of several important crop species (Tadele and Assefa 2012; Dube *et al.* 2016). (Funk *et al.* 2008) studied the satellite observations across Eastern and Southern Africa, and found that the rainfall has decreased by 15% in the principle growing seasons in food insecure countries of these areas. Predictions show that due to the warming of the Indian Ocean, the rainfall in continental Africa will decrease to a point where severe drought occurs, and is projected to increase the already malnourished population by approximately 50% by the year 2030. Due to the long-term changes in rainfall in Southern Africa, drought has become more intense and widespread (Fauchereau *et al.*, 2003).

However, in West Africa, a wetter environment is expected, with less severe drought conditions. Crops that have historically adapted to a drier environment will struggle to grow optimally in these altered conditions. These alterations in rainfall patterns will have some of the most drastic effects on agricultural practices across the African continent (Sarr 2012; Dube *et al.* 2016).

In general, most of the population in SSA depends on the few commercial crops as carbohydrate staples, but they lack sufficient amounts of essential micronutrients (vitamins and minerals), which are crucial in the human diet. Therefore, indigenous leafy vegetables are of high importance as they are often the only source of nutritional food (and essential micronutrients) especially in rural areas. The consumption patterns of leafy vegetables differ among households in different countries, and are influenced by several factors such as poverty, distance to fresh produce markets, degree of urbanization and seasonable availability (van Rensburg *et al.* 2007). Several studies have shown micronutrient status of these indigenous vegetables are as high as, or higher than, more conventionally grown leafy vegetables such as cabbage or spinach (Edmonds and Chweya 1997). People in SSA depend on staple foods prepared from cereal crops (millet, sorghum and maize), tubers (yam, cassava) and plantains (a vegetable like bananas) to obtain their daily requirements of micronutrients, but these crops do not always meet the WHO/FAO recommended daily intake (Oniang'o *et al.* 2004). A diversified diet is needed to reach the requirement of all micronutrients, and several indigenous vegetable species have the potential to contribute significantly in alleviating chronic problems associated with the lack of micronutrients (Alemayehu *et al.* 2015). Due to the diminished interest in wild leafy vegetables over a long period of time, their true potential had almost become lost, but a revival of research attention has led to renewed research opportunities (Afolayan and Jimoh 2009). Therefore, the shift back to the use of traditional vegetables was both inevitable and of utmost importance.

Amaranths are rich sources of micronutrients and dietary minerals with unique nutritional compositions that could enhance the composition of processed foods. Compared to other grain species, amaranth has superior amounts of protein (and limiting amino acids such as lysine and methionine), calcium and iron, all essential for a balanced, healthy diet. In addition, the gluten free starch is a healthy alternative to conventional grain species to people suffering from celiac disease to aid bone metabolism and relieve complications in the intestine. People following a vegetarian diet sometimes struggle to ingest sufficient amounts of protein due

to the lack of meat, but the protein levels within amaranth leaves are highly comparable to meat products (Jimoh *et al.*, 2018).

Amaranths have also been identified by the United States National Academy of Science of 1975 and the African Orphan Crops Consortium (2015) as a major potential crop with most promising economic value among several under-exploited tropical plants in Africa and across the world (www.africanorphancrops.org; Jimoh *et al.*, 2018).

Genus *Amaranthus* (L.)

The genus *Amaranthus* (L.) falls within the subfamily Amaranthoideae; family Amaranthaceae and order Caryophyllales. Within the genus *Amaranthus*, three subgenera are recognized, including subgenus *Amaranthus* (L.), subgenus *Acnida* (L.) Aellen ex K. R. Robertson and subgenus *Albersia* (Kunth) Gren & Godr (Mosyakin and Robertson 1996). The genus consists of approximately 60-75 species, incorporating mostly cosmopolitan weeds, but also cultivated species used as grain sources, leafy vegetables and garden ornamentals (Sauer, 1967; Mlakar *et al.* 2010; Stetter and Schmid, 2017; Waselkov *et al.*, 2018). The different species within the genus *Amaranthus* is often difficult to classify due to small morphological differences, intermediate forms of interspecific hybrids that occur in nature, environmental influences and their overall general similarity (Achigan-Dako *et al.*, 2014). Due to these reasons, researchers often interpret the species boundaries differently (Sauer, 1967). *Amaranthus* subgenus *Acnida* is exclusively composed of dioecious weeds, which represent a distinct morphological group from the remaining two monoecious subgenera *Acnida* and *Amaranthus*. Most of the *Amaranthus* species utilized as leafy vegetables fall within the *Amaranthus* subgenus *Albersia* (*A. blitum*, *A. graecizans*, *A. tricolor* and *A. viridis*) (Mosyakin and Robertson, 1996). In contrast, all the grain and weedy species fall within *Amaranthus* subgenus *Amaranthus*, including *A. hybridus*, *A. hypochondriacus*, *A. cruentus*, *A. caudatus*, *A. quitensis*, *A. retroflexus* and *A. powellii*. Due to the origin of *A. dubius* (an allopolyploid species thought to have originated from a cross between *A. spinosus* and *A. hybridus/A. quitensis*), a nothosection has been suggested for its classification, namely *Amaranthus* subgenus *Amaranthus* nothosection *Dubia* (Mosyakin and Robertson, 1996). Accessions of the *Amaranthus* genus are often referred to by the generic term, amaranths. The three grain amaranth species (*A. hypochondriacus*, *A. caudatus* and *A. cruentus*) together with their putative progenitors

(*A. quitensis* and *A. hybridus*) are often referred to as the Hybridus complex. These five species readily cross with each other and are considered closely related. Species within this complex is commonly difficult to identify, due to the lack of reliable, fixed diagnostic qualitative characters and the presence of intermediate morphological traits (Adhikary and Pratt, 2015). In the past, species within the Hybridus complex were identified based on three morphological characteristics, including bract to utricle length ratios, sepal posture and seed color (pale seeds assumed to belong to the cultivated grain amaranth types). However, recently it has been found that many populations of cultivated species have dark seeds (Sauer, 1967; Adhikary and Pratt, 2015).

Amaranths are found throughout the worlds' tropical and temperate regions, and most are annuals of naturally open habitats. They produce abundant amounts of seed, which aids to continuously colonize different areas such as disturbed sites with full sun and little competition (Sauer 1967). Sauer (1950, 1967) report the occurrence of amaranths in pre-Columbian times in North America, with original samples having been contributed from North and South America. Historically, the native geographic trail of grain amaranths distribution faintly starts with the Indian tribes in southwestern America, and became more prominent through northwestern Mexico (also known as the major Mexican stronghold of the crop) and tapered off again towards southern Mexico. An isolated center of amaranth was found in Guatemala, and a larger one, which started in Peru through Bolivia and ended in the northern parts of Argentina (Figure 1.1). Amaranth is also a popular crop grown in parts of Asia (China, Nepal, the Himalayas), Europe (Austria, Czech Republic, Slovakia, Germany, Hungary, Poland, Russia, Italy and Slovenia) as well as isolated farmers in Australia and Eastern, Western and Southern Africa (Council 1984; Chan and Sun 1997; Mlakar *et al.* 2010). It is hypothesized that Europeans settlers transported amaranths to Africa in the early nineteenth century. Their spread was rapid across the continent, most probably as contaminants in sorghum or millet seeds. The tropical climate of most of Africa made it an ideal setting for the growth and subsequent domestication of the amaranth species.



Figure 1.1: Global distribution of *Amaranthus* species. The native habitats of amaranths can be followed from southwestern United States through Mexico and in Guatemala (green circles). In South America, populations are found from Peru, through Bolivia and into the northern parts of Argentina (Council 1984).

Weedy Amaranth

Weedy plants have detrimental effects on social, economic or conservation values of an area by invading natural, rural and agricultural areas, which cause damage to natural ecosystems or completely changes the plant communities they colonize. Weeds also compete with natural flora for sunlight, soil and nutrients; which may lead to weak native vegetation (Group 2001). Several weedy species are found within the *Amaranthus* subgenus *Amaranthus* (which includes nothosection *Dubia* and section *Centrusa*); mainly *Amaranthus viridis* (slender amaranth, green amaranth), *A. spinosus* (spiny amaranth, thorny amaranth), *A. retroflexus* (red-root amaranth), *A. hybridus* (smooth amaranth, smooth pigweed, red amaranth), *A. powellii* (green amaranth, Powell amaranth) and *A. quitensis* (ataco, sangorache). The wild amaranth species are usually distinguished from the cultivated grain amaranth in that it has indeterminate growth, and produce

seed at different parts of the plant, not only at terminal inflorescence. Grain amaranth tends to mature over a short period of time and have a dominant seed head (Council 1984). The weedy amaranths are usually spotted in abandoned lots or disturbed areas within urban settlements, and occur as pests in crops, roadsides and pastures. *Amaranthus palmeri* (*Amaranthus* subgenus *Acnida*), a common weed in soybean fields in the USA, has evolved resistance to several types of acetolactate synthase (ALS)-inhibitor herbicides, glyphosate (5-enolpyruvylshikimate-3-phosphate synthase inhibitors) and protoporphyrinogen oxidase inhibiting herbicides, which limits the herbicide options to manage Palmer amaranth in agricultural cropping systems (Nakka *et al.* 2017). It has recently been found that although *A. palmeri* could be resistant to each herbicide individually, biotypes with multiple resistance mechanisms have also recently been identified (Legleiter and Bradley, 2008). The mechanism of resistance was found to be target-site based, due to an amino acid substitution within the ALS gene (p.P197L) that likely causes a reduced sensitivity of the target enzymes to the herbicide (Kohrt *et al.* 2017; Nakka *et al.* 2017). Even though herbicide resistance becomes problematic to manage amaranth weed populations among other crop species, the targeted genetic studies within *A. palmeri* reveals the potential of site-directed mutagenesis studies in grain or leafy amaranth species, which are grown as main crops for small subsistence farmers struggling with other weeds. Furthermore, triazine and ALS herbicide resistance has also been identified in *A. rudis*, *A. tuberculatus*, *A. powellii*, *A. retroflexus* as well as *A. hybridus* (Crespo *et al.*, 2017). The cultivated, grain amaranths usually do not carry the same potential for weediness, and are less aggressive in their habits (Council 1984).

***Amaranthus viridis* (2n=34)**

It is hypothesized that *A. viridis* may have originated in Asia but had migrated to several tropical and sub-tropical regions across the world such as Europe, North America and Australia where it is considered a cosmopolitan weed. In Africa, certain cultivars are farmed in parts of Nigeria, Gabon and the Democratic Republic of Congo (Jansen *et al.* 2004). *Amaranthus viridis* is an ascending or erect perennial herb, which grows around 1 m tall with slender, branched, glabrous to sparsely pubescent stems. The leaves are alternating, simple; laminas are deltoid-ovate to rhomboid-oblong and a sinuate margin. Leaf surfaces tend to be glabrous to pubescent. The inflorescence consists of agglomerated cymes, which are arranged in axillary and mostly terminal spikes up to 12 cm long. Flowers are unisexual, with female flowers more

abundant. The seeds are slightly compressed, have an acute margin and are glossy black (Mosyakin and Robertson 1996). *Amaranthus viridis* is a weed that grows on disturbed or cultivated land, and flowers continuously (Jansen *et al.* 2004). As this species is mostly considered a problematic weed, it is not currently purposefully cultivated. However, the high nutritional value in the leaves identifies this species as a potential vegetable for future domestication and cultivation (Council 1984).

***Amaranthus spinosus* (2n=34)**

Amaranthus spinosus originated from southern and central America but has been introduced into other parts of the world including tropical Africa. It is considered the most common and widespread weedy amaranth of the New World tropical lowlands, but is rarely cultivated (Sauer 1967; Jansen *et al.* 2004). The plants are annual, erect, monoecious herbs (100-130 cm tall), with branched, glabrous to slightly pubescent green to purplish stems. Leaves are alternate and simple, with an ovate-lanceolate to rhomboid shape, entire and glabrous to slightly pubescent on the veins when immature. The most distinct characteristic of *A. spinosus* is that their stems usually have sharp spines, which can grow up to 2 cm long. Flowers are unisexual, with the male flowers arranged in a terminal spike at the base of the inflorescences (Jansen *et al.* 2004). It is almost impossible to distinguish between *A. spinosus* and *A. dubius* based on morphological characters; one of the only differences being axillary spines present on the inflorescence clusters in *A. spinosus* which is not present in *A. dubius*. However, spineless *A. spinosus* has been observed in Nigeria and can only be distinguished from *A. dubius* based on chromosome numbers (*A. spinosus*: 2n=34; *A. dubius*: 2n=32) (Mosyakin and Robertson 1996; Achigan-Dako *et al.* 2014). *Amaranthus spinosus* grows in a very diverse set of environmental conditions, including extremely arid circumstances and is mostly found along roadsides, railway yards, cropped land up to an altitude of 1400 m. These species are propagated by seed and may produce as many as 235 000 seeds per single plant. *Amaranthus spinosus* is a noxious weed across the world and is not grown as a commercial vegetable. Due to its very bitter taste, it is considered more of a famine food and only consumed when no other leafy vegetable options are available (Jansen *et al.* 2004).

Grain Amaranths

Three species of *Amaranthus* subgenus *Amaranthus* section *Amaranthus* produce large seed heads filled with pale, ivory colored edible seeds. *Amaranthus caudatus*, *A. hypochondriacus* and *A. cruentus* are collectively known as the ‘grain amaranths’ and have historically been used as toasted or milled grains and are referred to as ‘pseudo-cereals’ due to their non-grass nature (Das 2011). Grain amaranths have been shown to have a higher protein, starch and lysine content than more conventional cereal grains such as rice and maize, which gives them a nutritional advantage (Venskutonis and Kraujalis 2013). *Amaranthus hypochondriacus* L. (prince’s feather) and *A. cruentus* L. (purple amaranth, Mexican grain amaranth) are native to Mexico and Guatemala, while the third species, *A. caudatus* L. (love-lies-bleeding) is native to Peru and several Andean countries (Council 1984). Sauer (1950, 1967) originally proposed two hypotheses pertaining to the grain amaranth evolution. The first hypothesis was that there had been one domestication event of *A. cruentus* from *A. hybridus*, and the remaining domesticates (*A. hypochondriacus* and *A. caudatus*) evolved secondarily with repeated crossing of *A. cruentus* with two other wild species, *A. quitensis* in Southern America/Mexico and *A. powellii* in Northern America. The alternative hypothesis stated that each of the three grain species was domesticated in different regions, independently from one another from different wild progenitors; *A. hypochondriacus* from *A. powellii* in Mexico, *A. caudatus* from *A. quitensis* in South America and *A. cruentus* from *A. hybridus* in Central America (Xu and Sun 2001).

Recently, a study by Stetter et al., (2017) revealed evidence of a weak domestication syndrome within the *Amaranthus* genus. The grain amaranths seem to have been subjected to partial selection pressures in terms of seed size (pale seeds versus dark-seed colors of the wild/leafy types), as well as larger and more compact inflorescences. However, since there are sporadic occurrences of dark seeds among all three grain amaranth types (*A. caudatus*, *A. cruentus* and *A. hypochondriacus*), it is hypothesized that the pale seed morphology is not a fixed trait. Rather, it could be related to partial domestication in specific geographical regions or ongoing gene flow between the different amaranth species (Stetter et al., 2017).

***Amaranthus caudatus* (2n=32)**

Amaranthus caudatus is a native crop in the Andean highlands of Peru, Argentina and Bolivia, and the oldest relic of this species was found in a late-prehistoric site in Salta, Argentina. It was introduced into Europe in the 16th century, and subsequently spread to the Asian and African continents (Agong 2006). Today, the growth and utilization of this crop spans North and South America, Europe and Asia (Council 1984) but the exact distribution pattern in Africa is not known. In Ethiopia and Eritrea, it is grown for its grain or as an ornamental, and has also been recorded in Kenya, Uganda and other central, east and southern African countries (Figure 1.2 A) (Agong 2006). The taxonomy of *Amaranthus* in general is unclear, and over time at different geographical locations, alternative names have been assigned to the same species. Several studies have reached a consensus on the general physiology of *A. caudatus* (Sauer 1950; Sauer 1967; Costea *et al.* 2001a; Brenner 2002). These include a distinctive inflorescence with an extremely long, thick, trailing terminal spike (Figure 1.2 B, C, D). The bracts are relatively short, and the laminas have broader bases. Tepals are broad towards the tips, and strongly recurved. The outer tepals are obovate but the inner tepals are spatulate with a narrow claw and obtuse tipped. The South American *A. caudatus* plants are normally pale or reddish-seeded, but some dark-seeded plants have been found in certain collections (especially those of Asiatic or African origin). Usually flowers are unisexual and male with 5 stamens and female flowers with superior, one-celled ovary with three stigmas (Figure 1.2 E) (Achigan-Dako *et al.* 2014). *Amaranthus caudatus* has frequently crossed with weedy populations of amaranth (e.g. *A. quitensis* in its native area), contributing significantly to complex hybrid weed populations (Sauer 1950; Sauer 1967; Agong 2006; Brenner *et al.* 2010). The plants grow well in cool, dry, high altitude conditions, with an annual rainfall of 500 mm (Council 1984). A single plant may yield up to 50 000 seeds (1 000 seed weight = 0.5 – 1.1 g) but their harvesting is hampered due to inconsistent ripening of the grain (O'Brien and Price 1983; Agong 2006). Leaves can be harvested at any stage during growth, and once grain has matured the whole plant can be harvested and dried. Yields vary substantially between growing regions due to environmental conditions; the seed yields can be as low as 500-800 kg/ha or as much as 2500-4000 kg/ha (Agong 2006).

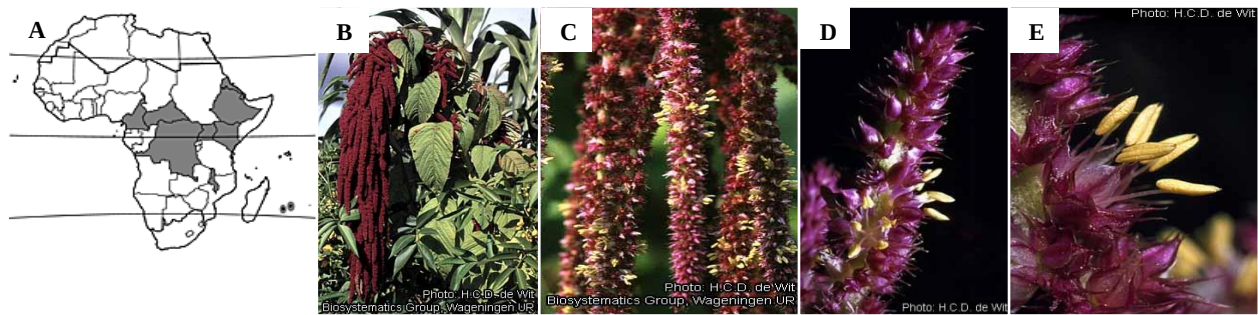


Figure 1.2: *Amaranthus caudatus*. (A) African distribution (B) flowering plant (C) inflorescence (D) detailed inflorescence (E) male flower (Agong 2006).

***Amaranthus cruentus* (2n=32)**

Amaranthus cruentus (purple or red amaranth) arose as a domesticated grain crop in parts of Mexico and Guatemala during the last part of the 19th century and is presently considered a vegetable crop in most countries across Africa (Council 1984). It is the most important leafy vegetable in Sierra Leone, Benin and Togo, and is cultivated to a lesser extent in southern parts of Nigeria, Democratic Republic of Congo, Kenya and Tanzania. Outside of Africa, it is also cultivated in areas of India, Sri Lanka, Bangladesh and the Caribbean. The grain cultivar of *A. cruentus* is grown commercially through Latin America, in Mexico, Guatemala, Colombia and Ecuador. Commercial production of this grain cultivar also occurs in the humid, hot areas of the United States, China and Argentina. In Africa, the grain type is not cultivated apart from a few try-outs in Kenya, Uganda, Ethiopia and Zimbabwe (Figure 1.3 A) (Grubben 1980). *Amaranthus cruentus* has two varieties, with the grain type having paler, white seeds and the vegetable type having darker, black seeds (Council 1984). This is an erect herb (2 m tall), with a thinly to moderately glabrous stem (Figure 1.3 B, C). Leaves are spirally arranged, with a lanceolate to rhomboid-ovate blade, and pinnately veined. Large inflorescences have numerous cymes arranged in axillary and terminal racemes spikes, with the terminal one being up to 45 cm long with many lateral, perpendicular, thin branches (Agong 2006). The flowers are unisexual and fruits an obovoid to rhombic capsule. *Amaranthus cruentus* has occasionally crossed with weedy amaranth populations of *A. powellii*, *A. retroflexus* and *A. hybridus* growing in close proximity, but the offspring were mostly sterile and only survived for a brief period (Sauer 1967; Brenner *et al.* 2010; Achigan-Dako *et al.* 2014). Germination of *A. cruentus* occurs in high light and temperature conditions, where the soil is moist and at temperatures of above 15°C. During one growing

season, the first leaf cuttings can yield up to 2 t/ha (1.0-1.5 kg/m²), while the grain types yield seed of up to 5 t/ha (Grubben 1980).

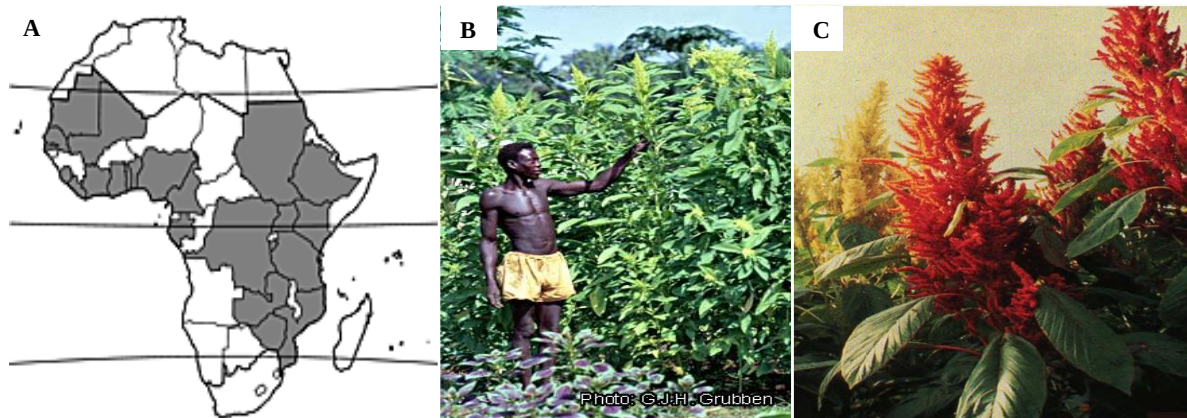


Figure 1.3: *Amaranthus cruentus*. (A) African distribution (B) seeds ready for harvest (Grubben 1980) (C) flowering plant (Council 1984).

***Amaranthus hypochondriacus* (2n=32)**

Amaranthus hypochondriacus was originally domesticated in Central Mexico, but the maximum cultivation today occurs in Central and South America, Asia and Africa (Jansen *et al.* 2004). It appears that the Spanish may have (inadvertently) taken the seed to Europe as indicated by 16th and 17th century herbalists, while in the 17th century, the species spread to Central Europe and Asia, where it was grown as a minor crop (Council 1984). Today, *A. hypochondriacus* is the highest yielding amaranth of the grain types, and is cultivated in several tropical, sub-tropical and temperate regions across the world (Figure 1.4 A). It is found in several countries across Africa, but the exact distribution is unknown (Jansen *et al.* 2004). This annual, erect herb often grows to 2 m tall, and is reddish tinted throughout. The stem is glabrous, and covered in moderately dense multicellular hairs. Leaves are arranged spirally and the blades are broadly lanceolate to rhombic ovate and glabrous. The inflorescence has large, complex, thick branches with numerous agglomerated cymes arranged in axillary and terminal spikes. (Figure 1.4 B, C). The seed ranges from pale to black, depending on the cultivar (Jansen *et al.* 2004). *Amaranthus hypochondriacus* can be grown as a grain crop at very high altitudes, up to 2000 m. *Amaranthus hypochondriacus* can be used in intercropping systems with other food crops and in home gardens. Many farmers harvest the whole plant by uprooting at 20-30 days

after transplanting and the yield is usually 20-25 t/ha for the leaves and approximately 6t/ha seeds (Jansen *et al.* 2004).

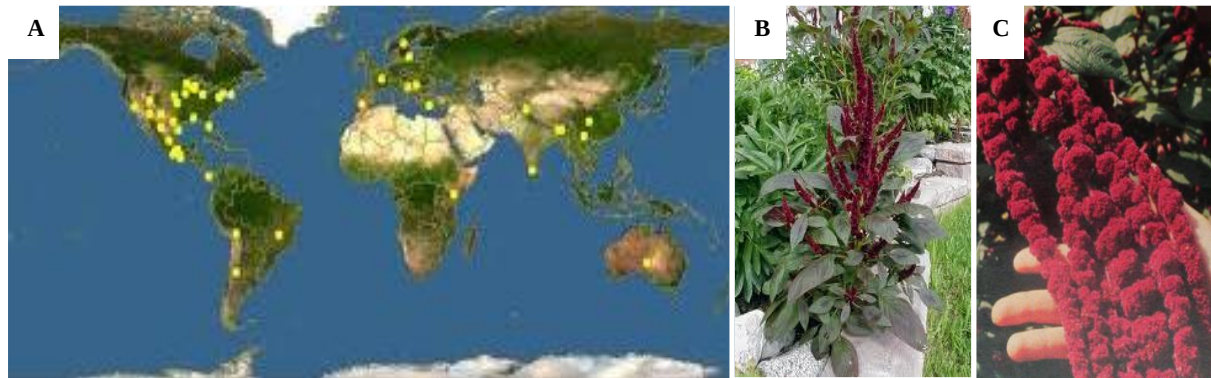


Figure 1.4: *Amaranthus hypochondriacus*. (A) global distribution (B) and (C) flowering plant (Council 1984).

Vegetable amaranths

Amaranth has been a popular leafy vegetable crop across the world, such as Central, South and North America, Southeast Asia and Africa (Venskutonis and Kraujalis 2013). The leaves of some amaranth species have a mild spinach-like flavor, which increase their commercial value as a potherb, or boiled greens crop (Council 1984). Communities, who mainly live in urban or peri-urban societies where adequate fresh fruits, vegetables and animal products are not available, often suffer from malnutrition as the optimal daily requirements of several minerals and vitamins are not ingested. Traditional leafy vegetables aid in the combat of ‘hidden hunger’, as they contribute significantly in supplying several essential micronutrients to the daily diet (Faber *et al.* 2010). The most popular leafy amaranths include *A. cruentus* (purple/red amaranth, Mexican grain amaranth), *A. dubius* (spleen amaranth), *A. hybridus* (smooth amaranth, pigweed), and also species that fall within the genus *Amaranthus* subgenus *Albersia*, namely *A. blitum* (synonym *A. lividis*, purple amaranth), *A. graecizans* and *A. tricolor* (Joseph's-coat). These leafy vegetables are known to have a high nutritional content in their leaves, with elevated protein, vitamin A and C, calcium and phosphorous levels (Venskutonis and Kraujalis 2013). Research has shown that up to 25% of the daily protein requirement can be obtained from consuming amaranth leaves during the harvesting season in several tropic countries. However, the leaves also contain high levels of anti-nutrients, with oxalic acid, cyanogenic

compounds, polyphenols and alkaloids such as betain being reported in several species (Council 1984). Boiling the leaves in water seems to lower the toxicity levels to below dangerous for humans, but may poison livestock when used as fresh animal fodder (Onyesom and Okoh 2007).

***Amaranthus dubius* (2n=64)**

Amaranthus dubius is originally from tropical America and is commonly found in the Caribbean and from northern South America to southern Mexico. The cultivated type of this species may have originated from the weedy type found in tropical Asia (possibly *A. spinosus*), and was introduced into several North American and African countries (Council 1984). It is a cultivated crop and pot-herb in eastern Africa (Kenya, Uganda), central Africa (Cameroon, Democratic Republic of Congo) and in western Africa (Sierra Leone, Ghana, Benin and Nigeria) (Figure 1.5 A) (Grubben 1980). *Amaranthus dubius* is an erect, annual herb (up to 150 cm tall), with branched, glabrous stems (Figure 1.5 B, C). Leaves are arranged spirally, lamina ovate to rhomboid ovate, glabrous and often the center of the lamina is blotched red. The inflorescence is axillary and terminal, with the terminal up to 25 cm long and agglomerated towards the end. The seeds are lenticular, compressed and black in color (Achigan-Dako *et al.* 2014). It is the only polyploid amaranth, possibly from the hybridization of *A. spinosus* with *A. hybridus* or *A. quitensis* (Sauer 1967; Rodríguez *et al.* 2011). *Amaranthus dubius* has a much smaller seed size, with 4000-6000 seeds/g. The first leaf cutting yields roughly 1.5-2.5 kg/m² and subsequent cuttings approximately 1 kg/m² (Grubben 1980). One of the few countries producing this species on a commercial scale is Kenya, but no statistical data on production is available.

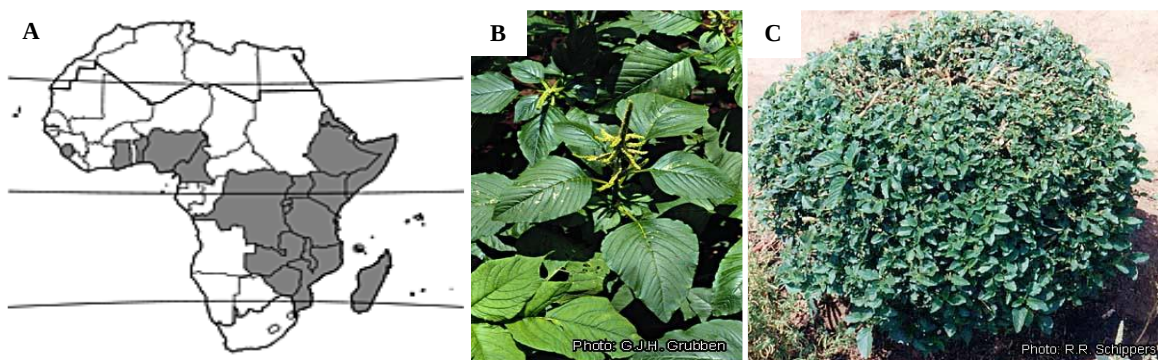


Figure 1.5: *Amaranthus dubius*. (A) African distribution (B) Inflorescence (C) Whole plant (Grubben 1980).

***Amaranthus blitum* (synonym *Amaranthus lividus*) (2n=34)**

Amaranthus blitum is hypothesized to have originated from the Mediterranean region, and had spread to other tropical, humid regions across the world. It is distributed across most of Africa, from Ethiopia to Senegal, South Africa, Cameroon, Kenya and Uganda (Figure 1.6 A) (Grubben 1980). *Amaranthus blitum* is a small, annual herb (growing around 100 cm tall), with simple or branched, glabrous stems. Leaves are arranged spirally, lamina angular ovate, entire, glabrous with a green or purple color (Figure 1.6 B, C). The inflorescence is an axillary cluster, forming a false spike at the apex of the plant. Flowers are unisexual; fruits are sub globular to broadly ovoid capsules, crowned by stigmas. The seeds are lenticular and glossy dark brown to black (Grubben 1980). The seeds of *A. blitum* species are larger than those of other amaranth species, with around 1000 seeds/g. Seeds can germinate after long periods in the soil, and are sown in a mixed cropping system with other indigenous vegetables in certain countries such as Kenya. Commercial growers harvest by uprooting or cutting at ground level, usually around 3-4 weeks after sowing. The uprooted plants can yield up to 1.0 kg/m² fresh leaf material after 4 weeks. Approximately 1.5 kg/m² can further be harvested after the second or third cutting from regenerated plantlets, but the yield obtained for *A. blitum* leaves are generally lower than that obtained for more popular species such as *A. cruentus* or *A. dubius*. In Africa, this crop is a popular vegetable in home gardens but is not cultivated as a commercial crop (Grubben 1980).

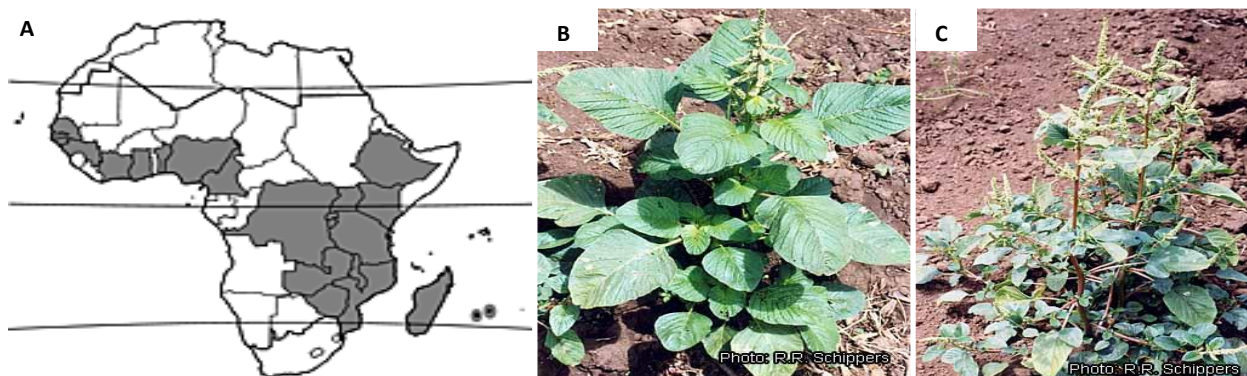


Figure 1.6: *Amaranthus blitum*. (A) African distribution (B) broad leaf form (C) weedy form (Grubben 1980).

***Amaranthus tricolor* (2n=34)**

Amaranthus tricolor originates from tropical Asia, and is considered an important leafy vegetable in south and southeastern Asian countries (Trucco and Tranel 2011). Several varieties of this species are also native to India, the Pacific islands and areas in China (Wu and Corke 1999). *Amaranthus tricolor* occurs as a cultivated vegetable in several African countries such as Benin, Nigeria, Kenya and Tanzania (Figure 1.7 A) (Grubben 1980). The plants are erect, annual herbs with a plant height of about 125 cm and a much branched, glabrous stem. Leaves are simple, arranged spirally, lamina ovate to rhomboid-ovate or broadly elliptical to lanceolate-oblong, very variable in size and glabrous on the lower surface and green to reddish in color (Figure 1.7 B). The inflorescence is an axillary, globose cluster up to 2.5 cm in diameter. The upper clusters sometimes form a terminal spike, with male and female flowers intermixed (Figure 1.7 C). The seeds can be faintly reticulate and shining black or brown (Achigan-Dako *et al.* 2014). The seed weight reaches around 1200-2900 seeds/g, and market seed production fields yield seed amounting to approximately 300-600 kg/ha over five months. Plants can yield up to 1.0-1.5 kg/m² fresh leaf material after 4 weeks of growth. Approximately 1.0-1.5 kg/m² can further be harvested after the second or third cutting from regenerated plantlets. If the crops are continuously cropped through the year, it may yield up to 20 kg/m² product. Very little to no data is available on the commercial production of *A. tricolor* in any country, but in Africa it is grown on a limited scale in home gardens and appears to be of some economic importance (Grubben 1980).



Figure 1.7: *Amaranthus tricolor*. (A) African distribution (B) green to reddish leaf color (C) inflorescence (Jansen *et al.* 2004).

***Amaranthus graecizans* (2n=34)**

Amaranthus graecizans is scattered across several countries in Africa, southern Europe and tropical and subtropical Asia. It is considered a weed in home gardens in Kenya, Uganda, Tanzania, Malawi and southern African countries (Figure 1.8 A) (Grubben 1980). *Amaranthus graecizans* is a small, annual herb growing around 45-70 cm tall. Branches originate from the base of the stem, and are glabrous with short to long, multicellular hairs. Leaves are arranged spirally, simple; lamina is broadly ovate to rhomboid-ovate or narrowly lanceolate, glabrous or with a few short glandular hairs on the lower surface. The inflorescence is an axillary cluster, with male and female flowers intermixed but male flowers more prominent towards the upper parts of the cluster (Figure 1.8 B). Seed are compressed, faintly reticulate and black in color (Costea and DeMason 2001; Costea *et al.* 2003). *Amaranthus graecizans* plants are resistant to unfavorable soil and climatic conditions, and are found in a large variety of different habitats such as waste or cultivated ground, arid areas or flood lands (Costea *et al.* 2003; Sithanantham *et al.* 2004; Maundu 2006). Young, tender shoots and leaves are picked as vegetables, but whole plants can also be uprooted after 4-5 weeks post transplanting. This crop is occasionally sold on small scale at local markets, but no data exists for international trade (Sithanantham *et al.* 2004; Maundu 2006).

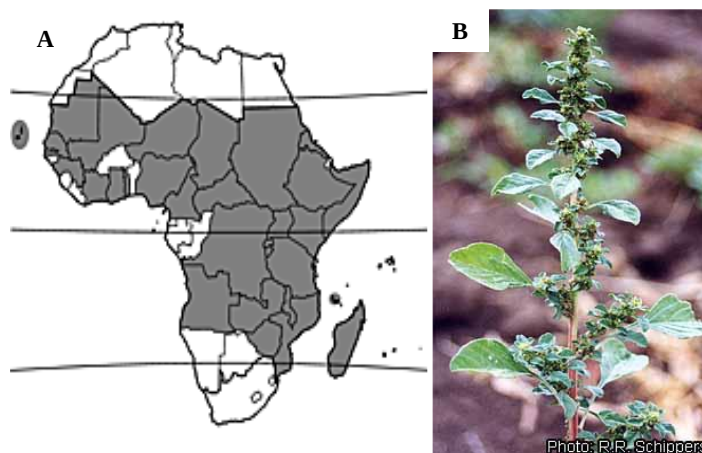


Figure 1.8: *Amaranthus graecizans*. (A) African distribution (B) inflorescence (Grubben 1980).

***Amaranthus hybridus* (2n=32/34)**

Amaranthus hybridus is mostly native to moister, milder regions in eastern North America, Central America, Mexico and through to South America. Recently, it has become a naturalized weed in western North America, Australia, South Africa and Asia. Isolated collections have also been made in Bermuda, Brazil, the Dominican Republic, Algeria, Japan, China and Morocco (Sauer 1967; Weaver and McWilliams 1980). *Amaranthus hybridus* are tall (0.1-2 m) annual plants with a shallow taproot, erect growth habit and green/red stems with branches. Leaves are arranged in an alternate manner, ovate to rhombic-ovate, shiny green to reddish green. Flowers are abundant, small, green and arranged on spikes which form a long, terminal panicle growing from 10 cm to about 25 cm tall (less than 1 cm thick, compared to the 1-1.5 cm of *A. powellii*), with smaller spikes in the leaf axils below. Flowers are unisexual, and each is enclosed by stiff bracts (up to 3-4 mm long), giving the flowers a bristly appearance (less so than for *A. powellii* and *A. retroflexus* due to smaller flowers and shorter bracts). Seeds are round, shiny, dark-brown and somewhat flattened (Weaver and McWilliams 1980; Mosyakin and Robertson 1996; Costea *et al.* 2001b). *Amaranthus hybridus* matures within 11-13 weeks, and produces up to 100 000 seeds per plant. In addition, at maturity, this species had attained a significantly greater plant height, seed weight, leaf, shoot and root weight than other amaranth species (Weaver and McWilliams 1980). *Amaranthus hybridus* mostly occurs along disturbed roadsides, waste sites and orchards and cultivated fields where they are considered a weed, and are not cultivated on a commercial scale, but their potential as a commercial crop due to the highly nutritional herbage and valuable fodder has recently become evident. Currently, they are nurtured in home gardens as pot herbs, dyes and ornamental plants in Zimbabwe, South Africa and parts of India (Weaver and McWilliams 1980). *Amaranthus hybridus* is one of the few species with natural resistance to stem decay and root rot (caused by *Fusarium oxysporum*) in South Africa (Chen and Swart, 2001). Information regarding the resistance of amaranth species to insects and pathogens are relatively scarce, but some defense related genes identified in amaranth include proteases, amylase inhibitors (*A. hypochondriacus*), agglutinins (transgenic cotton that express an agglutinin gene from *A. caudatus* confers resistance to aphids) and cystatins (*A. hypochondriacus*) that inhibit the growth of phytopathogenic fungi (Broekaert *et al.*, 1992; Chagollalopez *et al.*, 1994; Sanchez-Hernandez *et al.*, 2004; Wu *et al.*, 2006; Roy *et al.*, 2006; Valdes-Rodriguez *et al.*, 2010).

Economic Importance

Tolerance to climatic conditions

Plants from the *Amaranthus* genus are well adjusted to sub-optimal soil quality, elevated temperatures and the presence of different pests and diseases. Due to their high genetic diversity and range of phenotypic flexibility, they hold significant potential for further adaptation and development (Rastogi and Shukla 2013; Alemayehu *et al.* 2015). *Amaranthus* species have been found to grow in adverse environmental conditions, most importantly dry, very warm areas where more exotic crops fail. The high adaptability to drought makes them an ideal candidate to explore for cultivation in semi-arid regions (Alemayehu *et al.* 2015). Moisture stress and soil salinity, which are typical of dry, arid regions, adversely affect plants with the traditional C₃ photosynthetic apparatus. *Amaranthus* species are of the few dicot species that utilizes the more efficient C₄ photosynthetic pathway, which is modified to more effectively utilize CO₂ from the air (Kellogg 2013; Sunil *et al.* 2017). The C₄ plants achieve this by concentrating the CO₂ in the chloroplasts of specialized cells, which surround the leaf vascular bundles (Mlakar *et al.*, 2010; Alemayehu *et al.*, 2014). Therefore, carbohydrate production is not compromised through the photorespiratory loss of CO₂ and C₄ plants can convert a higher ratio of atmospheric carbon to plant sugars per unit of water lost. Even when stomata are closed in hot climates to restrict water loss, these plants can maintain high rates of CO₂ fixation (Council 1984). During drought stress in amaranth, it is hypothesized that protein elements help the plant to cope with excessive periods without water. (Huerta-Ocampo *et al.* 2009) found that during water stress, amaranth leaves produce an abundance of chloroplast chaperonin proteins. Chaperonins protect other proteins in the plant cell during folding and in reducing aggregation in stressful conditions. The decreased abundance of rubisco, the cytochrome b6f complex and the enzyme fructose-bisphosphate aldolase during drought stress, indicate that the electron transport chain could be related to the reduction of carbon metabolism (Huerta-Ocampo *et al.* 2009). Soils are often characterized by an excess of organic salts, due to high evaporative water losses during warm conditions (Omamt *et al.* 2006). Such saline soils can affect the growth, yield and dry matter accumulation in plants (Asch *et al.* 2000). In a study by (Omamt *et al.* 2006), the effect of salinity was tested on several amaranth species. Their results indicate that even though reductions in growth parameters was observed, the effect of salinity was less severe on the grain amaranth accessions (A.

hypochondriacus and *A. cruentus*) than their leafy counterparts. For all amaranth genotypes used during the study, water-use efficiency increased with increased salt concentration, probably due to an adaptive mechanism where there is a larger decrease in transpiration than photosynthetic rate (Omamt *et al.* 2006).

Phytoremediation

Globally, agricultural soil is becoming increasingly polluted by inorganic heavy metals and radionuclides due to industrial practices, which cannot be removed from the environment or soil through breakdown by biological processes (Thakur *et al.* 2016). The harmful compounds can spread to and contaminate underground water resources, or be taken up by plants and become hazardous to humans and animals (Alamgir *et al.* 2011). Conventional methods of physio-chemical and engineering techniques utilized in the past have been relatively successful, but their effects on soil properties, the surrounding natural biodiversity and cost associations often makes it an unpractical route for soil remediation (Thakur *et al.* 2016). Phytoremediation using plant species is a practical approach to extracting heavy metals from the soil in a natural way without disturbing the biodiversity of the surrounding areas. *Amaranthus* is proven to be a hyperaccumulator of heavy metals from the soil, extracting the metals through their roots and accumulate them in the shoots and leaves without suffering from phytotoxicity (Huang *et al.* 1998; Chen *et al.* 2000; Alamgir *et al.* 2011). It is therefore assumed that *Amaranthus* species have evolved their hyperaccumulating ability by overexpressing genes that support efficient sequestration of the metals to inactive compartments such as the vacuoles and cell walls, inhibiting potential damage to the photosynthetic apparatus (Rascio and Navari-Izzo 2011).

In the Ukrainian Chernobyl disaster of 1986, a large amount of radionuclides including cesium, plutonium and strontium was released into the surrounding environment. After a few successful cases of phytoremediation with different plants species had been reported for these compounds, it was found that amaranth was especially useful for removing cesium from the soils (Marmioli *et al.* 2006). In Bangladesh, results show that amaranth is a very efficient hyper-accumulator of cadmium and lead from polluted soils (Alamgir *et al.* 2011). In Ohio, USA, a large problem exists with uranium-contaminated soils in industrial areas. After investigating the phytoremediation properties of 30 different plant species, it was found that

three species of *Brassica* as well as *Amaranthus cruentus* demonstrated significant potential to accumulate uranium in their shoots (Huang *et al.* 1998). Two amaranth species (*A. viridis* and *A. caudatus*) have been tested for their ability to extract metals such as chromium, copper, zinc, nickel and chromium from polluted soils in Taiwan, and were found to compare favorably with other species more commonly used (fireweed – *Crassocephalum crepidioides*, black nightshade – *Solanum nigrum* and rape – *Brassica campstris*) (Chen *et al.* 2000). Once such heavy metals have been removed from the soil, the “contaminated” plants can be disposed of and the land could be utilized for agricultural practices.

Role in biofuels

A broad definition of agricultural biomass includes the food-based portion (the simple carbohydrates and oils of crops) and the non-food based portion of crops (complex carbohydrates such as leaves, stalks and cobs of corn, rice husks and straw), as well as perennial grasses and animal waste (Koçar and Civaş 2013). Agricultural biomass is a potential renewable energy resource capable of displacing large amounts of gaseous, liquid or fossil fuels currently utilized as the main energy sources (Demirbas 2009). Besides the production of bioethanol and biofuels, the production of biogas has become a suitable source of alternative energy. Anaerobic digestion of organic wastes (animal manure, blood, fish wastes, stalks and leaves of plants) to produce methane, is possibly a solution to recover waste and produce energy and fertilizers (Koçar and Civaş 2013). *Amaranthus* species have been under scrutiny for possible use in biogas production for several years. Since most *Amaranthus* species have a high yield under marginal conditions, low energy input to produce, high photosynthetic rate, low nutrient requirements, adaptation of tolerance to drought conditions and resistance to pests and diseases and a C₄ photosynthetic pathway, it yields more biomass under sub-optimal conditions and is therefore considered an attractive biomass source (Akond *et al.* 2013). The energy generated from amaranth biomass based biofuel has the potential to reduce CO₂ emissions and decrease the dependency on fossil fuels (Viglasky *et al.* 2009). In a study in the Republic of Moldova, it was found that several herbaceous plants, shrubs and trees (including amaranth) could have the potential to produce biomass of industrial scale to produce biofuel (Timofte *et al.* 2009). (Sitkey *et al.* 2013) reveals the ability of amaranth silage (solely or together with other organic agricultural material) to produce suitable raw material for biogas production. During their research, the authors found that up to 100-200 t/ha of fresh amaranth

biomass can be produced per year, with a minimum dry matter content of 20%. However, the success of amaranth production depends on the choice of species, varieties or hybrids and the authors suggest the use of original seeds obtained from the Andean regions. In addition, amaranth biomass can also contribute as a renewable energy source as a solid fuel with a heating value of 17 MJ/kg, which compares favorably with currently used sources such as mixed household waste (6-10 MJ/kg), wood (6-18 MJ/kg), coal (22.7 MJ/kg) and sugarcane bagasse (15.05 MJ/kg) (Spliethoff 2010); <https://greet.es.anl.gov/>). Two grain amaranth species, *A. caudatus* and *A. cruentus*, shows the highest potential as raw material for biogas production as they have optimum cell wall composition with high cellulose and low lignin content for easy biodegradability (Akond *et al.* 2013; Sitkey *et al.* 2013). *Amaranthus cruentus* produces more normal liter biomethane per kilogram volatile solids (NL/kg VS) than artichoke and sugarbeet, but less than sunflower or sorghum substrates (Mursec *et al.* 2009). However, this study was performed in Slovenia, Europe; and similar measurements of field trials in other geographical regions such as Africa, may lead to different results since amaranth is highly dependent on environmental conditions.

Nutritive value

Low-income populations in developing countries have some of the highest instances of inadequate dietary intake and subsequent disease rate (Popkin 2014). As nutrient dense animal foodstuffs can rarely be afforded on a continuous basis, the high-quality proteins and important minerals iron and zinc are not consumed on a regular basis (Smith *et al.* 2012). Green leafy vegetables contain most essential micronutrients in high levels, and can be locally sourced at market stalls or self-sustaining rural farming posts (Icard-Verniere *et al.* 2015). *Amaranthus* is one of the most important African leafy vegetables, due to their high nutritional content and drought resistance (de la Rosa *et al.* 2009). Grain amaranth cultivation is more common than leafy amaranth farming in Africa, with several farmers growing grain amaranth for local millers and supermarkets in Kenya, Ethiopia, Uganda and Zimbabwe (Achigan-Dako *et al.* 2014). Both the World Vegetable Centre (AVRDC) and the Biodiversity International's African leafy vegetables program has increased awareness among farmers of the superior micronutrient rich leafy amaranth varieties available for agricultural farming (Achigan-Dako *et al.* 2014). *Amaranthus* varieties commonly referred to as leafy

amaranths include *A. tricolor*, *A. cruentus* (which is also considered a grain amaranth), *A. blitum* and *A. hybridus*, while weedy amaranths generally include *A. viridis*, *A. dubius* and *A. blitoides*. The leaves of all the aforementioned species have been shown to contain significant amounts of proteins, β -carotene, iron, calcium, vitamin C and folic acid (Kunyanga *et al.* 2013; Venskutonis and Kraujalis 2013; Achigan-Dako *et al.* 2014). Protein content of leaf samples has been measured as 3.5 – 4.6 g/100g fresh weight (Venskutonis and Kraujalis 2013; Achigan-Dako *et al.* 2014) to approximately 0.74 g/100g dry weight (Icard-Verniere *et al.* 2015). On average, it was found that the protein content for the wild amaranths, and a few leafy amaranths, were up to 2-fold higher than for grain amaranths (Andini *et al.* 2013). Leafy amaranths also have high concentrations of β -carotene, vitamin B6, vitamin C, and essential dietary minerals such as calcium, phosphorous, iron, magnesium, potassium, copper, zinc and manganese (Andini *et al.* 2013; Venskutonis and Kraujalis 2013). Amaranth grains are a rich source of iron, calcium, magnesium, zinc, ascorbic acid, niacin, thiamine and riboflavin (Mnkeni *et al.* 2007; Mlakar *et al.* 2010; Kunyanga *et al.* 2013). In addition, the presence of high concentrations of betalains in the leaves (not anthocyanins, as commonly found in colored leaf species), have strong antioxidant activities (Li *et al.* 2015). However, the presence of these micronutrients and proteins are altered by different cooking methods. Boiling normally lowers the levels of β -carotene, proteins and some minerals, but the antinutritional factors, which may diminish the bioavailability of certain nutrients, are also reduced (Achigan-Dako *et al.* 2014). A very important system to pursue in future research is the potential for breeding grain amaranth crops with the nutritional leafy profile of wild amaranths, to produce a nutritionally enhanced hybrid for domestication.

The proteins found in the grain structures of amaranths are 90% digestible and rich in the essential amino acid lysine (Venskutonis and Kraujalis 2013). Lysine acts as a precursor to produce carnitine, which in turn plays an important role in maintaining acceptable blood cholesterol levels (Tomé and Bos 2007). Lysine is available in significant quantities in animal derived products (eggs, cheese and meat), but deficiencies can occur in human populations that depend mostly on plant-derived products. Since the lysine quantities are extremely low in more conventional grain crop species such as rice (3.8 mg/g protein), maize (2.5 mg/g protein) and wheat (2.3 mg/g protein), amaranth offers a potential lysine content of 55 mg/g total protein without the need for bio-fortification or genetic manipulation of wild type species (Caselato-Sousa and Amaya-Farfan 2012; Tang *et al.* 2013). Amaranth grains are also a good source of tryptophan as well as

sulfur containing amino acids; methionine and cysteine contents ranged from 24-51 mg/g (Mlakar *et al.* 2010; Amare *et al.* 2015). In addition, the proteins in grain amaranths are mainly composed of globulins and albumins (high in lysine and low in glutamic acid) and very little to no prolamin proteins (gluten). Therefore, amaranth grains are an ideal substitute for wheat/barely/rye/oats in gluten intolerant individuals suffering from celiac disease (Council 1984).

The starch granules of grain amaranth seeds mainly consist of amylopectin (93.6-95.2%) and are extremely small compared to starch molecules of other grain types, 1-3 μm compared to 10-25 μm (Venskutonis and Kraujalis 2013). Smaller starch granules generally have better water binding capacity, a lower gelatinization temperature and a higher swelling power. These characters lead to good freeze/thaw capabilities and gelatinization properties, which are important in the commercial food industry (Venskutonis and Kraujalis 2013).

Table 1.1: Comparison of Amaranth with Recommended Daily Allowances (Food and Nutrition Board, Institute of Medicine, National Academies, 2005;

	RDA		Grain Amaranth		Leaf Amaranth	
Folate	400	$\mu\text{g/day}$	53-78	$\mu\text{g/100g}$	85	$\mu\text{g/100g}$
Niacin	15	mg/day	28	mg/kg	0.7	mg/100g
Riboflavin (Vitamin B ₂)	1.2	mg/day	2.4	mg/kg	0.44	mg/100g
Thiamin (Vitamin B ₁)	1.2	mg/day	0.7	mg/100g	0.42	mg/100g
Vitamin A	800	mg/day	0.07	mg/100g	11.3	mg/100g
Vitamin B ₆	1.6	mg/day	4.5	mg/kg	-	-
Vitamin C	80	mg/day	3.86	mg/100g	28	mg/100g
Vitamin E	15	mg/day	63.7	mg/kg	31.43	mg/g
Calcium	1000	mg/day	105	mg/100g	264	mg/100g
Phosphorous	800	mg/day	522	mg/100g	481	mg/100g
Zinc	10	mg/day	3.6-4	mg/100g	15	mg/100g
Magnesium	400	mg/day	253	mg/100g	1320	mg/100g
Manganese	7	mg/day	5.9-6.8	mg/100g	-	-
Iron	8	mg/day	16.8-21	mg/100g	45	mg/100g

Amaranthus species generally contain approximately 73% unsaturated fatty acids; composed of linoleic fatty acids (37-56%), palmitic acid (19-23%), oleic acid (22%), stearic acid (4%) and linolenic acid (0.3-2.2%). Squalene is a naturally occurring polyprenyl compound which plays an important role during cholesterol synthesis, is a natural antioxidant and can act as a “sink” for lipophilic xenobiotic agents in the body helping the release of un-ionized drugs (Kelly 1999). The traditional source of squalene is from whale (*Physeter microcephalus*) and shark (*Centrophorus squamosus*) liver oil, which contain on average 30-45% squalene (Tikekar *et al.* 2008). However, due to diminishing numbers of shark and whale populations, consequently the protection of marine animals has limited the harvesting of squalene from these sources. Amaranth grain

oil has a high squalene content (7-11%) compared to other commercially important oils such as sunflower, rapeseed, rice, corn (0.01-0.4%) and olive oil (~1%) and has been regarded as important in the medical and cosmetic fields due to its anti-oxidant and conditioning properties (He and Corke 2003; Tikekar *et al.* 2008; Mlakar *et al.* 2010).

The recommended daily allowance for important vitamins and minerals, compared to what is found in daily portions of leafy and grain amaranth are indicated in Table 1. Several traditional vegetables (among them, amaranth) have been found to provide 10-50% of the RDA for several minerals and vitamins (Odhav *et al.* 2007), indicating the nutritional importance of these vegetables, especially in countries where malnutrition and disease are common occurrences. Table 2 indicates nutrient levels for specific grain, leafy and wild amaranth compared to more conventional crops.

Table 1.2: Nutritional properties of selected *Amaranthus* species per 100 g edible portion (Grubben 1980; Jansen *et al.* 2004; Agong 2006; Odhav *et al.* 2007; Tikekar *et al.* 2008; Kunyanga *et al.* 2013)

	Water	Energy	Protein	Fat	Carbohydrate	Dietary	Ca	Mg	P	Fe	Zn	β-	Thiamin	Riboflavin	Niacin	Vitamin	Folate	Ascorbic
Grain Amaranth																		
<i>Amaranthus caudatus</i>	9.8	15	14.5	6.5	66.2	15.2	153	266	455	7.6	3.2	*	0.08	0.21	1.29	0.22	49	4.2
<i>Amaranthus cruentus</i>	*	2006	14.7	8.2	74.2	7.6	282	65	148	3.8	0.9	28.5	0.16	0.36	1.1	0.113	22	0
<i>Amaranthus hypochondriacus</i>	84	176	4.6	0.2	8.3	1.8	410	66	103	8.9	0.7	1.7	0.05	0.42	1.2	0.21	85	64
Vegetable and Weedy																		
<i>Amaranthus dubius</i>	84	176	4.6	0.2	8.3	1.8	410	73	103	8.9	1.2	5.7	0.05	0.42	1.2	64	75	54.3
<i>Amaranthus lividus</i>	88.9	75	3.5	0.3	0.3	2.6	270	130	65	3	1.6	1.7	0.07	0.22	0.7	0.23	85	42
<i>Amaranthus hybridus</i>	83	205	6	0.5	6.09	2.81	264	1320	481	45	15	4.2	0.42	0.44	0.7	*	85	62
Maize (<i>Zea mays</i>)	*	1525	10	4.8	73.6	2	13	20.8	219	4.9	0	0.1	0.32	0.12	1.7	0.164	27	4
Wheat (<i>Triticum</i> spp.)	11	1391	12.4	1.8	72.9	2.2	54	23.3	279	6.1	12	0	0.36	0.09	3.8	1.3	281	0
Rice (<i>Oryza sativa</i>)	12	1479	6.2	2	76.4	10.3	27	33.6	312	7.6	1.1	0	0.34	0.05	5.8	0.164	8	0
Potato (<i>Solanum tuberosum</i>)	82	343	1.7	0.1	18.9	0.6	0.1	52	51	1.1	0.3	0.025	0.07	0.03	1.3	0.2	18	21
Spinach (<i>Spinacia oleracea</i>)	*	109	2.1	0.2	5.3	0.8	61	35.2	46	1.7	0.5	5.73	0.3	0.27	0.217	0.06	58	46

Antinutritional factors

Amaranth has been implicated in several studies of livestock poisoning, due to the relatively high nitrate and oxalic acid contents of fresh leaf material (Alegbejo 2013). Plants accumulate nitrates in their leaves due to high nitrogen applications during agricultural practices, or when the process of photosynthesis is slowed down due to herbicide applications, frost or drought conditions. In the gastrointestinal tract of humans and animals, nitrates are converted to nitrites, which interact with the hemoglobin in blood and results in elevated methemoglobin levels that cannot bind to oxygen. The absorbed nitrates can also form carcinogenic nitrosamines (Koh *et al.* 2012). If adequate levels of ascorbic acid are ingested together with the nitrates, the acidic gastric juices can counteract the symptoms of nitrate poisoning by converting the nitrates to nitric oxide, which are absorbed by the stomach lining and metabolized (O'Brien and Price 1983; Mowat and McColl 2001). Oxalate is a naturally occurring anti-nutrient in several leafy vegetables (Libert and Franceschi 1987). Oxalates can bind to minerals such as magnesium and calcium forming insoluble calcium and magnesium oxalate compounds, which cannot be effectively absorbed by the body (Flyman and Afolayan 2006; G  linas and Seguin 2007). However, cooking and leaching of leaves high in oxalate has shown that the levels decrease substantially, to a level where the detrimental effect is eliminated (Noonan and Savage 1999).

In South Africa, the leaves of amaranth species collected across the Eastern Cape Province have been analyzed for antinutritive factors and revealed a nitrate content of 790-1413 mg/100g dry weight. It was concluded from this study that certain amaranth species have higher nitrate content than others, although the exact identities of the accessions were not available (Mnkeni *et al.* 2007). In contrast, the average nitrate content of the more conventional leafy vegetable, spinach, ranges from 470-1342 mg/100g dry weight, compared to lettuce (468-842 mg/150g), kale (402-787 mg/150g), cabbage (478 mg/150g) and chard (648-682 mg/150g) (Umar *et al.*, 2006, Brki   *et al.*, 2017). The total leaf oxalate content of amaranths grown in Canada was approximately 180-250 mg/100g edible portions (G  linas and Seguin 2007), although it can rise to 1270 mg/100g edible portion in species occurring in India (Gupta *et al.* 2005). The oxalate content is comparable to levels found within spinach (750 mg/100g) and beet greens (610 mg/100g) (Sinha and Khare, 2017). In a study performed by (Kariuki *et al.* 2013), grain amaranth was found to contain 204-302 mg/100g phytic acid, 150-210 mg/100g tannins and 178-220 mg/100g oxalate, which is within levels that can be tolerated by the human body. However, if large quantities of fresh leaves are consumed, the levels may become toxic.

In addition, the amounts of trypsin inhibitors, phenolics and saponine contents are fairly low within amaranth grains (Correa *et al.*, 1986; Chávez-Jáuregui *et al.*, 2000). Amaranth grains are commonly cooked, extruded, roasted or popped before being ingested, which also has an influence on the amounts of antinutritional compounds present in the grain. Njoki *et al.* (2017) found reduced amounts of antinutritional compounds after moist heat was applied, 56.1% reduced oxalates, 75.3% less phytates and a slight reduction in tannins (8.7%).

Observed levels of anti-nutritional factors depend on the plant species, age of the plant, the plant tissue and the amount of nitrate in the growth medium/soil (Gupta and Wagle 1988). According to (Grubben 1980), boiling the leaves reduces the amount of nitrates and oxalic acid in the ingested portions, and most of the harmful compounds are discarded in the cooking water. Other studies indicate that consumption of 200 g of cooked amaranth leaves per day has no adverse health effects (Escudero *et al.* 1999).

Amaranthus species also contain saponins, tannins and phytates in the leaves, which are known to influence the bioavailability of micronutrients by forming insoluble complexes (Flyman and Afolayan 2006). Tannin levels are relatively low in amaranth when compared to more conventional grain crops such as sorghum and millet and dehulling of the grains can be used to further reduce the tannin content. Phytates present in the grain can bind to bio-available calcium, iron, zinc and magnesium in the body, rendering them insoluble to the gastric lining of humans and animals.

***Amaranthus* Breeding**

Molecular hybridization is a frequent and important part of plant evolution and speciation, and the study of evolutionary relations and gene pools are essential for breeding programs (Rieseberg *et al.* 1993; Trucco *et al.* 2005). In all plant species, the degree of outcrossing and selfing are important factors to consider during breeding, as the mode and degree of success is mostly dependent on the biology and mating system of the crop (Hauptli and Jain 1985). The crossing potential of *Amaranthus* species is erratic, depending heavily on accession or variety used as maternal or paternal parent, environmental conditions and geographical locations (Hauptli and Jain 1984; Agong and Ayiecho 1991; Trucco *et al.* 2005). According to Brenner (1990), the

three grain species (*Amaranthus caudatus*, *A. hypochondriacus* and *A. cruentus*) have some crossing possibilities with the leafy/wild species *A. dubius*, *A. hybridus*, *A. palmeri*, *A. powellii*, *A. quitensis*, *A. retroflexus* and *A. spinosus*. Although several intra-species hybridization attempts have been made with different amaranth species or varieties (e.g. from different geographical locations), only a few have been successfully commercialized as breeding lines (Pal 1972; Kauffman and Weber 1990; Gupta and Gudu 1991; Trucco *et al.* 2005). Gupta and Gudu (1991) report the successful crossing of *A. hypochondriacus* and *A. caudatus* varieties, where the F₁ seed and seedlings were without any visible deformity and grew to maturity. The F₁ plants had low pollen fertility, but still went on to produce F₂ and even F₃ seeds. This result is in concordance with what was found by (Trucco *et al.* 2007; Trucco and Tranel 2011), but in contrast to the findings of (Murray 1940) and (Pal and Khoshoo 1973). The last mentioned found almost complete sterility in the *A. hypochondriacus*/*A. caudatus* crosses, with no fertile F₁ offspring. Hybridizations between *A. caudatus* and *A. cruentus* were more difficult to obtain and large amounts of pollen had to be used to generate a few F₁, which germinated well without any deformities. However, soon after germination the seedlings showed retarded growth and died before reaching maturity. This result was also observed found when crosses of *A. cruentus* and *A. hypochondriacus* were attempted (Gupta and Gudu 1991). From these results, it could be assumed that *A. hypochondriacus* and *A. caudatus* share a more recent common ancestor, while *A. cruentus* evolved independently and consequently developed unique/alternative hybridizational barriers.

Amaranthus hybridus is hypothesized to be the monophyletic parent of the grain amaranths (*A. caudatus*, *A. cruentus* and *A. hypochondriacus*), leading to the assumption of close genetic relatedness and therefore a high probability of successful crosses between *A. hybridus* and the grain amaranths (Sauer 1967). *Amaranthus caudatus* can be crossed with *A. hybridus* with no observed hybridizational barriers, and normal seed was produced. However, the F₁ does not proceed after the first or second leaf stage, and the plants later died without reaching maturity (Coons 1982). (Greizerstein and Poggio 1994) also reported partial fertile hybrids when *A. caudatus* was crossed with *A. hybridus*, as well as when crossed with *A. quitensis*. In contrast, a cross between *A. hypochondriacus* and *A. hybridus* produced fully fertile F₁, and normal seed germination in the F₂ (Pal and Khoshoo 1973). In fact, a cross between these two species resulted in one of the most widely commercialized species in the US, *A. hypochondriacus* var. Plainsman (crossed with a Pakistani *A. hybridus*) (Trucco *et al.* 2005). Very few hybridization studies have been performed between grain and wild amaranths,

one of the few examples being the cross between *A. hypochondriacus* and *A. powellii*. In their study, (Pal and Khoshoo 1973) found the *A. hypochondriacus/A. powellii* cross to produce completely sterile pollen. However, when crosses were attempted with *A. retroflexus* and *A. caudatus* or *A. spinosus*, the offspring were fully fertile and proceeded to an F₂ generation (Murray 1940).

Barriers to hybridization are difficult to determine for the amaranths due to the (sometimes microscopic) morphological features, but the most common reasons for hybrid sterility seem to include (i) prezygotic reproductive barriers, (ii) the proximal arrangement of the male and female flowers (which promote selfing), (iii) genetic differentiation at the interspecific level due to small structural and genic changes in the chromosomes (iv) weight, stickiness and aerodynamic properties of the pollen and (v) the variable contributions of insects and wind in pollen distributions (Pal and Khoshoo 1973; Hauptli and Jain 1985; Trucco and Tranel 2011). In addition, geographical locations and environmental conditions can have a large influence on the hybridization ability of amaranth species (Coons 1982).

Since there is almost no molecular data available for the different amaranth species, the naturally beneficial traits locked up in their genome cannot be fully exploited. The current study focuses on the most important locally cultivated leafy amaranth in Africa, *A. tricolor*. Several molecular tools developed here will aid future researchers to harness these traits and potentially develop superior genotypes to combat chronic hunger and establish food security in resource poor nations.

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
- Adhikary D, Pratt DB (2015) Morphologic and taxonomic analysis of the weedy and cultivated *Amaranthus hybridus* species complex. *Systematic Biology* 40:604-610
- Afolayan A, Jimoh F (2009) Nutritional quality of some wild leafy vegetables in South Africa. *International Journal of Food Sciences and Nutrition* 60:424-431
- Agong S (2006) *Amaranthus caudatus* L. PROTA (Plant Resources of Tropical Africa/Ressources végétales de l'Afrique tropicale). Backhuys Publishers, Wageningen
- Agong SG, Ayiecho P (1991) The rate of outcrossing in grain amaranths. *Plant Breeding* 107:156-160
- 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 Sciences and Environmental Management* 2:237-240
- Alegbejo JO (2013) Nutritional Value and Utilization of *Amaranthus* (*Amaranthus* spp.)—A Review. *Bayero Journal of Pure and Applied Sciences* 6:136-143
- 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
- Amare E, Mouquet-Rivier C, Servent A, Morel G, Adish A, Haki GD (2015) Protein quality of amaranth grains cultivated in Ethiopia as affected by popping and fermentation. *Food and Nutrition Sciences* 06:38-48 doi:10.4236/fns.2015.61005
- Andini R, Yoshida S, Yoshida Y, Ohsawa R (2013) *Amaranthus* genetic resources in Indonesia: morphological and protein content assessment in comparison with worldwide amaranths. *Genetic Resources and Crop Evolution* 60:2115-2128
- Asch F, Dingkuhn M, Dorffling K (2000) Salinity increases CO₂ assimilation but reduces growth in field-grown, irrigated rice. *Plant and Soil* 218:1-10
- Brenner D (2002) Non-shattering grain amaranth populations. *Trends in New Crops and New Uses*. ASHS Press, Alexandria:104-106
- Brenner D, Baltensperger D, Kulakow P, Lehmann J, Myers R, Slabbert M, Sleugh B (2010) Genetic resources and breeding of *Amaranthus*. *Plant Breeding Reviews*, Volume 19:227-285
- Broekaert WF, Marien W, Terras FRG, Debolle MFC, Proost P, Vandamme J, Dillen L, Claeys M, Rees SB, Vanderleyden J, Cammue BPA (1992) Antimicrobial Peptides from *Amaranthus caudatus* Seeds with

Sequence Homology to the Cysteine Glycine-Rich Domain of Chitin-Binding Proteins. *Biochemistry-U*s 31:4308-4314

- Brkić D, Bošnjir J, Bevardi M, *et al.* (2017) Nitrate in leafy green vegetables and estimated intake. *African Journal of Traditional, Complementary, and Alternative Medicines*. 14:31-41
- Caselato-Sousa VM, Amaya-Farfan J (2012) State of knowledge on amaranth grain: a comprehensive review. *Journal of Food Science* 77:R93-104 doi:10.1111/j.1750-3841.2012.02645.
- 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
- Chagollalopez A, Blancolabra A, Patthy A, Sanchez R, Pongor S (1994) A Novel Alpha-Amylase Inhibitor from Amaranth (*Amaranthus hypochondriacus*) Seeds. *Journal of Biological Chemistry* 269:23675-23680
- Chávez-Jáuregui, Silva MEMP, Arêas JAG (2006) Extrusion cooking process for Amaranth (*Amaranthus caudatus* L.). *Journal of Food Science* 65:1009-1015
- Chen H, Zheng C, Tu C, Shen Z (2000) Chemical methods and phytoremediation of soil contaminated with heavy metals. *Chemosphere* 41:229-234
- Chen W-Q, Swart WJ (2001) The in vitro phytotoxicity of culture filtrates of *Fusarium oxysporum* to five genotypes of *Amaranthus hybridus*. *Euphytica* 127:61-67
- Coons MP (1982) Relationships of *Amaranthus caudatus*. *Economic Botany* 36:129-146
- Correa AD, Jokl L, Carlsson R. (1986) Chemical constituents, in vitro protein digestibility, and presence of antinutritional substances in amaranth grains. *Archivos latinoamericanos de nutricion*, 36:319-326.
- Costea M, DeMason DA (2001) Stem morphology and anatomy in *Amaranthus* L.(Amaranthaceae), taxonomic significance. *Journal of the Torrey Botanical Society*:254-281
- Costea M, Sanders A, Waines G (2001a) Notes on some little known *Amaranthus* taxa (Amaranthaceae) in the United States. *SIDA, Contributions to Botany*:975-992
- Costea M, Sanders A, Waines G (2001b) Preliminary results toward a revision of the *Amaranthus hybridus* species complex (Amaranthaceae). *SIDA, Contributions to Botany*:931-974
- Costea M, Tardif FJ, Brenner DM (2003) The identity of a cultivated *Amaranthus* from Asia and a New Nomenclatural combination. *Economic Botany* 57:646-649
- Costinot A, Donaldson D, Smith C (2016) Evolving comparative advantage and the impact of climate change in agricultural markets: Evidence from 1.7 million fields around the world. *Journal of Political Economy* 124:205-248
- Council NR (1984) Amaranth: modern prospects for an ancient crop vol 47. National Academies
- Crespo R., Wingeyer A, Kruger G, Riggins C, Tranel P and Bernards M. (2017) Multiple-herbicide resistance in a 2,4-D-resistant waterhemp (*Amaranthus tuberculatus*) population from Nebraska. *Weed Science*, 65: 743-754
- 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 doi:10.1007/s10722-011-9684-7

- De la Rosa AB *et al.* (2009) Amaranth (*Amaranthus hypochondriacus*) as an alternative crop for sustainable food production: Phenolic acids and flavonoids with potential impact on its nutraceutical quality. *Journal of Cereal Science* 49:117-121
- Demirbas A (2009) Progress and recent trends in biodiesel fuels. *Energy Conversion and Management* 50:14-34 doi:10.1016/j.enconman.2008.09.001
- Dube T, Moyo P, Ncube M, Nyathi D (2016) The impact of climate change on agro-ecological based livelihoods in Africa: A review. *Journal of Sustainable Development* 9:256-267
- 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
- Edmonds JM, Chweya JA (1997) Black nightshades: *Solanum nigrum* L. and related species Vol 15. Bioversity International
- Escudero N, Albarracin G, Fernández S, De Arellano L, Mucciarelli S (1999) Nutrient and antinutrient composition of *Amaranthus muricatus*. *Plant Foods for Human Nutrition* 54:327-336
- Faber M, Van Jaarsveld P, Wenhold F, Van Rensburg J (2010) African leafy vegetables consumed by households in the Limpopo and KwaZulu-Natal provinces in South Africa. *South African Journal of Clinical Nutrition* 23:30-38
- Flyman MV, Afolayan AJ (2006) The suitability of wild vegetables for alleviating human dietary deficiencies. *South African Journal of Botany* 72:492-497 doi:10.1016/j.sajb.2006.02.003
- Funk C, Dettinger MD, Michaelsen JC, Verdin JP, Brown ME, Barlow M, Hoell A (2008) Warming of the Indian Ocean threatens eastern and southern African food security but could be mitigated by agricultural development. *Proceedings of the National Academy of Sciences* 105:11081-11086
- Gélinas B, Seguin P (2007) Oxalate in grain amaranth. *Journal of Agricultural and Food Chemistry* 55:4789-4794
- Greizerstein E, Poggio L (1994) Karyological studies in grain Amaranths. *Cytologia* 59:25-30
- Group SWPS (2001) A weed plan for Western Australia. Department of Agriculture
- Grubben G (1980) Germplasm of vegetable and grain amaranths. *Plant Genetic Resources Newsletter*:33-35
- Gupta K, Wagle D (1988) Nutritional and antinutritional factors of green leafy vegetables. *Journal of Agricultural and Food Chemistry* 36:472-474
- Gupta S, Jyothi Lakshmi A, Manjunath MN, Prakash J (2005) Analysis of nutrient and antinutrient content of underutilized green leafy vegetables. *LWT - Food Science and Technology* 38:339-345 doi:10.1016/j.lwt.2004.06.012
- Gupta V, Gudu S (1991) Interspecific hybrids and possible phylogenetic relations in grain amaranths. *Euphytica* 52:33-38
- Hauptli H, Jain S (1984) Allozyme variation and evolutionary relationships of grain amaranths (*Amaranthus* spp.). *Theoretical and Applied Genetics* 69:153-165
- Hauptli H, Jain S (1985) Genetic variation in outcrossing rate and correlated floral traits in a population of grain amaranth (*Amaranthus cruentus* L.). *Genetica* 66:21-27

- He H-P, Corke H (2003) Oil and squalene in amaranthus grain and leaf. *Journal of Agricultural and Food Chemistry* 51:7913-7920
- Huang JW, Blaylock MJ, Kapulnik Y, Ensley BD (1998) Phytoremediation of uranium-contaminated soils: role of organic acids in triggering uranium hyperaccumulation in plants. *Environmental Science & Technology* 32:2004-2008
- Huerta-Ocampo José Á, Briones-Cerecero Erika P, Mendoza-Hernández G, De León-Rodríguez A, Barba de la Rosa Ana P (2009) Proteomic Analysis of Amaranth (*Amaranthus hypochondriacus* L.) leaves under drought stress. *International Journal of Plant Sciences* 170:990-998
- Icard-Verniere C, Olive F, Picq C, Mouquet-Rivier C (2015) Contribution of leafy vegetable sauces to dietary iron, zinc, vitamin A and energy requirements in children and their mothers in Burkina Faso. *Plant Foods Hum Nutr* 70:63-70
- Jansen P, Grubben G, Denton O, Messiaen C, Schippers R, Lemmens R, Oyen L (2004) *Amaranthus hypochondriacus* L. *Plant Resources of Tropical Africa* 2:78-80
- Jimoh MO, Afolayan AJ, Lewu FB. (2018) Suitability of *Amaranthus* species for alleviating human dietary deficiencies. *South African Journal of Botany* 115:65-73.
- Kariuki S, Sila D, Kenji G (2013) Nutritional profile of Amaranth grain varieties grown in Kenya. *Food Science and Quality Management* 17:19-24
- Kauffman CS, Weber LE (1990) Grain amaranth. Timber Press, Portland, Oregon
- Kellogg EA (2013) C₄ photosynthesis. *Current Biology* 23:R594-R599
- Kelly GS (1999) Squalene and its potential clinical uses. *Alternative Medicine Review* 4:29-36
- Koçar G, Civaş N (2013) An overview of biofuels from energy crops: Current status and future prospects. *Renewable and Sustainable Energy Reviews* 28:900-916
- Koh E, Charoenprasert S, Mitchell AE (2012) Effect of organic and conventional cropping systems on ascorbic acid, vitamin C, flavonoids, nitrate, and oxalate in 27 varieties of spinach (*Spinacia oleracea* L.). *Journal of Agricultural and Food Chemistry* 60:3144-3150
- Kohrt JR, Sprague CL, Nadakuduti SS, Douches D (2017) Confirmation of a three-way (glyphosate, ALS, and atrazine) herbicide-resistant population of Palmer amaranth (*Amaranthus palmeri*) in Michigan. *Weed Science* 65:327-338
- Kunyanga CN, Imungi JK, Vellingiri V (2013) Nutritional evaluation of indigenous foods with potential food-based solution to alleviate hunger and malnutrition in Kenya. *Journal of Applied Biosciences* 67:5277-5288
- Legleiter TR, Bradley KW (2008) Glyphosate and multiple herbicide resistance in common waterhemp (*Amaranthus rudis*) populations from Missouri. *Weed Science* 56:582-587
- Lesk C, Rowhani P, Ramankutty N (2016) Influence of extreme weather disasters on global crop production. *Nature* 529:84-87
- Li R, Zhai H, Kang C, Liu D, He S, Liu Q (2015) *De novo* transcriptome sequencing of the orange-fleshed sweet potato and analysis of differentially expressed genes related to carotenoid biosynthesis. *International Journal of Genomics* 2015:843802

- Libert B, Franceschi VR (1987) Oxalate in crop plants. *Journal of Agricultural and Food Chemistry* 35:926-938
- Marmioli N, Marmioli M, Maestri E (2006) Phytoremediation and phytotechnologies: A review for the present and the future. *Soil and Water Pollution Monitoring, Protection and Remediation*:403-416
- Maundu P (2006) Promotion of under-utilized food plants in Sub-Saharan Africa: Experiences with African leafy vegetables. International Plant Genetic Resources Institute
- 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* in South Africa. *Water SA* 33:377-380
- Mosyakin SL, Robertson KR (1996) New infrageneric taxa and combinations in *Amaranthus* (Amaranthaceae). *Annales Botanici Fennici* 33:275-281
- Mowat C, McColl KE (2001) Alterations in intragastric nitrite and vitamin C levels during acid inhibitory therapy. *Best Practice & Research Clinical Gastroenterology* 15:523-537
- Murray MJ (1940) The genetics of sex determination in the family Amaranthaceae. *Genetics* 25:409
- Mursec B, Vindis P, Janzekovic M, Brus M, Cus F (2009) Analysis of different substrates for processing into biogas. *Journal of Achievements in Materials and Manufacturing Engineering* 37:652-659
- Nakka S, Godar AS, Wani PS, Thompson CR, Peterson DE, Roelofs J, Jugulam M (2017) Physiological and molecular characterization of hydroxyphenylpyruvate dioxygenase (HPPD)-inhibitor resistance in Palmer amaranth (*Amaranthus palmeri* S. Wats.). *Frontiers in Plant Science* 8:555
- Njoki JW, Sila DN, Onyango AN (2014) Impact of processing techniques on nutritional composition and anti-nutrient content of grain amaranth. In *Scientific Conference Proceedings*.
- Noonan S, Savage G (1999) Oxalate content of foods and its effect on humans. *Asia Pacific journal of clinical nutrition* 8:64-74
- O'Brien GK, Price M (1983) Amaranth: Grain and vegetable types. Educational Concerns for Hunger Organization (ECHO) Technical Note. ECHO, North Fort Myers, FL
- Odhav B, Beekrum S, Akula U, Baijnath H (2007) Preliminary assessment of nutritional value of traditional leafy vegetables in KwaZulu-Natal, South Africa. *Journal of Food Composition and Analysis* 20:430-435
- Oh JE, Kwon Y, Kim JH, Noh H, Hong S-W, Lee H (2011) A dual role for MYB60 in stomatal regulation and root growth of *Arabidopsis thaliana* under drought stress. *Plant Molecular Biology* 77:91-103
- Omamt E, Hammes P, Robbertse P (2006a) Differences in salinity tolerance for growth and water - use efficiency in some amaranth (*Amaranthus* spp.) genotypes. *New Zealand Journal of Crop and Horticultural Science* 34:11-22
- Oniang'o R, Allotey J, Malaba SJ (2004) Contribution of indigenous knowledge and practices in food technology to the attainment of food security in Africa. *Journal of Food Science* 69:CRH87-CRH91
- Onyesom I, Okoh PN (2007) Quantitative analysis of nitrate and nitrite contents in vegetables commonly consumed in Delta State, Nigeria. *British Journal of Nutrition* 96 doi:10.1017/bjn20061915

- Pal M (1972) Evolution and improvement of cultivated amaranths. I. Breeding system and inflorescence structure. *Proceedings of the Indian National Science Academy* 38:28-37
- Pal M, Khoshoo T (1973) Evolution and improvement of cultivated amaranths: VI. Cytogenetic relationships in grain types. *Theoretical and Applied Genetics* 43:242-251
- Popkin BM (2014) Nutrition, agriculture and the global food system in low and middle income countries. *Food Policy* 47:91-96
- Rascio N, Navari-Izzo F (2011) Heavy metal hyperaccumulating plants: how and why do they do it? And what makes them so interesting? *Plant Science* 180:169-181
- Rastogi A, Shukla S (2013) Amaranth: a new millennium crop of nutraceutical values. *Critical Reviews in Food Science and Nutrition* 53:109-125
- Rieseberg L, Ellstrand N, Arnold M (1993) What can molecular and morphological markers tell us about plant hybridization? *Critical Reviews in Plant Sciences* 12:213-241
- Rodríguez P, Pérez E, Romel G, Dufour D (2011) Characterization of the proteins fractions extracted from leaves of *Amaranthus dubius* (*Amaranthus* spp.). *African Journal of Food Science* 5:417-424
- Roy S, Sadhana P, Begum M, Kumar S, Lodha ML, Kapoor HC: Purification, characterization and cloning of antiviral/ribosome inactivating protein from *Amaranthus tricolor* leaves. *Phytochemistry* 2006, 67:1865-1873
- Sanchez-Hernandez C, Martinez-Gallardo N, Guerrero-Rangel A, Valdes-Rodriguez S, Delano-Frier J (2004) Trypsin and alpha-amylase inhibitors are differentially induced in leaves of amaranth (*Amaranthus hypochondriacus*) in response to biotic and abiotic stress. *Physiologia Plantarum* 122:254-264
- Sarr B (2012) Present and future climate change in the semi - arid region of West Africa: a crucial input for practical adaptation in agriculture. *Atmospheric Science Letters* 13:108-112
- Sauer JD (1950) The grain amaranths: a survey of their history and classification. *Annals of the Missouri Botanical Gardens* 37:561-632
- Sauer JD (1967) The grain amaranths and their relatives: a revised taxonomic and geographic survey. *Annals of the Missouri Botanical Gardens* 54:103-137
- Sinha K, Khare V (2017) Review on: Antinutritional factors in vegetable crops. *The Pharma Innovation Journal* 6:353-358
- Sithanantham S, Matoka C, Maundu M, Jakari M, Agong S (2004) Integrated crop protection research for sustainable production of indigenous vegetable crops in Eastern Africa. *Sustainable Horticultural Production in the Tropics* 37
- Sitkey V, Gaduš J, Kliský L, Dudák A (2013) Biogas production from Amaranth biomass. *Acta Regionalia et Environmentalica* 10 doi:10.2478/aree-2013-0013
- Smith FP, Gorrard R, House AP, McIntyre S, Prober SM (2012) Biodiversity and agriculture: production frontiers as a framework for exploring trade-offs and evaluating policy. *Environmental Science & Policy* 23:85-94
- Spliethoff H (2010) Power Generation from Biomass and Waste. *Power Generation from Solid Fuels*:361-467

- 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
- Sunil M, Hariharan N, Dixit S, Choudhary B, Srinivasan S (2017) Differential genomic arrangements in Caryophyllales through deep transcriptome sequencing of *A. hypochondriacus*. *PLoS One* 12:e0180528
- Tadele Z, Assefa K (2012) Increasing food production in Africa by boosting the productivity of understudied crops. *Agronomy* 2:240-283
- Tang M, He X, Luo Y, Ma L, Tang X, Huang K (2013) Nutritional assessment of transgenic lysine - rich maize compared with conventional quality protein maize. *Journal of the Science of Food and Agriculture* 93:1049-1054
- Thakur S, Singh L, Ab Wahid Z, Siddiqui MF, Atnaw SM, Din MFM (2016) Plant-driven removal of heavy metals from soil: uptake, translocation, tolerance mechanism, challenges, and future perspectives. *Environmental Monitoring and Assessment* 188:206
- Tikekar RV, Ludescher RD, Karwe MV (2008) Processing stability of squalene in amaranth and antioxidant potential of amaranth extract. *Journal of Agricultural and Food Chemistry* 56:10675-10678
- Timofte I, Timofte N, Brega V (2009) Development of bioenergy in Moldova. *Problemele Energeticii Regionale* 2:1-12
- Tomé D, Bos C (2007) Lysine requirement through the human life cycle. *The Journal of Nutrition* 137:1642S-1645S
- Trucco F, Tatum T, Rayburn AL, Tranel PJ (2005) Fertility, segregation at a herbicide-resistance locus, and genome structure in BC hybrids from two important weedy *Amaranthus* species. *Molecular Ecology* 14:2717-2728
- Trucco F, Tranel PJ (2011) *Amaranthus*. In: C. Kole (ed) *Wild Crop Relatives: Genomic and Breeding Resources: Vegetables*. Springer, New York pp. 11-21
- Trucco F, Zheng D, Woodyard AJ, Walter JR, Tatum TC, Lane Rayburn A, Tranel PJ (2007) Nonhybrid progeny from crosses of dioecious *Amaranthus*: Implications for gene-flow research. *Weed Science* 55:119-122
- Umar AS, Iqbal M, Abrol YP (2006) Are nitrate concentrations in leafy vegetables within safe limits? *Current Science* 92:355-360
- Valdes-Rodriguez S, Cedro-Tanda A, Aguilar-Hernandez V, Cortes-Onofre E, Blanco-Labra A, Guerrero Rangel A (2010) Recombinant amaranth cystatin (AhCPI) inhibits the growth of phytopathogenic fungi. *Plant Physiology and Biochemistry* 48:469-475.
- 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
- Viglasky J, Andrejcek I, Huska J, Suchomel J (2009) Amaranth (*Amaranthus* L.) is a potential source of raw material for biofuels production. *Agronomy Research* 7:865-873
- Waselkov KE, Boleda AS, Olsen KM (2018) A phylogeny of the genus *Amaranthus* (Amaranthaceae) based on several low copy nuclear loci and chloroplast regions. *Systematic Botany* 43:439-458
- Weaver SE, McWilliams EL (1980) The biology of Canadian weeds: 44. *Amaranthus retroflexus* L., *A. powellii* S. Wats. and *A. hybridus* L. *Canadian Journal of Plant Science* 60:1215-1234
- Wu H, Corke H (1999) Genetic diversity in physical properties of starch from a world collection of *Amaranthus*. *Cereal Chemistry* 76:877-883
- 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

CHAPTER 2

REVIEW ON CROP GENETICS AND IMPROVEMENT WITH EMPHASIS ON ORPHAN CROPS FOR SUSTAINABLE GLOBAL FOOD SECURITY

Abstract

Attaining global food security is a rising concern internationally. Worldwide, a few major domesticated crop species provide most of the calorific and protein intake by humans and animals, but future demands require additional sources of nutrition. Orphan crops are emerging as a valuable source of untapped genetic diversity for nutritional content and natural resistance to biotic and abiotic factors and add immense benefits to urban and rural human livelihoods. However, rapid progress in developing genomic tools for these crops is lagging behind more conventional and domesticated crop species. The advent of affordable, less time-consuming, high throughput genomic techniques such as full genome and transcriptome sequencing, genome editing, high throughput genetic marker development and marker-assisted breeding used together with more conventional plant breeding techniques, have broadened the scope of orphan crop development. This review focuses on genome-aided molecular breeding and transgenic technologies of several orphan crops. Specifically, marker assisted breeding of amaranth; pigeon pea, tef and sweet potato are discussed. Management of genetic diversity in these crops depends on, and is influenced by, several disciplines including environmental factors, availability of natural resources and investment by public and private research sectors. These disciplines interact continuously during development and are dependent on government investment, which may differ between high and low income countries. By combining next generation molecular techniques in the trait development of orphan crops, they could be utilized as a feasible, alternative source contributing to achieving global sustainable food security.

Introduction

Achieving food system sustainability is challenging due to global issues such as population growth, climate change and unstable political environments. During the 1960's, it became apparent that the food-population balance was becoming unstable (Searchinger *et al.* 2014). The average life expectancy of humans increased due to modern medicine and health care, while the amount of arable land remained constant or was reduced to accommodate population growth. Rising food crises prompted influential government officials and scientists around the world to launch the Green Revolution, which included a large-scale increase in crop productivity, global mobilization of resources and a drive to raise awareness in underdeveloped nations (Pingali 2012). Projections of the world's population predict an increase from 7.2 billion in 2014 to approximately 9.6 billion by 2050 (FAO 2015). Although the growth rate is slower than in the past (2.19% in 1963, compared to 1.14% estimated in 2009), the increasing population will nevertheless have a large impact on the earth's ecosystems. It is predicted that the majority of the population growth will be concentrated in developing countries, where food security is already problematic. More than 1 billion people in these countries survive on less than one US\$ per day, resulting in 800 million undernourished adults and more than 200 million underweight children (Tilman 2015). The globalization of agriculture has led to the technical development of only a few commercial plant species to meet global food demands. Of the 10 000 edible plant species in the world, only 150 are produced on a commercial scale. Twelve of these species are responsible for providing 80% of dietary energy and more than 60% of the dietary proteins and calories requirement. The four most important crop species globally are rice, wheat, maize and potato (Mlakar *et al.*, 2010).

Industrialized countries may have contributed to the global climate change phenomena, leading to warmer temperatures, increased frequency and severity of extreme weather events and changes to rainfall patterns (Ziervogel *et al.* 2014). Agricultural development is extremely sensitive to weather variability and shifting micro- and macroclimate environments will force urban and rural farming communities to adopt alternative cultivation techniques. Changes in global temperatures may lead to the increased occurrence of drought and flooding in previously unaffected areas. Species grown on a commercial scale in such areas will ultimately show a decrease in yield due to sub-optimum environmental conditions. If this phenomenon occurs in a number of important agricultural sectors in the world, it can have a significant influence on global crop production (Mlakar *et al.* 2010; Oh *et al.* 2011). An increase of production by almost 70% is required in order

to sustain the estimated 35% population increase projected for 2050 (Ebert 2014).

Stringently managed agricultural environments have led to well adapted domesticated plant species that lack sufficient genetic diversity to efficiently adjust to novel environmental stressors. Current breeding strategies are dependent on the ‘shuffling’ of genes to improve desirable or preferred agricultural traits in crops, a strategy that is failing due to restricted recombination patterns, linkage drag and the predominant inheritance of the same alleles (Kole *et al.* 2015). Untapped genetic diversity needs to be exploited/utilized to increase yields or infer disease resistance in crop species (Kole *et al.* 2015). It is therefore of utmost importance to investigate alternative options for food and nutritional security, whether it be through the cultivation of alternative crops, transgenic-based approaches or molecular breeding strategies.

Role of traditional crops and their wild relatives

The Food and Agriculture Organization (FAO) of the United Nations described food security as “*The right to adequate food (is) a universal human right that is realized when all people have physical and economic access at all times to adequate food or the means for its procurement, without discrimination of any kind*” (FAO 2015). There are four broad dimensions within food security; (i) food availability (ii) ease of accessibility to food (iii) safe and healthy utilization of food and (iv) food system stability. All four of these dimensions are required to efficiently establish food security in a region (Ruane and Sonnino 2010). Due to a lack of diversity within the dietary intake of people in developing nations, many are struggling with noticeable and progressive nutrient deficiencies. In 2011, more than 40% of globally available agricultural land was occupied by only three crops; wheat, rice and maize (Ebert 2014). Micronutrient and vitamin deficiencies or ‘hidden hunger’ is as a result of an imbalanced diet. These deficiencies are pronounced in rural communities where most of the staple foods are carbohydrate based, with little to no supplementation with protein or essential micronutrients (Silva Dias 2014). Dependence on a few non-diverse crop species exacerbates the difficulties associated with food security, health, effective nutrition and ecosystem sustainability. Limited access to vegetable crops leads to deficiencies in vitamins (A, C, B1, B6, B9 and E), minerals, dietary fiber and essential phytochemicals. Consistently imbalanced diets may lead to prominent health concerns such as heart disease, poor gastrointestinal health, impaired vision, diabetes and other chronic diseases. Micronutrient

deficiencies in particular can lead to increased risk of infectious diseases and compromised growth and development in children (Silva Dias 2014).

Neglected, under-utilized and orphan crops are normally regarded as crops that are grown on a small scale by rural farmers for their own food security. These species have undergone little to no scientific research and development. There are, however, thousands of plant species globally that provide edible leaves, grains, nuts, oils, roots and fruits for human consumption thereby supporting community health, livelihoods and environments (Dawson *et al.* 2009; Daru *et al.* 2015). A subset of prioritized orphan crops for nutritional development to aid food security is presented in Table 2.1 (Fox 2013; Arora 2014; Gahukar 2014; Galluzzi and López Noriega 2014).

Table 2.1: Globally prioritized orphan crop species for research and development*

Species	Common name	Continental Development
<i>Abelmoschus caillei</i>	Okra	Africa
<i>Adansonia digitata</i>	Baobab	Sub-Saharan Africa
<i>Allium schoenoprasum</i>	Chives	China
<i>Amaranthus</i> spp.	Amaranth	Sub-Saharan Africa, North and South America, Asia,
<i>Anacardium occidentale</i>	Cashew	Latin America, Africa, India
<i>Aronia melanocarpa</i>	Black chokeberry	Central Asia
<i>Arracacia xanthorrhiza</i>	White carrot	Latin America
<i>Averrhoa carambola</i>	Star fruit	Southeast Asia, India, Australia
<i>Bertholletia excelsa</i>	Brazil nut	South America
<i>Cajanus cajan</i>	Pigeon pea	South Asia, Africa, Latin America
<i>Carica papaya</i>	Papaya	Africa, India, Asia
<i>Celosia argentea</i>	Celosia	Africa, South East Asia
<i>Chenopodium quinoa</i>	Quinoa	India, East Asia, South America
<i>Cicer arietinum</i>	Chickpeas	India, Australia, Western Asia, Middle East
<i>Cichorium intybus</i>	Common chickory	China
<i>Citrullus colocynthis</i>	Bitter cucumber	West Africa, Middle East, Asia, India
<i>Cleome gynandra</i>	Spiderplant	Southern Africa, Asia
<i>Dacryodes edulis</i>	African pear/plum	Central and Western Africa
<i>Dioscorea alata</i>	Yams	Southeast Asia, India, South America, Africa
<i>Eleusine coracana</i>	Finger Millet	Africa, Asia
<i>Eragrostis tef</i>	Ethiopian tef	East Africa
<i>Fagopyrum esculentum</i>	Common buckwheat	South America, Asia
<i>Ipomoea batatas</i>	Sweet potato	Africa, Latin America
<i>Lathyrus sativus</i>	Grass pea	India, East Asia, Pacific
<i>Manihot esculenta</i>	Cassava	Africa, Asia, Latin America, India
<i>Moringa oleifera</i>	Drumstick tree	India, Africa
<i>Opuntia monacantha</i>	Prickly pear	South America, Southern Africa, Australia
<i>Oxalis tuberosa</i>	Oca	Central and South America
<i>Pennisetum glaucum</i>	Pearl Millet	India, Africa
<i>Phaseolus lunatus</i>	Lima bean	India, Central and South America
<i>Plectranthes esculentus</i>	African potato	South America, Africa
<i>Psidium guajava</i>	Guava	Latin America, Africa
<i>Rubus chamaemorus</i>	Blackberry	East Asia, India, South America
<i>Solanum nigrum</i>	Nightshade	Asia, Africa, South America, Mexico
<i>Sorghum bicolor</i>	Sorghum	Africa, India, Asia, Latin America, North America
<i>Vicia faba</i>	Fababean	North and South America, Africa, India, Middle East, South Asia
<i>Vigna subterranea</i>	Bambara groundnut	Southern Africa
<i>Vigna umbellata</i>	Rice bean	India, Africa, East Asia
<i>Vigna unguiculata</i>	Cowpea	Asia, Central and South America, Africa
<i>Ziziphus mauritiana</i>	Jujube	Africa, India, South Asia

*Fox 2013; Arora 2014; Gahukar 2014; Galluzzi and López Noriega 2014

It is also well known that these lesser known crops contain a wealth of genetic diversity and nutritional and health benefits. Orphan crops are a nutritious alternative to conventional crops and provide an important source of income to rural farmers, as well as offer medicinal benefits and ecological sustainability (Jaenicke and Höschle-Zeledon 2006). Many traditional farmers prefer growing vegetable crops that they are familiar with, which have established growing conditions, management strategies and financial value to the farmer among the community members. Crop productivity can be significantly enhanced through the use of traditional plants in crop rotation systems, as they often inhibit the build-up of pests and diseases, which can lead to devastating crop losses. Their use as ground cover may prevent soil erosion, reduce the growth of weeds and loss of water via evaporation from the soil (Oniang'o *et al.* 2004). Orphan crops aid in providing

food security during the off-season when conventional crops such as maize, rice and green leafy vegetables are not cultivated. Commonly referred to as ‘poor-man’ crops, these under-utilized vegetables are only considered viable food crop in the absence of other food sources (Cloete and Idsardi 2013).

Important factors in orphan crops research

Despite the positive effects orphan crops are likely to have on future food security, a number of challenges remain. One of the major constraints affecting the success of newly commercialized orphan crop species is their adoption in mainstream food supply chains. The interaction of genotypes (amount of genetic diversity within a genus, hybridization and domestication events), environmental pressures (population growth, climate change) and management applications (investment of funds, labor and scientific training) will work interchangeably to drive the commercial development of neglected and underutilized crop species (Figure 2.1). Balancing and managing these interchanges would ultimately determine the success with which the crop is integrated into main commercial supply chains.

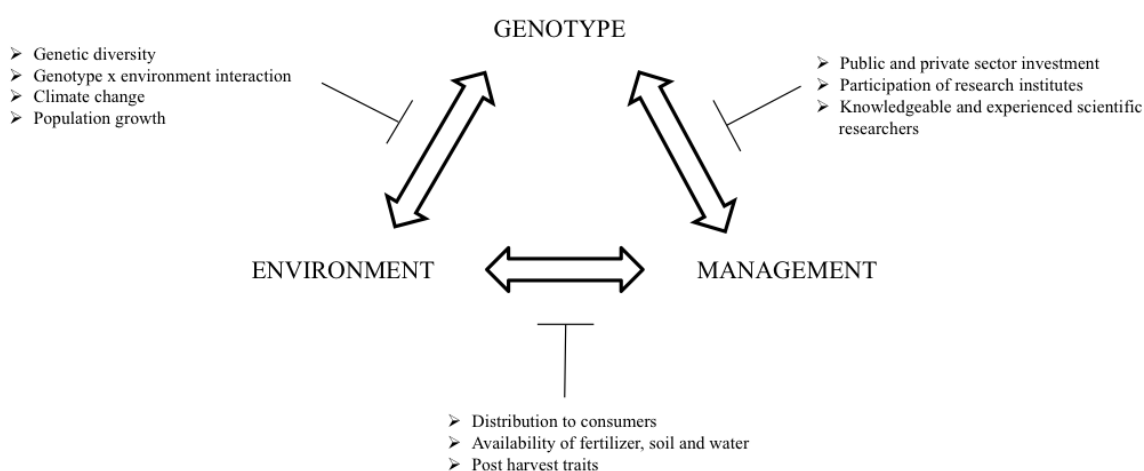


Figure 2.1: Interacting strategies between crop genotypes, environmental influences and application of efficient management strategies. Bulleted points indicate possible constraints on the interaction between disciplines.

Orphan crops have to compete with well-known conventional crop species that are historically popular with consumers. In addition, orphan crops lack the breadth of scientific research that traditionally supports

commercial crops. Another factor to consider during orphan crop commercialization relates to the market value of the crop. For example, to establish if productivity is sustainable by farmers to meet future demand and predicting whether the crops will be sold at competitive retail prices in order to be sufficiently profitable to support farming communities. Initial research, development, agricultural constraints and scientific interventions may result in increased production costs, implying that the final retail prices are unaffordable to most consumers.

Despite orphan crops being locally adapted to their environmental conditions, it is unclear if they are likely to succeed in novel environments outside of their native range, emphasizing the need to conduct thorough research on these species. A lack of knowledge pertaining to genetic variation in these taxa, difficulties in propagation, limited access to diverse germplasm, high post-harvest losses, low yields and susceptibility to strange pests and diseases (outside their native environment) could lead to significant crop losses prior to the products reaching marketable stage (Dawson *et al.* 2009). However, effectively utilizing the high levels of genetic diversity during plant breeding with the support of sufficient capital investment by national and international public sectors will enable the commercial development and release of orphan crops to the retail industry.

Traditional plant breeding

Traditionally, plant breeders have crossed, or hybridized plants based on superior or preferred traits, followed by screening of progeny to choose hybrids exhibiting the enhanced traits. The genetic make-up of the parents and the progeny are not considered by farmers as they typically do not have the expertise to genetically screen for insufficient recombinations, linkage drag or promising gene combinations (Searchinger *et al.* 2014). Given that many crop species have extended maturation cycles and that the integration of traits has to be tested in concurrent generations of progeny, fixing the preferred traits in crop populations are often time and resource intensive processes. Domestication events of leading crop species such as maize, wheat and soybean have diminished the genetic diversity held within their genomes (Mir *et al.* 2013; Lopes *et al.* 2015; Zhou *et al.* 2015b). This negatively affects their resistance to pests when growing outside their ideal environments, as well as their ability to adapt to atmospheric and resource fluctuations.

Many rural farming communities do not have access to commercial crop varieties that have been enhanced through conventional breeding or genetic modification (Azadi *et al.* 2015). Consequently, a number of farmers endeavor to select their own varieties of plants and often focus on traditional, underutilized or neglected species (Mavengahama *et al.* 2013). Rural farmers view the following traits as the most important: total yield, height, resistance to biotic and abiotic stresses and uniformity of growth. Not only do farmers prefer certain traits, but consumers are demanding better quality produce (especially leafy, root and grain species) with enhanced appearance, flavor, shelf life and nutritional quality (Khush 2001; Silva Dias 2014).

The main disadvantages of conventional breeding include the effort of obtaining genetically diverse lines available for certain conventional crops (narrow germplasm sets); the need for trained staff to perform crosses between species that are difficult to hybridize; long time periods to grow and screen progeny over many generations; lack of high throughput phenotyping technologies to consistently score traits; the difficulty in evaluating the effects of variable gene recombination and segregation; and the cost and effort of glasshouse or field maintenance during progeny screening. The use of conventional breeding techniques can be supplemented with knowledge obtained through molecular breeding and can possibly accelerate the plant breeding process.

Genomics-assisted plant breeding

The development of molecular techniques has gained increased investment especially in developed countries, for the description and manipulation of plant genomes. These innovations have enabled researchers to investigate gene structures and biochemical pathways, which govern important traits such as nutrition, yield, biotic resistance and abiotic tolerance (Murphy 2015). The two major molecular approaches commonly used in trait development are transgenic technologies and marker assisted breeding.

Transgenic Technologies

The concept of transgenic plants began in the 1980's, with the insertion of useful genes from multiple sources

into plant genomes, utilizing techniques such as particle bombardment, electroporation and *Agrobacterium tumefaciens* transformation (Prado *et al.* 2014). The term ‘agricultural biotechnology’ arose with the successful transfer of DNA from one species to another using *A. tumefaciens* as a vector for transformation (Chilton *et al.* 1982; Caplan *et al.* 1983; Fraley *et al.* 1985). Crops genomes can be altered to have increased yields (Yamori *et al.* 2016), tolerance to adverse climatic conditions (Rahman *et al.* 2016) and resistance to economically important diseases (Guo *et al.* 2016); which in turn aids significantly in the fight for global food security (www.worldfoodprize.org). Traditional plant breeding involves sexual crosses of whole genomes to produce random mixing of genes in the offspring, while inserting one or more pre-defined genes produces transgenic plants. In recent years, direct genome editing has become a feasible approach to alter the genetic profile of organisms (Khatodia *et al.* 2016). Targeted gene inactivation conclusively provides information about the functions of genes and can be exploited by gene modifying systems such as zinc finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs) (Gaj *et al.* 2013). Synthetic nucleases can be engineered to target any nucleotide pattern in the genome and together with the substantial progress made in improving and observing genomic modification, it is possible to alter genes in favor of a specific trait (Perez-Pinera *et al.* 2012). Another biological system exploited for genome editing is the CRISPR/Cas (clustered regularly interspaced short palindromic repeats-CRISPR associated proteins) structure, naturally encoded by prokaryotic cells for immune protection against invading threats (Cong *et al.* 2013). Particularly, researchers are using the Cas9 endonuclease to focus on specific genomic sequences using a chimeric guide RNA, which will recognize a target motif in the endogenous DNA. Precise modification, insertions, deletions and mutations of the target sequence can be performed (Makarova *et al.* 2011). As this system only introduces small nucleotide base modifications (through the Cas9 and RNA structural components) and not foreign full-length gene regions, it is unlikely that the resulting genotypes will be regarded as true transgenics (Khatodia *et al.* 2016). However, there may still be new challenges in marketing, acceptance and regulation of genome edited (GE) crops. These new, enhanced plant varieties could be commercialized at a much faster rate, because the set of strict regulations pertaining to transgenics would not completely apply (Hefferon 2015; Kole *et al.* 2015). To date, the CRISPR-cas genome-editing system has been applied to many different species groups, both in commercial species such as wheat (Shan *et al.* 2013), rice (Miao *et al.*, 2013), and maize (Liang *et al.* 2014); and in orphan crops including *Medicago trunculata* (Michno *et al.* 2015) and sweet orange (Jia and Wang 2014).

The ability to specifically transform or edit genes related to resistance, nutritional traits or yield has significantly impacted agricultural development in many different plant species, especially maize, soybean and cotton (Searchinger *et al.* 2014). The first transformation event into tobacco using *A. tumefaciens* was performed by (Barton *et al.* 1983), but has since developed into one of the most important tools for crop improvement.

In 2014 it was reported that 18 million farmers across 28 countries have commercialized and are growing more than 180 million hectares of genetically modified (GM) crops. The majority of agricultural land across the world is planted with herbicide tolerant and insect resistant soybean (48%), followed by maize (33%), cotton (24%) and canola (5%). These crops are also marketed with stacked traits of herbicide tolerance (different types of herbicide) and insect resistance (to different pest species) and has seen an 8% increase in planting from the year 2012 to 2013 (James 2015). Other traits and crops, which have been genetically modified and sold commercially are: (i) increased abiotic stress tolerance (maize, soybean, sugarcane); (ii) altered growth/yield (Eucalyptus, maize, soybean); (iii) disease resistance (bean, papaya, plum, potato, squash, sweet pepper, tomato); (iv) herbicide tolerance (alfalfa, canola, chickory, cotton, flax, maize, canola, potato, rice, soybean, sugar beet, tobacco, wheat); (v) insect resistance (cotton, eggplant, maize, poplar, potato, rice, soybean, tomato); (vi) modified product quality (alfalfa, apple, canola, maize, melon, potato, rice, soybean, tobacco, tomato) and (vii) pollination control/male sterility/fertility restoration systems (canola, chickory, maize) (James 2015).

Genetic modification methods have been optimized in many laboratories around the world and therefore a skilled team can relatively easily identify and transfer genes for traits such as nutritional improvement, herbicide tolerance or abiotic stress resistance into a crop of interest, including orphan crops (Chetty *et al.* 2013). However, several technical challenges exist during the development of transgenic orphan crops, including the need to develop protocols for new plant species from scratch, the problem of genotype specificity (i.e. that only certain genotypes of a species can be transformed), availability of effective transformation techniques, sufficient funding and scientific capacity to perform the experiments and governmental and legislature regulations in different countries to adopt the commercialization of transgenic crops (Naylor *et al.* 2004). Several research studies have produced transgenic orphan crop lines, but few have translated beyond the laboratory and field assessments to be fully commercialized (Table 2.2).

Table 2.2: Examples of transgenic orphan/neglected crop species

Orphan crop	Transgenic trait	Reference
Amaranth (<i>Amaranthus tricolor</i>)	Antibiotic resistant	(Jofre-Garfias <i>et al.</i> 1997)
Adzuki bean (<i>Vigna angularis</i>)	Resistant to soybean mosaic virus	(Chen <i>et al.</i> 2013)
Cassava (<i>Manihot esculenta</i>)	Resistance to brown streak virus	(Yadav <i>et al.</i> 2011)
	Increased vitamin B6 bioavailability	(Li <i>et al.</i> 2015a)
	Accumulation of iron in roots	(Narayanan <i>et al.</i> 2015)
Cowpea (<i>Vigna unguiculata</i>)	Resistance to beetles	(Solleti <i>et al.</i> 2008)
	Herbicide resistance	(Citadin <i>et al.</i> 2013)
Chickpeas (<i>Cicer arietinum</i>)	Resistance to bruchid beetles	(Ehlers <i>et al.</i> 2013)
	Increased drought tolerance	(Hajyzadeh <i>et al.</i> 2015)
	Enhanced salt tolerance	(Ghanti <i>et al.</i> 2011)
Eggplant (<i>Solanum melongena</i>)	Resistance to root and shoot borer	(Rai <i>et al.</i> 2013)
	Fungal resistance	(Darwish <i>et al.</i> 2014)
	Enhanced chilling tolerance	(Wan <i>et al.</i> 2014)
Faba bean (<i>Vicia faba</i>)	Salt and drought tolerance	(Hanafy <i>et al.</i> 2013)
Finger millet (<i>Eleusine coracana</i>)	Resistance to leaf blast disease	(Ignacimuthu and Ceasar 2012)
	Salt tolerance	(Anjaneyulu <i>et al.</i> 2014)
	Herbicide resistance	(Bayer <i>et al.</i> 2014)
Papaya (<i>Carica papaya</i>)	Resistance to papaya ringspot virus**	(Gonsalves 1998)
	Delayed ripening trait	(Cabanos <i>et al.</i> 2014)
Pigeon pea (<i>Cajanus cajan</i>)	Insect resistance	(Ramu <i>et al.</i> 2012)
	Increased proline accumulation and salt tolerance	(Surekha <i>et al.</i> 2014)
Sorghum (<i>Sorghum bicolor</i>)	Resistance to stem borer	(Visarada <i>et al.</i> 2014)
	Tolerance to anthracnose	(AKosambo-Ayoo <i>et al.</i> 2013)
Sweet potato (<i>Ipomoea batatas</i>)	Increased starch quality and content	(Zhou <i>et al.</i> 2015a)
	Resistance to black rot disease	(Muramoto <i>et al.</i> 2012)
	Tolerance to various abiotic stresses	(Fan <i>et al.</i> 2012)
Tef (<i>Eragrostis tef</i>)	In vitro regeneration	(Gugsa and Kumlehn 2011)
	Induce semi-dwarfism	(Gebre <i>et al.</i> 2013)

Although transgenic technologies show great potential for contributing to sustainable food security, some obstacles remain. Public concern based on perceived risks with regard to food and animal feed safety, ecological impact as well as socio-economic factors has held many governments back in establishing biosafety regulatory frameworks or approving commercial release (Whitty *et al.* 2013). The majority of research and development of agro-biotechnology has been limited to developed countries. However, in countries such as South Africa, local seed companies have licensed the main GM events such as Bt insect resistant maize or herbicide tolerant soybean. The genes responsible for resistance have been introgressed into locally adapted germplasm by conventional breeding combined with DNA markers for the transgene (Carstens and Berger 2016). This model has been followed for high value field crops in South America and Asia; however, applying GM technology to orphan crops with small markets is often the responsibility of local governments. Developing countries usually have limited financial resources to support advanced scientific research and therefore fail to develop their own GM technology (Azadi *et al.* 2015). Socio-economic concerns also need to be addressed, such as the effect on traditional cultivation systems. Building

** Commercialized in Hawaii, 1998

of North-South partnerships with a strong emphasis on capacity building offers one solution to implementation of GM crops in developing countries.

Genomic Resources and Trait Mapping

Molecular markers

Genetic markers are heritable polymorphisms within a gene or intergenic region at a known location on a chromosome, that could be used to identify individuals or traits of interest (Utomo *et al.* 2012). They are abundant throughout the genome, making them popular to elucidate population structure, germplasm diversity and allelic richness in orphan and commercial crops (Govindaraj *et al.* 2015). The majority of the markers are in the non-coding regions of nuclear DNA, due to errors in replication of tandem repeats, point mutations or insertions and deletions, whereas the coding regions are under much stronger selection pressure to remain constant. Environmental fluctuations and the developmental stage of a plant do not affect the presence of these markers, since they are genetically heritable (Govindaraj *et al.* 2015). Measurements of marker patterns and phenotypic traits enable researchers to form an association between genotype and phenotype. Once a marker-trait association has been established, researchers can screen progeny early in the growth cycle (e.g. sub-sample of seed or seedling) for the DNA marker using a rapid, cost effective molecular technique (e.g. polymerase chain reaction (PCR)) as a representation for the trait. This saves resources by discarding negative plants early in the season and not having to wait till they grow to maturity before scoring the trait (Searchinger *et al.* 2014). It also decreases the amount of samples for downstream analysis of yield or nutritional content, which may require expensive techniques (Tadele and Assefa 2012). Unprecedented growth in the field of biology resulted from molecular technologies becoming more accessible and cheaper to perform (Searchinger *et al.* 2014).

Different marker technologies exist, from conventional hybridization-based restriction fragment length polymorphism (RFLP), PCR based random amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP) and simple sequence repeats (SSRs) technologies to more recent sequence based single nucleotide polymorphisms (SNPs). The choice of marker technology depends on the research objectives, as well as the technical procedure for assays, reliability, level of polymorphism and the cost involved. RFLPs,

RAPDs and AFLPs are becoming redundant mainly due to their expensive and time-consuming protocols, their limited ability to distinguish between closely related lines and their low reproducibility. The advent of next generation sequencing (NGS) technologies and its high throughput application towards SSR and SNP markers has revolutionized the application of marker genotyping (Gimode *et al.* 2016; Vartia *et al.* 2016). Instead of randomly amplifying segments of DNA and possibly detecting polymorphic markers for species, it is now possible to sequence entire chromosomes or genomes to search for not hundreds, but thousands to ten thousands of markers simultaneously (Grover and Sharma 2016).

Next generation sequencing

First generation sequencing involved the chain termination method, established by Sanger and Coulson in 1975 (Niedringhaus *et al.* 2011). Sanger sequencing has been the preferred method utilized during sequence based projects, due to the accuracy, long read length and ability to resolve homopolymer regions and repeats (Morozova and Marra 2008; Niedringhaus *et al.* 2011). However, low throughput and high costs associated with sample preparation necessitated the development of NGS. Entire genomes can now be sequenced from minimal starting material in only a few days or hours. New methods stemming from NGS include transcriptome sequencing (RNA-seq), epigenetic sequencing (methyl-seq) and chromatin immunoprecipitation followed by sequencing (ChIP-seq). These technologies are all contributing to the ultimate understanding of cellular processes and the complex interactions between DNA, RNA and protein molecules in plants, animals and humans (Treangen and Salzberg 2012; Omboki *et al.* 2015).

Due to the significant decrease in the cost of sequencing, determining the complete genome of an individual has become a more realistic approach for researchers, even with moderate budget constraints (Varshney 2015). Complete sequencing, assembly and annotation of the initial genome of a species provide the platform for subsequent functional characterization of gene regions using bioinformatics resources. These include the rapid identification of structural variants in terms of markers (SNPs), copy number variants (CNVs), genome duplication events and chromosomal rearrangements (Omboki *et al.* 2015). Genomic resources are essential for genomics-enabled molecular breeding of crops with desirable traits (Varshney *et al.* 2009). High throughput sequencing has revolutionized plant breeding by revealing syntenic properties between species groups. Current evidence indicates that plant genomes share gene content, chromosomal organization and

biochemical pathways (Naylor *et al.* 2004). Several ‘model’ species (so named due to their popularity in cultivation or ease of study), have been sequenced at great depth and syntenic studies with their ‘orphan’ brothers led to accelerated research projects related to trait development.

Trait Mapping

Utilizing molecular marker data during the breeding process is referred to as marker assisted selection (MAS). The concept of MAS is based on the linkage disequilibrium between markers and genes, or the proximity of the markers towards genes responsible for phenotypic trait variation (Muranty *et al.* 2014). The first step in MAS is to identify and characterize putative parental accessions for a biological trait of interest (Figure 2.2 A). It is important to produce a mapping population where the trait of interest is segregated into the progeny. The majority of mapping populations are F₂s, double haploids, backcrosses, recombinant inbred lines and near isogenic lines (Figure 2.2 B) (Moose and Mumm 2008; Davey *et al.* 2011). Due to high throughput technologies (whole genome sequencing, identification of abundant molecular markers, high-density genetic maps, contig-based physical maps) being applied to several crop species, MAS is an essential technique to accelerate precision plant breeding (Figure 2.2 C).

Technologies utilizing NGS and genomic tool development (markers and gene annotations) have reached unprecedented growth in the last decade, compared to trait screening and phenotyping technologies. Based on the demand for the development of the latter technologies, researchers have developed high throughput phenotyping tools using computer vision-assisted analysis tools and robotic-assisted imaging platforms (Figure 2.2 D) (Fahlgren *et al.* 2015). Coupling genomic data with phenotypic data will enable the production of linkage maps to identify quantitative trait loci (QTL), which represent a statistical association between the DNA and a variation observed in phenotype (quantitative trait). Usually QTL will contain genes responsible for the control of the trait variation. A large number of markers in a segregating population will lead to the identification of highly informative recombination maps (genetic maps) and therefore accurate QTL predictions. This will enable researchers to select markers associated with traits through the population (Figure 2.2 E).

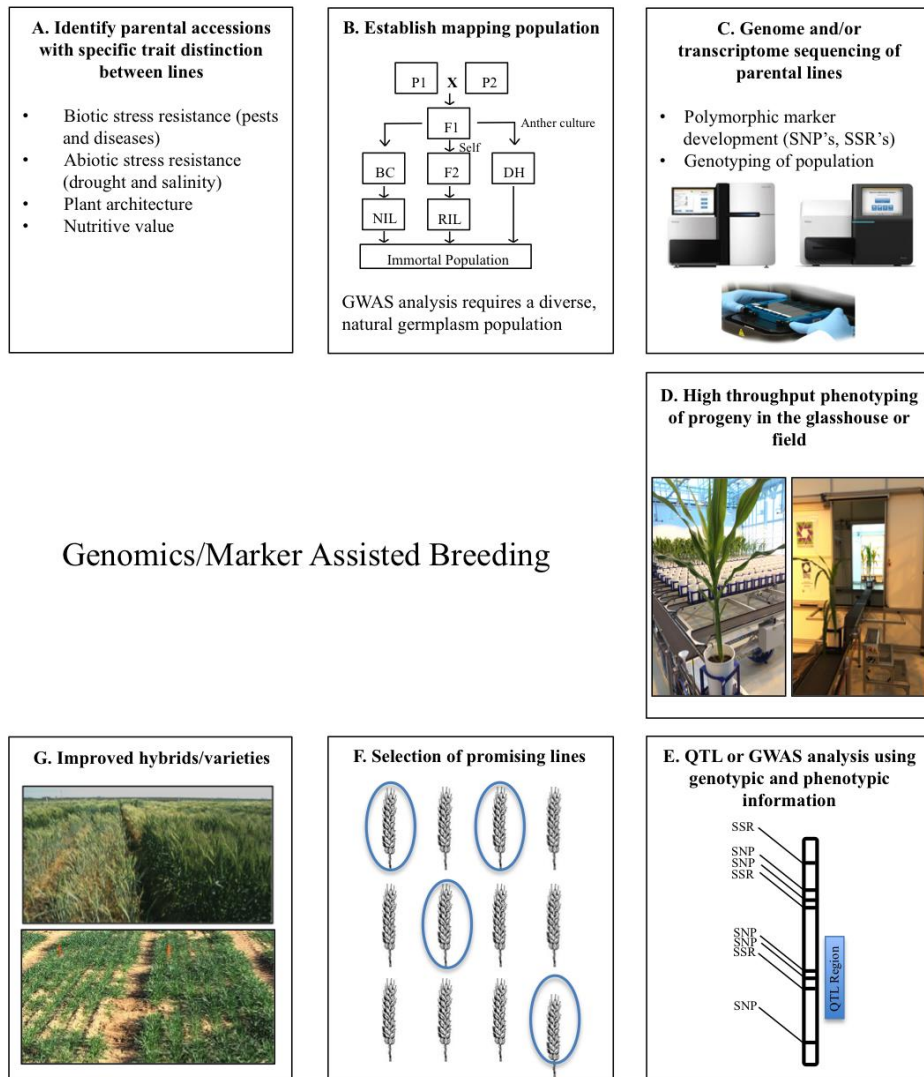


Figure 2.2: Simplified diagram of genomics/marker-assisted selection steps. (1) Identify parental lines which show clear trait differences (2) Produce crosses for applicable mapping population (3) Perform precise phenotyping of the progeny for the trait of interest (4) Once the genotypic profile of the parental lines are known, high throughput genotyping is performed on the progeny to monitor genetic segregation of markers (Images: www.illumina.com) (5) Statistical analysis is performed using the genotypic and phenotypic data, to identify possible QTL regions explaining the trait variation (6) Lines showing promising traits are chosen for further development (7) The production and commercialization of improved varieties or hybrids (Images: www.nytimes.com; www.phys.org)

High throughput DNA sequencing and high-density SNP arrays have resulted in an immense amount of genetic data for studies in both eukaryotic and prokaryotic systems. The combination and relationship of genotype and phenotype data can be used to predict important traits within a species complex (Korte and

Farlow 2013). Genome wide association studies (GWAS) evaluate the relationship between each genotypic marker and a specific phenotype in a large population, to determine if any variant is associated with that trait (Evangelou and Ioannidis 2013). These associations can be combined with genetic maps and QTL analyses results to accurately pinpoint the cause of variation of a trait between individuals or populations. In rice, GWAS has been used to find 10 candidate genes related to plant architecture traits; a finding that was supported additional overlapping QTLs previously identified. The authors were therefore able to propose a genetic model for panicle traits, as well as identify links between yield performance and panicle size (Crowell *et al.* 2016). Similar important findings have been reported for (among others) soybean (Sonah *et al.* 2015; Chang *et al.* 2016), tomato (Sauvage *et al.* 2014), maize (Chen *et al.* 2015) and banana (Sardos *et al.* 2016).

Once information is available pertaining to the gene-marker association of an agronomic trait, MAS and/or Genomics Assisted Breeding (GAB) can be used to introgress or pyramid genes of interest for elite cultivar development. In contrast to the long growth period required during conventional breeding, breeders can now predict whether seedlings contain the necessary genes to produce a multiple trait breeding line through a combination of DNA map and statistical analysis (Figure 2.2 F and G) (Searchinger *et al.* 2014).

Molecular breeding in orphan crops

In the past, scarce genomic information in orphan crops necessitated the need for conventional breeding techniques (natural crosses, domestication) to exploit their natural diversity. With the development of advanced molecular techniques, routine screening of genetic markers in a breeding population resulted in precision breeding for agricultural traits. Several studies have been performed in different orphan crop species in recent years. A selection of significant research outputs pertaining to four commercially important African orphan crops is discussed in the following section, while an in-depth list is presented in Table 2.3.

Amaranthus

Amaranthus species (amaranths) are an emerging and promising nutritious traditional leafy and grain dicot vegetable food source in Africa, India and South America (Achigan-Dako *et al.* 2014). The leaves of these

crops are high in proteins, vitamin A precursors and iron, while the grains contain significant amounts of essential oils and important scarce amino acids such as methionine and lysine (Rastogi and Shukla 2013). They thrive under extreme abiotic stresses, are naturally resistant to several insects and herbicides and have a developing role in biofuel production (Akond *et al.* 2013). To initiate breeding programs for the amaranths, more knowledge is needed on their species relationships, genome structure, trait associations and hybridization potential. A novel genomic reduction protocol to sequence and identify SNP markers between amaranth mapping populations had identified 27 658 informative SNPs. These are utilized during subsequent association studies, genetic diversity analyses, population genetics and genetic linkage mapping experiments within the *Amaranthus* genus to lay the foundation for subsequent breeding programs (Maughan *et al.* 2009). The first draft genome and transcriptome of a grain amaranth (*Amaranthus hypochondriacus*) was published by (Sunil *et al.* 2014), identifying additional genomic SNP markers and important repeat elements (microsatellite markers). Due to the syntenic studies with *Arabidopsis thaliana*, *Vitis vinifera* and *Solanum lycopersicum*, it was possible to decipher and annotate more than 80% of the genes within a gene ontology (GO) category from *A. hypochondriacus*. Furthermore, the complete lysine biosynthetic pathway genes could be annotated using comparative genomics, explaining the high lysine content of the grains in this species. The authors conclude by emphasizing the importance of a genome sequence to exploit their unique traits in yield, high nutritional value and resistance to biotic and abiotic stresses (Sunil *et al.* 2014).

Table 2.3: Genomic tools developed for selected orphan crops

Orphan crop	Genomic tools	Reference
Amaranth (<i>Amaranthus</i> spp.)	SNP discovery	(Maughan <i>et al.</i> 2009)
	Genome and transcriptome sequencing	(Sunil <i>et al.</i> 2014)
	Draft genome and genome re-sequencing	(Clouse <i>et al.</i> 2015)
Adzuki bean (<i>Vigna angularis</i>)	Draft genome sequence	(Kang <i>et al.</i> 2015)
Bambara nut (<i>Vigna subterranea</i>)	Genomic and transcriptomic SSR markers	(Mayes <i>et al.</i> 2013)
	Linkage map and QTLs identified for phenotypic traits	(Ahmad <i>et al.</i> 2011)
Cassava (<i>Manihot esculenta</i>)	Draft genome sequence	(Prochnik <i>et al.</i> 2012)
	Genome wide SNP discovery	(Sakurai <i>et al.</i> 2013)
	Genetic map localizing immunity related genes	(Soto <i>et al.</i> 2015)
Chickpea (<i>Cicer arietinum</i>)	Molecular markers, diversity assessments, genetic maps	(Upadhyaya <i>et al.</i> 2011)
	Draft genome sequence	(Varshney 2015)
	Draft transcriptome sequence	(Garg <i>et al.</i> 2011)
Cowpea (<i>Vigna unguiculata</i>)	Molecular markers and linkage map construction	(Singh <i>et al.</i> 2014)
	Genetic mapping for aphid resistance	(Huynh <i>et al.</i> 2015)
Eggplant (<i>Solanum melongena</i>)	Draft genome sequence	(Hirakawa <i>et al.</i> 2014)
	SNP and SSR marker discovery	(Barchi <i>et al.</i> 2011)
	Linkage map construction	(Fukuoka <i>et al.</i> 2012)
Faba bean (<i>Vicia faba</i>)	SSR marker analysis	(Yang <i>et al.</i> 2012)
	QTL development for blight resistance	(Avila <i>et al.</i> 2004)
Finger millet (<i>Eleusine coracana</i>)	SSR marker analysis for blast resistance	(Babu <i>et al.</i> 2014)
	Transcriptome data	(Kumar <i>et al.</i> 2015)
Papaya (<i>Carica papaya</i>)	Draft genome sequence	(Ming <i>et al.</i> 2008)
	QTL development for fruit size and shape	(Blas <i>et al.</i> 2012)
Pigeon pea (<i>Cajanus cajan</i>)	Draft genome sequence	(Varshney <i>et al.</i> 2012a)
	Genetic map construction using DArT markers	(Yang <i>et al.</i> 2011)
Sorghum (<i>Sorghum bicolor</i>)	Draft genome sequence	(Paterson <i>et al.</i> 2009)
	SSR marker development	(Li <i>et al.</i> 2009)
Sweet potato (<i>Ipomoea batatas</i>)	De novo root transcriptome assembly	(Wang <i>et al.</i> 2010)
	QTL development for dry matter content	(Zhao <i>et al.</i> 2013)
Tef (<i>Eragrostis tef</i>)	Linkage map construction	(Girma <i>et al.</i> 2014)
	Draft genome sequence	(Cannarozzi <i>et al.</i> 2014)
Quinoa (<i>Chenopodium quinoa</i>)	SNP marker and linkage map development	(Maughan <i>et al.</i> 2012)

Pigeon pea

Pigeon pea (*Cajanus cajan*) is an important dicot legume food crop in Asia, South America, Africa, Central America and the Caribbean, predominantly cultivated by smallholder and resource-poor farmers for their high protein content (Varshney *et al.* 2012). Despite the importance of this crop globally to ensure food security, little research has been published (Varshney *et al.* 2012). Current yields by farmers are below projected yields, due to its susceptibility to various biotic and abiotic stresses. Varshney *et al.*, (2012) reported on the first draft genome sequence of pigeon pea (variety Asha). This draft genome sequence represented approximately 72% of the total estimated genome size and it was possible to identify putative gene regions involved in drought resistance (Varshney *et al.* 2012). Utilizing the large numbers of SNP and SSR markers identified during this study, various factors influencing pigeon pea yield, biotic resistance and specifically drought tolerance can be studied and manipulated through marker-assisted breeding and marker selection. Pigeon pea researchers were able to use the SNP markers (isolated from the draft genome) to investigate seven

genes related to determinacy in 142 pigeon pea lines. QTL analysis revealed a strong association between determinacy and the *CcTFL1* gene, which corresponded to the results of syntenic studies with soybean and common bean genomes. These results indicate that gene-based markers, which can be identified with relative ease from a draft genome assembly, can aid immensely in accelerated trait manipulation in different crop species (Mir *et al.* 2014).

Tef

Eragrostis tef (commonly known as tef) is a monocot cereal crop domesticated in several parts of eastern Africa, especially Ethiopia where the center of diversity lies (Cannarozzi *et al.* 2014). The crop is a critical source of food security in African nations, as they are rich in essential amino acids (especially lysine), minerals and protein (Cannarozzi *et al.* 2014). The main constraint of growing tef is its susceptibility to lodging, which affects the total yield harvested. Insignificant increases in yield were obtained through “superior” germplasm accession propagation (0.8% yield increase) as well as some conventional breeding techniques (up to 9% yield increase). This is inadequate to fulfill the growing demand of rural based farmers to feed the nearly 68 million residents of Ethiopia (Assefa *et al.* 2009). Due to constraints of conventional breeding, tef research has benefitted immensely from genomics based approaches. In 2011 three tef linkage maps had been produced with several sets of markers, but the 30 linkage groups had low marker density and therefore the QTLs which have been identified were not validated (Girma *et al.* 2014). However, (Cannarozzi *et al.* 2014) published the first genome sequence of tef. Their results include the identification of approximately 50 000 SSR markers, some of which lie in close proximity with gene clusters controlling yield or drought resistance/tolerance. Collinear relationships to closely related species such as sorghum revealed putative target genes for abiotic stress resistance studies. It was shown that epitopes causing a severe immune response in celiac patients are absent from the tef genome. These results illustrate how genome sequences for orphan crops, which possess increased natural diversity and unexploited agricultural traits, can be utilized.

Sweet Potato

Sweet potato (*Ipomoea batatas*) is an important agricultural dicot crop across the world, especially in tropical, sub-tropical and temperate zones. Uganda is the highest producer of orange-fleshed sweet potato in Africa. It contains high levels of carbohydrate, riboflavin, calcium and antioxidants (carotenoids and anthocyanins) (Naylor *et al.* 2004; Li *et al.* 2015b). Sweet potato is a popular starchy crop, often supplementing or replacing other popular starchy crops such as cassava or bananas in large areas of Uganda and other African countries dependent on less popular crops species for enhanced food security (Naylor *et al.* 2004). Due to the complexity of the sweet potato genome (hexaploid, large number of chromosomes; up to 90), mapping and QTL development has been limited. (Cervantes-Flores *et al.* 2011) produced one of the first genetic linkage maps for sweet potato using AFLP markers and identified several QTLs thought to play a role in dry matter, starch and β -carotene content. Additional linkage maps were produced using SSR markers and QTLs were identified for root dry matter content (Zhao *et al.* 2013) and storage root yield (Li *et al.* 2015b). The genome of two lines which most likely represent the diploid ancestors of hexaploid sweet potato (*Ipomoea trifida*), were recently *de novo* sequenced. Essential information concerning putative gene identifications as well as 1.5 million SNP markers and approximately 16 000 CNVs were identified, which could be applied to genetic studies of *I. trifida*, *I. batatas* and even the *Ipomoea* genus in general (Hirakawa *et al.* 2015).

Conclusion

Studies performed on a variety of orphan crops illustrate their potential use in biotic and abiotic stress resistance breeding, nutritional enhancement and yield improvement. The constraints orphan crop researchers are faced with are not necessarily always scientific in nature, but rather include (i) limited research budgets; (ii) poor communication between agricultural farmers and scientific researchers on respective needs; (iii) lack of government interventions pertaining to effective marketing and distribution of alternative crops and (iv) public concerns on health and environmental safety pertaining to genetically modified crops (Pray and Naseem 2007). Such restraints prohibit the development of alternative, enhanced or orphan crops for resource poor countries.

Biotechnology is rapidly developing into a prominent area of research in order to mitigate emerging challenges of food insecurity, poor nutrition and limited food supply/production. The field encompasses various techniques, such as tissue culture and propagation, molecular marker development, genetic maps, marker-assisted selection and genomic-assisted selection as well as transgenic technologies. Numerous research studies have been performed on major staple crops such as maize, wheat, rice and potato. However, declining genetic diversity within these crops and major constraints such as population growth and climate change has necessitated the need of research into less utilized, orphan crops. It is proposed that developing countries may benefit more from increased investment in locally adapted orphan crops, rather than stretching the limits of extant major crops that may be close to the limit of their genetic potential. Food security, especially in poorer countries, is an essential research objective for all nations and the exploitation of neglected or orphan crops have huge potential in attaining this common goal.

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
- Ahmad N et al. (2011) Developing genetic mapping and marker-assisted breeding techniques in Bambara groundnut (*Vigna subterranean* L.). In: II International Symposium on Underutilized Plant Species: Crops for the Future-Beyond Food Security 979:437-449
- 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
- AKosambo-Ayoo L, Bader M, Loerz H, Becker D (2013) Transgenic sorghum (*Sorghum bicolor* L. Moench) developed by transformation with chitinase and chitosanase genes from *Trichoderma harzianum* expresses tolerance to anthracnose. *African Journal of Biotechnology* 10:3659-3670
- Anjaneyulu E, Reddy PS, Sunita MS, Kishor PBK, Meriga B (2014) Salt tolerance and activity of antioxidative enzymes of transgenic finger millet overexpressing a vacuolar H⁺-pyrophosphatase gene (SbVPPase) from *Sorghum bicolor*. *Journal of Plant Physiology* 171:789-798
- Arora R (2014) Diversity in Underutilized Plant Species: An Asia-Pacific Perspective. Biodiversity International, New Delhi, India pp. 203
- Assefa K, Belay G, Tefera H, Yu J, Sorrells M, Tadele Z (2009) Breeding tef: conventional and molecular approaches. In: New approaches to plant breeding of orphan crops in Africa. Proceedings of an International Conference, Bern, Switzerland, 19-21 September 2007. Organizing Committee of the International Conference on New Approaches to Plant Breeding of Orphan Crops in Africa, pp 21-41
- Avila CM, Satovic Z, Sillero JC, Rubiales D, Moreno MT, Torres AM (2004) Isolate and organ-specific QTLs for ascochyta blight resistance in faba bean (*Vicia faba* L.). *Theoretical and Applied Genetics* 108:1071-1078
- Azadi H, Samiee A, Mahmoudi H, Jouzi Z, Rafiaani Khachak P, De Maeyer P, Witlox F (2015) Genetically modified crops and small-scale farmers: main opportunities and challenges. *Critical Reviews in Biotechnology* 36:434-446
- Babu BK, Dinesh P, Agrawal PK, Sood S, Chandrashekara C, Bhatt JC, Kumar A (2014) Comparative genomics and association mapping approaches for blast resistant genes in finger millet using SSRs. *PLoS One* 9:e99182

- Barchi L, Lanteri S, Portis E, Acquadro A, Vale G, Toppino L, Rotino GL (2011) Identification of SNP and SSR markers in eggplant using RAD tag sequencing. *BMC Genomics* 12:304
- Barton KA, Binns AN, Matzke AJ, Chilton M-D (1983) Regeneration of intact tobacco plants containing full length copies of genetically engineered T-DNA, and transmission of T-DNA to R1 progeny. *Cell* 32:1033-1043
- Bayer GY, Yemets A, Blume YB (2014) Obtaining the transgenic lines of finger millet *Eleusine coracana* (L.). with dinitroaniline resistance. *Cytology and Genetics* 48:139-144
- Blas AL, Yu Q, Veatch OJ, Paull RE, Moore PH, Ming R (2012) Genetic mapping of quantitative trait loci controlling fruit size and shape in papaya. *Molecular Breeding* 29:457-466
- Cabanos CS, Sajise AG, Garcia RN, Siar SV, Tecson-Mendoza EM (2014) Compositional analysis of transgenic papaya with delayed ripening trait. *The Philippine Agricultural Scientist* 96:331-339
- Cannarozzi G et al. (2014) Genome and transcriptome sequencing identifies breeding targets in the orphan crop tef (*Eragrostis tef*). *BMC Genomics* 15:581
- Caplan A, Herrera-Estrella L, Inze D, Van Haute E, Van Montagu M, Schell J, Zambryski P (1983) Introduction of genetic material into plant cells. *Science* 222:815-821
- Carstens M, Berger DK (2016) Biotechnology and Crop Disease Resistance in South Africa. *Plant Pathogen Resistance Biotechnology*:305
- Cervantes-Flores J et al. (2011) Identification of quantitative trait loci for dry-matter, starch, and β -carotene content in sweetpotato. *Molecular Breeding* 28:201-216
- Chang H-X, Brown PJ, Lipka AE, Domier LL, Hartman GL (2016) Genome-wide association and genomic prediction identifies associated loci and predicts the sensitivity of Tobacco ringspot virus in soybean plant introductions. *BMC Genomics* 17:1
- Chen G, Wang X, Hao J, Yan J, Ding J (2015) Genome-wide association implicates candidate genes conferring resistance to maize rough dwarf disease in maize. *PLoS One* 10:e0142001
- Chen H, Chen X, Gu H, Yuan X, Zhang H, Cui X (2013) Disease-resistant transgenic adzuki bean plants obtained through an efficient transformation system. *Crop and Pasture Science* 63:1090-1096
- Chetty C, Rossin C, Gruissem W, Vanderschuren H, Rey M (2013) Empowering biotechnology in southern Africa: establishment of a robust transformation platform for the production of transgenic industry-preferred cassava. *New Biotechnology* 30:136-143
- Chilton M-D, Tepfer DA, Petit A, David C, Casse-Delbart F, Tempé J (1982) *Agrobacterium rhizogenes* inserts T-DNA into the genomes of the host plant root cells. *Nature* 295:432

- Citadin C, Cruz A, Aragão F (2013) Development of transgenic imazapyr-tolerant cowpea (*Vigna unguiculata*). Plant Cell Reports 32:537-543
- Cloete PC, Idsardi EF (2013) Consumption of indigenous and traditional food crops: Perceptions and realities from South Africa. Agroecology and Sustainable Food Systems 37:902-914
- Clouse J, Adhikary D, Page JT *et al.*, (2015) The Amaranth Genome: Genome, Transcriptome, and Physical Map Assembly. The Plant Genome 9
- Cong L *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. Science 339:819-823
- Crowell S, Korniliev P, Falcão A, Ismail A, Gregorio G, Mezey J, McCouch S (2016) Genome-wide association and high-resolution phenotyping link *Oryza sativa* panicle traits to numerous trait-specific QTL clusters. Nature Communications 7
- Daru BH, Berger DK, Wyk AE (2015) Opportunities for unlocking the potential of genomics for African trees. New Phytologist 210:772-778
- Darwish NA, Khan RS, Ntui VO, Nakamura I, Mii M (2014) Generation of selectable marker-free transgenic eggplant resistant to *Alternaria solani* using the R/RS site-specific recombination system. Plant Cell Reports 33:411-421
- Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics 12:499-510
- Dawson IK, Hedley PE, Guarino L, Jaenicke H (2009) Does biotechnology have a role in the promotion of underutilised crops? Food Policy 34:319-328
- 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
- Ehlers J, Álvarez-Alfageme F, Higgins T, Lüthi C, Romeis J (2013) Resistance of α AI-1 transgenic chickpea (*Cicer arietinum*) and cowpea (*Vigna unguiculata*) dry grains to bruchid beetles (Coleoptera: Chrysomelidae). Bulletin of Entomological Research 103:373-381
- Evangelou E, Ioannidis JP (2013) Meta-analysis methods for genome-wide association studies and beyond. Nature Reviews Genetics 14:379-389
- Fahlgren N, Gehan MA, Baxter I (2015) Lights, camera, action: high-throughput plant phenotyping is ready for a close-up. Current Opinion in Plant Biology 24:93-99
- Fan W, Zhang M, Zhang H, Zhang P (2012) Improved tolerance to various abiotic stresses in transgenic sweet potato (*Ipomoea batatas*) expressing spinach betaine aldehyde dehydrogenase. PLoS One 7:e37344

- FAO I (2015) WFP. 2015. The State of Food Insecurity in the World
- Fox JL (2013) Mars collaborates to sequence Africa's neglected food crops. *Nature Biotechnology* 31:867-867
- Fraley RT *et al.* (1985) The SEV system: a new disarmed Ti plasmid vector system for plant transformation. *Nature Biotechnology* 3:629-635
- Fukuoka H *et al.* (2012) Development of gene-based markers and construction of an integrated linkage map in eggplant by using *Solanum* orthologous (SOL) gene sets. *Theoretical and Applied Genetics* 125:47-56
- Gahukar R (2014) Potential of Minor Food Crops and Wild Plants for Nutritional Security in the Developing World. *Journal of Agricultural & Food Information* 15:342-352
- Gaj T, Gersbach CA, Barbas CF (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in Biotechnology* 31:397-405
- Galluzzi G, López Noriega I (2014) Conservation and use of genetic resources of underutilized crops in the Americas—a continental analysis. *Sustainability* 6:980-1017
- Garg R, Patel RK, Tyagi AK, Jain M (2011) *De novo* assembly of chickpea transcriptome using short reads for gene discovery and marker identification. *DNA Research* 18:53-63
- Gebre E, Gugsu L, Schlüter U, Kunert K (2013) Transformation of *tef* (*Eragrostis tef*) by *Agrobacterium* through immature embryo regeneration system for inducing semi-dwarfism. *South African Journal of Botany* 87:9-17
- Ghanti SKK, Sujata K, Kumar BV, Karba NN, Janardhan Reddy K, Rao MS, Kishor PK (2011) Heterologous expression of P5CS gene in chickpea enhances salt tolerance without affecting yield. *Biologia Plantarum* 55:634-640
- Gimode D *et al.* (2016) Identification of SNP and SSR markers in finger millet using next generation sequencing technologies. *PLoS One* 11:e0159437
- Girma D, Assefa K, Chanyalew S, Cannarozzi G, Kuhlemeier C, Tadele Z (2014) The origins and progress of genomics research on *Tef* (*Eragrostis tef*). *Plant Biotechnology Journal* 12:534-540
- Gonsalves D (1998) Control of papaya ringspot virus in papaya: a case study. *Annual Review of Phytopathology* 36:415-437
- Govindaraj M, Vetriventhan M, Srinivasan M (2015) Importance of genetic diversity assessment in crop plants and its recent advances: An overview of its analytical perspectives. *Genetics Research International* 2015:1-14
- Grover A, Sharma P (2016) Development and use of molecular markers: past and present. *Critical Reviews in Biotechnology* 36:290-302

- Gugsa L, Kumlehn J (2011) Somatic embryogenesis and massive shoot regeneration from immature embryo explants of tef. *Biotechnology Research International* 2011
- Guo Wc *et al.* (2016) Development of selectable marker - free transgenic potato plants expressing cry3A against the Colorado potato beetle (*Leptinotarsa decemlineata* Say). *Pest Management Science* 72:497-504
- Hajyzadeh M, Turktas M, Khawar KM, Unver T (2015) miR408 overexpression causes increased drought tolerance in chickpea. *Gene* 555:186-193
- Hanafy MS, El-Banna A, Schumacher HM, Jacobsen H-J, Hassan FS (2013) Enhanced tolerance to drought and salt stresses in transgenic faba bean (*Vicia faba* L.) plants by heterologous expression of the PR10a gene from potato. *Plant Cell Reports* 32:663-674
- Hefferon K (2015) Nutritionally enhanced food crops; progress and perspectives. *International Journal of Molecular Sciences* 16:3895-3914
- Hirakawa H *et al.* (2015) Survey of genome sequences in a wild sweet potato, *Ipomoea trifida* (HBK) G. Don. *DNA Research*:dsv002
- Hirakawa H *et al.* (2014) Draft genome sequence of eggplant (*Solanum melongena* L.): the representative solanum species indigenous to the old world. *DNA Research* 21:649-660
- Huynh B-L, Ehlers JD, Ndeve A, Wanamaker S, Lucas MR, Close TJ, Roberts PA (2015) Genetic mapping and legume synteny of aphid resistance in African cowpea (*Vigna unguiculata* L. Walp.) grown in California. *Molecular Breeding* 35:1-9
- Ignacimuthu S, Ceasar SA (2012) Development of transgenic finger millet (*Eleusine coracana* (L.) Gaertn.) resistant to leaf blast disease. *Journal of Biosciences* 37:135-147
- Jaenicke H, Höschle-Zeledon I (2006) Strategic framework for underutilized plant species research and development: with special reference to Asia and the Pacific, and to Sub-Saharan Africa. *Biodiversity International*
- James C (2015) ISAAA Brief 49–2014: Executive summary. Global status of commercialized biotech/GM crops: 2014. International Service for the Acquisition of Agri-Biotech Applications.
- Jia H, Wang N (2014) Targeted genome editing of sweet orange using Cas9/sgRNA. *PLoS One* 9:e93806
- Jofre-Garfias A, Villegas-Sepúlveda N, Cabrera-Ponce J, Adame-Alvarez R, Herrera-Estrella L, Simpson J (1997) *Agrobacterium*-mediated transformation of *Amaranthus hypochondriacus*: light-and tissue-specific expression of a pea chlorophyll a/b-binding protein promoter. *Plant Cell Reports* 16:847-852
- Kang YJ *et al.* (2015) Draft genome sequence of adzuki bean, *Vigna angularis*. *Scientific Reports* 5:8069

- Khatodia S, Bhatotia K, Passricha N, Khurana SP, Tuteja N (2016) The CRISPR/Cas Genome-Editing Tool: Application in Improvement of Crops. *Frontiers in Plant Science* 7
- Khush GS (2001) Green revolution: the way forward. *Nature Reviews Genetics* 2:815-822
- Kole C et al. (2015) Application of genomics-assisted breeding for generation of climate resilient crops: progress and prospects. *Frontiers in Plant Science* 6:563
- Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9:1
- Kumar A, Gaur VS, Goel A, Gupta AK (2015) *De novo* assembly and characterization of developing spikes transcriptome of finger millet (*Eleusine coracana*): A minor crop having nutraceutical properties. *Plant Molecular Biology Reporter* 33:905-922
- Li K-T et al. (2015a) Increased bioavailable vitamin B6 in field-grown transgenic cassava for dietary sufficiency. *Nature Biotechnology* 33:1029-1032
- Li M, Yuyama N, Luo L, Hirata M, Cai H (2009) *In silico* mapping of 1758 new SSR markers developed from public genomic sequences for sorghum. *Molecular Breeding* 24:41-47
- Li R, Zhai H, Kang C, Liu D, He S, Liu Q (2015b) *De novo* transcriptome sequencing of the orange-fleshed sweet potato and analysis of differentially expressed genes related to carotenoid biosynthesis. *International Journal of Genomics* 2015:843802
- Liang Z, Zhang K, Chen K, Gao C (2014) Targeted mutagenesis in *Zea mays* using TALENs and the CRISPR/Cas system. *Journal of Genetics and Genomics* 41:63-68
- Lopes MS et al. (2015) Exploiting genetic diversity from landraces in wheat breeding for adaptation to climate change. *Journal of Experimental Botany*:erv122
- Makarova KS et al. (2011) Evolution and classification of the CRISPR–Cas systems. *Nature Reviews Microbiology* 9:467-477
- Maughan P et al. (2012) Single nucleotide polymorphism identification, characterization, and linkage mapping in quinoa. *The Plant Genome* 5:114-125
- 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
- Mavengahama S, McLachlan M, De Clercq W (2013) The role of wild vegetable species in household food security in maize based subsistence cropping systems. *Food Security* 5:227-233

- Mayes S *et al.* (2013) Molecular analysis of Bambara Groundnut, an underutilised african legume crop as part of the BAMLINK Project–What lessons can we learn. In: 2nd International symposium on underutilized species: Crops for the future-beyond food security pp 451-458
- Michno J-M, Wang X, Liu J, Curtin SJ, Kono TJ, Stupar RM (2015) CRISPR/Cas mutagenesis of soybean and *Medicago truncatula* using a new web-tool and a modified Cas9 enzyme. *GM Crops & Food* 6:243-252
- Ming R *et al.* (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 452:991-996
- Mir C *et al.* (2013) Out of America: tracing the genetic footprints of the global diffusion of maize. *Theoretical and Applied Genetics* 126:2671-2682
- Mir RR *et al.* (2014) Candidate gene analysis for determinacy in pigeonpea (*Cajanus* spp.). *Theoretical and Applied Genetics* 127:2663-2678
- 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
- Moose SP, Mumm RH (2008) Molecular plant breeding as the foundation for 21st century crop improvement. *Plant Physiology* 147:969-977
- Morozova O, Marra MA (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics* 92:255-264
- Muramoto N *et al.* (2012) Transgenic sweet potato expressing thionin from barley gives resistance to black rot disease caused by *Ceratocystis fimbriata* in leaves and storage roots. *Plant Cell Reports* 31:987-997
- Muranty H, Jorge V, Bastien C, Lepoittevin C, Bouffier L, Sanchez L (2014) Potential for marker-assisted selection for forest tree breeding: lessons from 20 years of MAS in crops. *Tree Genetics & Genomes* 10:1491-1510
- Murphy DJ (2015) Roles of genomics in addressing global food security. *Genomics and Society: Ethical, Legal, Cultural and Socioeconomic Implications*:259
- Narayanan N, Beyene G, Chauhan RD, Gaitan-Solis E, Grusak MA, Taylor N, Anderson P (2015) Overexpression of *Arabidopsis* VIT1 increases accumulation of iron in cassava roots and stems. *Plant Science* 240:170-181
- Naylor RL, Falcon WP, Goodman RM, Jahn MM, Sengooba T, Tefera H, Nelson RJ (2004) Biotechnology in the developing world: a case for increased investments in orphan crops. *Food Policy* 29:15-44

- Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE (2011) Landscape of next-generation sequencing technologies. *Analytical Chemistry* 83:4327-4341
- Oh JE, Kwon Y, Kim JH, Noh H, Hong S-W, Lee H (2011) A dual role for MYB60 in stomatal regulation and root growth of *Arabidopsis thaliana* under drought stress. *Plant Molecular Biology* 77:91-103
- Omboki RB, Wu W, Xie X, Mamadou G (2015) Applications of genomics in Plant Breeding. *International Journal of Agriculture and Crop Sciences* 8:560
- Oniang'o R, Allotey J, Malaba SJ (2004) Contribution of indigenous knowledge and practices in food technology to the attainment of food security in Africa. *Journal of Food Science* 69:CRH87-CRH91
- Paterson AH *et al.* (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 457:551-556
- Perez-Pinera P, Ousterout DG, Gersbach CA (2012) Advances in targeted genome editing. *Current Opinion in Chemical Biology* 16:268-277
- Pingali PL (2012) Green Revolution: Impacts, limits, and the path ahead. *Proceedings of the National Academy of Sciences* 109:12302-12308
- Prado JR *et al.* (2014) Genetically engineered crops: from idea to product. *Annual Review of Plant Biology* 65:769-790
- Pray CE, Naseem A (2007) Supplying crop biotechnology to the poor: Opportunities and constraints. *Journal of Development Studies* 43:192-217
- Prochnik S *et al.* (2012) The cassava genome: current progress, future directions. *Tropical Plant Biology* 5:88-94
- Rahman H, Ramanathan V, Nallathambi J, Duraialagaraja S, Muthurajan R (2016) Over-expression of a NAC 67 transcription factor from finger millet (*Eleusine coracana* L.) confers tolerance against salinity and drought stress in rice. *BMC Biotechnology* 16:7
- Rai NP, Rai GK, Kumar S, Kumari N, Singh M (2013) Shoot and fruit borer resistant transgenic eggplant (*Solanum melongena* L.) expressing cry1Aa3 gene: Development and bioassay. *Crop Protection* 53:37-45
- Ramu S *et al.* (2012) Expression of a synthetic cry1AcF gene in transgenic pigeon pea confers resistance to *Helicoverpa armigera*. *Journal of Applied Entomology* 136:675-687
- Rastogi A, Shukla S (2013) Amaranth: a new millennium crop of nutraceutical values. *Critical Reviews in Food Science and Nutrition* 53:109-125

- Ruane J, Sonnino A (2010) Agricultural biotechnologies in developing countries and their possible contribution to food security. *Journal of Biotechnology* 156:356-363
- Sakurai T, Mochida K, Yoshida T, Akiyama K, Ishitani M, Seki M, Shinozaki K (2013) Genome-wide discovery and information resource development of DNA polymorphisms in cassava. *PLoS One* 8:e74056
- Sardos J *et al.* (2016) A genome-wide association study on the seedless phenotype in banana (*Musa* spp.) reveals the potential of a selected panel to detect candidate genes in a vegetatively propagated crop. *PLoS One* 11:e0154448
- Sauvage C *et al.* (2014) Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits. *Plant Physiology* 165:1120-1132
- Searchinger T, Hanson C, Lacape J-M (2014) Crop Breeding: Renewing the Global Commitment. World Resources Institute Working Paper: Installment 7
- Shan Q *et al.* (2013) Targeted genome modification of crop plants using a CRISPR-Cas system. *Nature Biotechnology* 31:686-688
- Silva Dias JCD (2014) Guiding strategies for breeding vegetable cultivars. *Agricultural Sciences* 05:9-32
- Singh B, Timko MP, Aragao FJ (2014) Advances in cowpea improvement and genomics. In: *Legumes in the Omic Era*. Springer, pp 131-153
- Solleti SK, Bakshi S, Purkayastha J, Panda SK, Sahoo L (2008) Transgenic cowpea (*Vigna unguiculata*) seeds expressing a bean α -amylase inhibitor 1 confer resistance to storage pests, bruchid beetles. *Plant Cell Reports* 27:1841-1850
- Sonah H, O'Donoghue L, Cober E, Rajcan I, Belzile F (2015) Identification of loci governing eight agronomic traits using a GBS - GWAS approach and validation by QTL mapping in soya bean. *Plant Biotechnology Journal* 13:211-221
- Soto JC *et al.* (2015) A genetic map of cassava (*Manihot esculenta* Crantz) with integrated physical mapping of immunity-related genes. *BMC Genomics* 16:190
- Sunil M *et al.* (2014) The draft genome and transcriptome of *Amaranthus hypochondriacus*: A C4 dicot producing high-lysine edible pseudo-cereal. *DNA Research*:dsu021
- Surekha C, Kumari KN, Aruna L, Suneetha G, Arundhati A, Kishor PK (2014) Expression of the *Vigna aconitifolia* P5CSF129A gene in transgenic pigeonpea enhances proline accumulation and salt tolerance. *Plant Cell, Tissue and Organ Culture* 116:27-36

- Tadele Z, Assefa K (2012) Increasing food production in Africa by boosting the productivity of understudied crops. *Agronomy* 2:240-283
- Tilman D (2015) Food & Health of a Full Earth. *Daedalus* 144:5-7
- Treangen TJ, Salzberg SL (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics* 13:36-46
- Upadhyaya HD, Thudi M, Dronavalli N, Gujaria N, Singh S, Sharma S, Varshney RK (2011) Genomic tools and germplasm diversity for chickpea improvement. *Plant Genetic Resources* 9:45-58
- Utomo HS, Wenefrida I, Linscombe SD (2012) Progression of DNA marker and the next generation of crop development. INTECH Open Access Publisher
- Varshney R *et al.* (2012) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nature Biotechnology* 30:83-89
- Varshney RK, Nayak SN, May GD, Jackson SA (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends in Biotechnology* 27:522-530
- Varshney RK (2015) Exciting journey of 10 years from genomes to fields and markets: Some success stories of genomics-assisted breeding in chickpea, pigeonpea and groundnut. *Plant Science* 242:98-107
- Vartia S *et al.* (2016) A novel method of microsatellite genotyping-by-sequencing using individual combinatorial barcoding. *Royal Society Open Science* 3:150565
- Visarada K, Padmaja P, Saikishore N, Pashupatinath E, Royer M, Seetharama N, Patil J (2014) Production and evaluation of transgenic sorghum for resistance to stem borer. *In Vitro Cellular & Developmental Biology-Plant* 50:176-189
- Wan F *et al.* (2014) Heterologous expression of *Arabidopsis* C-repeat binding factor 3 (AtCBF3) and cold-regulated 15A (AtCOR15A) enhanced chilling tolerance in transgenic eggplant (*Solanum melongena* L.). *Plant Cell Reports* 33:1951-1961
- Wang Z *et al.* (2010) *De novo* assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweetpotato (*Ipomoea batatas*). *BMC Genomics* 11:726
- Whitty CJM, Jones M, Tollervey A, Wheeler T (2013) Biotechnology: Africa and Asia need a rational debate on GM crops. *Nature* 497:31-33
- Yadav JS *et al.* (2011) RNAi - mediated resistance to Cassava brown streak Uganda virus in transgenic cassava. *Molecular Plant Pathology* 12:677-687

- Yamori W, Kondo E, Sugiura D, Terashima I, Suzuki Y, Makino A (2016) Enhanced leaf photosynthesis as a target to increase grain yield: insights from transgenic rice lines with variable Rieske FeS protein content in the cytochrome b6/f complex. *Plant, Cell & Environment* 39:80-87
- Yang S *et al.* (2011) The first genetic map of pigeon pea based on diversity arrays technology (DArT) markers. *Journal of Genetics* 90:103-109
- Yang T *et al.* (2012) High-throughput novel microsatellite marker of faba bean via next generation sequencing. *BMC Genomics* 13:1
- Zhao N *et al.* (2013) A genetic linkage map based on AFLP and SSR markers and mapping of QTL for dry-matter content in sweetpotato. *Molecular Breeding* 32:807-820
- Zhou W, Yang J, Hong Y, Liu G, Zheng J, Gu Z, Zhang P (2015a) Impact of amylose content on starch physicochemical properties in transgenic sweet potato. *Carbohydrate Polymers* 122:417-427
- Zhou Z *et al.* (2015b) Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature Biotechnology* 33:408-414
- Ziervogel G *et al.* (2014) Climate change impacts and adaptation in South Africa. *Wiley Interdisciplinary Reviews: Climate Change* 5:605-620

CHAPTER 3

A DRAFT GENOME AND TRANSCRIPTOME ASSEMBLY OF THE ECONOMICALLY IMPORTANT ORPHAN CROP *AMARANTHUS* *TRICOLOR*

Abstract

The effects of global warming have major impacts on agricultural sectors across the world. Rise in temperatures and the growing population size are negatively impacting the yield and growth of conventional crop species, consequently resulting in widespread food insecurity and scarcity. Researchers are renewing their interest and research on indigenous crop species due to their natural genetic diversity and adaptations to the climatic atmosphere in each region across the world. *Amaranthus* species are considered important, edible grain and leafy vegetables in several countries across the world, including the Americas, Asia and especially in Africa, where access to nutritional food is severely hampered by poor economic growth. Particularly, *Amaranthus tricolor* is a naturally occurring leafy vegetable with natural adaptation to water scarcity, growing with sufficient leaf yield even in extreme climatic conditions. Coupled with the high concentrations of essential minerals and micronutrients, *A. tricolor* is considered a strong candidate for future food development. In the current study, the whole genome sequence was determined for *A. tricolor* and together with supporting transcriptomic data from leaf, root and shoot material; a draft annotation was performed. In total, 135.5 Gb reads were produced from paired end and long mate pair NGS (Illumina) sequencing data, resulting in a total scaffold size of 323 Mb, with 15.72% gapped regions and an N50 size of 51 Kb. Repetitive element analysis revealed that approximately 54% of the genome is composed of repeated elements. CEGMA analysis indicated the presence of 77% complete core eukaryotic genes, and an average of two orthologues per CEG in the scaffolded assembly. BUSCO analysis revealed 88.4% completely assembled orthologues within the genome assembly, while 8.6% and 3% of the reference genes were missing or fragmented, respectively. A *de novo* transcriptome assembly was obtained from RNA-Seq analysis of root, shoot and leaf tissues. After an FPKM filtering was applied to remove lowly expressed genes or artefacts of an incomplete assembly, 79 380 expressed contigs were assembled with an N50 of 1 036bp. Using the transcriptomic data as evidence of gene predictions, the genome was annotated with 16 495 coding regions with a median length of 4 461 bp. Whole genome sequencing of traditional vegetables such as *A. tricolor* will enable researchers to specifically target agriculturally important traits such as nutritional content and drought tolerance and exploit them for future food security.

Introduction

Understanding the way DNA is structured has been a fundamental breakthrough in determining how alterations in the DNA sequence influences disease, population structures and evolutionary lineages in many species. First generation DNA sequencing was established by Sanger and Coulson in 1975, and to a lesser extent the chemical method developed by Maxam and Gilbert in 1976-1977 (Niedringhaus *et al.* 2011; Raza and Ahmad 2016). For at least three decades after its inception, the Sanger method has been the most popular choice for sequence-based projects, due to the accuracy, long read length, and its capacity to read homopolymer regions and motif repeats (Morozova and Marra 2008; Niedringhaus *et al.* 2011; Levy and Myers 2016). However, limitations in throughput and high costs associated with sample preparation and base determination have led to the development of more targeted sequencing technologies.

Next generation sequencing (NGS) technologies refer to innovative DNA base determination strategies, similar in principle to Sanger technologies but extending to massively parallel sequencing of millions of fragments simultaneously. The NGS technology greatly reduces the cost of sample preparation, the cost per base sequenced and greatly increases large dataset output time (Deschamps and Campbell 2009; Raza and Ahmad 2016). The methods involved in different NGS platforms rely on sequencing by ligation or sequencing by synthesis, including reversible chain termination and pyrosequencing (Niedringhaus *et al.* 2011). Since the inception of NGS in the years 2000-2005, several different platforms, technologies and protocols have been developed, each striving to provide more in-depth coverage of genomes at a reduced associated sample preparation time and cost per sequence, while simultaneously increasing the number and length of sequence reads produced. The most popular platforms to date include Roche (currently the 454 GS-FLX Titanium, discontinued in 2013), Illumina (MiSeq – 2011; HiSeq – 2012; NovaSeq – 2017) and Thermo Fisher Scientifics' Ion Torrent (PII – 2015; Ion S5 – 2016). Recently the emergence of single, long read, real-time sequencers have emerged, which produces even more, longer read data in a shorter time period, including Pacific Biosciences (PacBio RS II – 2011, Sequel – 2015) and Oxford Nanopore (MinION – 2014).

The analysis and biological interpretation of NGS data into usable information remains a major drawback of high-throughput sequencing. Appropriate knowledge of computational tools and a high-performance computing infrastructure is necessary to analyse NGS data, which is not often available to researchers with

limited resources. Bioinformatic challenges to be considered include (i) standardized analysis protocols or integrative software resources that should be available to researchers for consistent data analysis such as BioExtract (Lushbough *et al.* 2010) and Galaxy (Goecks *et al.* 2010) and (ii) the ability of researchers to consistently reproduce experiments from previous studies using published source codes and scripts. An additional challenge for bioinformaticians is the storage of immense datasets, due to limited computer resources at institutions and the high cost of local data storage and maintenance. Online database storage in the form of Sequence Read Archives (NCBI), RNA-Seq Atlas and NGS Catalog are gaining popularity for depositing large sequence dataset for easy access by other researchers (Raza and Ahmad 2016).

To date, researchers have been able to sequence and annotate thousands of whole genomes spanning all kingdoms, including eukaryotes (5 247), prokaryotes (131 202), viruses (14 025) and organelles (11 371) (www.ncbi.nlm.nih.gov/genome - 20 February 2018). The downstream application of these genomes is limitless in identifying mutations or structural variants in hereditary or novel diseases, modes of infection and replication in disease causing bacteria and viruses as well as the identification of crucial points in biological pathways that can serve as targets for future genetic manipulation (Goodwin *et al.* 2016; Raza and Ahmad 2016). In addition to genomic information, the development of next-generation deep sequencing technologies has also had an impact on transcriptome/expressed genome studies. In particular, exome capture/sequencing (recently termed exomics) has gained increasing popularity due to the relative ease of sequencing entire transcripts, identifying gene splice regions or targeted mutations in the protein coding regions of a genome (Hashmi *et al.* 2015). In plant genomes, the exome makes up approximately 1.2% of the genome, making exomics more manageable in terms of data generation and analysis (Kaur and Gaikwad 2017).

The genus *Amaranthus* belongs to the order Caryophyllales and consists of approximately 60 species that are considered globally important vegetable crops. Several species within the *Amaranthus* genus possess desirable traits for sustainable food security such as high tolerance to changing environmental conditions, high seed yield, natural pest and insect resistance, high seed lysine content, elevated protein and iron levels in the leaves, high biomass for bio-fuel production and the potential to use amaranth for phytoremediation of disturbed land (Sitkey *et al.* 2013; Venskutonis and Kraujalis 2013; Achigan-Dako *et al.* 2014; Alemayehu *et al.* 2015; Bansah and Addo 2016).

Two main types of amaranth plants are currently recognized, the pale seeded varieties that are used as grain crops and the broad, leafy (dark-seeded) varieties that are utilized as leafy vegetable crops (Achigan-Dako *et al.* 2014). The grain crops (*Amaranthus caudatus*, *Amaranthus cruentus* and *Amaranthus hypochondriacus*) have been recognized for their high protein levels in the grains, especially since these proteins contain high levels of the essential amino acids lysine and methionine. Lysine acts as a precursor to produce carnitine, which in turn plays an important role in maintaining acceptable blood cholesterol levels in humans (Tomé and Bos 2007). The leafy amaranths (including, but not limited to, *A. tricolor*, *A. cruentus*, *A. hybridus*, *A. blitum* and *A. viridis*) are considered excellent sources of a variety of dietary components (Gupta *et al.* 2005). The leaves contain significant amounts of protein, β -carotene, iron, calcium, vitamin C and folic acid (Kunyanga *et al.* 2013; Venskutonis and Kraujalis 2013; Achigan-Dako *et al.* 2014). Particularly, *A. tricolor* (a popular edible leafy crop in Africa) has the potential to sustainably produce edible, nutritional leaves both during the rainy and dry seasons, which ensures a food source for resource poor farmers throughout the year (Grubben 1980; Jansen *et al.* 2004; Ribeiro *et al.* 2017). To date, farmers have been using *A. tricolor* mainly as a subsistence crop, and little to no data is available on the commercial production of this crop (Achigan-Dako *et al.*, 2014). Although it was found that reproductive growth was sensitive to low water levels, a higher concentration of calcium and crude proteins were observed in *A. tricolor* leaves under these conditions. Only when less than 20% of total available water was applied, was a decrease in leaf yield and vegetative growth of *A. tricolor* crops observed (Ribeiro *et al.* 2017). In terms of mineral content of the leaves, *A. tricolor* has been found to contain high levels of calcium, zinc and iron that serve as important cofactors for metabolic processes. A long-term aim for *Amaranthus* breeding includes crossing a grain species with an edible leafy species, such as *A. tricolor*, to incorporate the respective beneficial traits (edible grain and palatable leaves) in one genotype without compromising yield.

To date, the sequencing of an *Amaranthus* genome has been focused on a specific grain species, *A. hypochondriacus*. During a study by Sunil *et al.* (2014), a draft genome and transcriptome for *A. hypochondriacus* was sequenced and assembled (Sunil *et al.* 2014). The genome was further mined for genic data relating to lysine production in the grains, identifying potential gene loci that play a role within the lysine biosynthetic pathway. The authors hypothesize that knowledge on the polymorphisms within these loci may shed light on the underlying mechanisms that control lysine production in amaranth compared to other cereal

crops. The sequencing of an alternative cultivar of *A. hypochondriacus* (Plainsman, a popular commercial variety in the USA), revealed a high-quality draft genome of approximately 377 Mb contained in 3518 scaffolds (Clouse *et al.* 2016). In addition, seven grain amaranths including varieties of *A. hypochondriacus*, *A. caudatus*, *A. cruentus* and *A. hybridus* were re-sequenced at a low coverage. Ensuing SNP marker analysis enabled the researchers to prove phylogenetically that *A. hybridus* is the most probable progenitor of the grain amaranths (Clouse *et al.* 2016). The initial two draft genomes were significantly improved upon by contig gap closing using PacBio data, as well as scaffold assemblies using Hi-C mapping data (Lightfoot *et al.* 2017). The result was a chromosome scale assembly of *A. hypochondriacus*, with contig and scaffold N50s of 1.25 Mb and 24.4 Mb respectively. Approximately 98% of the assembly was contained in 16 scaffolds, likely representing the haploid number of 16 chromosomes for *A. hypochondriacus*. Using the genome at this level of assembly can facilitate numerous downstream studies such as investigating the speciation and whole genome evolution within the genus *Amaranthus*, in-depth gene discovery for important agricultural traits (through breeding mapping populations or genome wide marker association programs) and targeted manipulation of gene regions responsible for abiotic and biotic resistance (Lightfoot *et al.* 2017).

During the present study, the hypothesis that a good quality draft genome can be assembled for the important leafy vegetable *Amaranthus tricolor*, using a combination of paired-end and long mate pair Illumina NGS data was addressed. The genome was annotated using supporting transcriptomic data generated from leaf, stem and root material, as well as the existing protein database SwissProt. Complementing the genome sequences with transcriptome sequences could enable more accurate assemblies and annotations, since the addition of protein coding sequences aid gene prediction algorithms. The draft genome was compared to the published genomes of an important grain amaranth, *A. hypochondriacus* as well as a close relative within the Caryophyllales order, *Beta vulgaris*. Comparative genomics could potentially identify loci linked to advantageous genes associated with important agricultural traits and would be a key step in breeding crops incorporating species-specific traits in superior genotypes. Downstream applications during the present study included the identification of potential microsatellite marker loci for genetic diversity analysis (Chapter 4), as well as extracting the full-length chloroplast sequence to use in phylogenetic analysis (Chapter 5) within the *Amaranthus* genus.

Materials and Methods

Plant growth

Seed for the leafy vegetable, *Amaranthus tricolor* (accession SAG29), was obtained from the plant germplasm bank at the Agricultural Research Council Vegetable and Ornamental Plant Institute (ARC-VOPI) in Pretoria, South Africa. The original seeds were collected in the USA (exact location unknown) and had undergone several self-pollinating generation cycles under glasshouse conditions at ARC-VOPI before being deposited in the germplasm bank (Willem Janse van Rensburg, personal communication). The seeds were germinated in irrigated glasshouses at ARC-VOPI with natural light intensity, day temperatures of 25°C-35°C and night temperatures of 20°C-25°C. Leaf, stem and root material were harvested at approximately 30-40 days post germination (Figure 3.1). All tissues were stored immediately at -80°C until use.



Figure 3.1: Mature *Amaranthus tricolor* plants, seeds and inflorescence grown under general glasshouse conditions at ARC-VOPI, Pretoria, South Africa.

Extraction of DNA

Total genomic DNA was extracted from 100 mg fresh leaf material from a single 30-40-day old *A. tricolor* plant, using a DNeasy Plant Mini DNA Isolation kit (Qiagen, Valencia CA, USA) following the protocol outlined by the manufacturer. The genomic DNA was quantified using the Qubit® 2.0 Fluorometer Broad Range dsDNA quantification assay (Invitrogen, Life Technologies, CA, USA) and visualized on a 1% agarose gel stained with 0.5 µg/ml ethidium bromide (Figure 3.2).

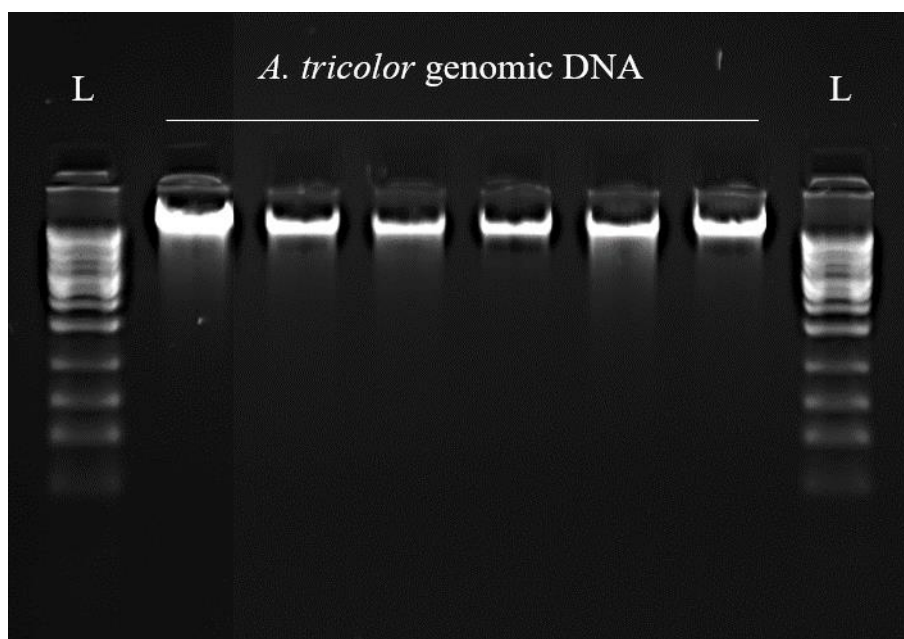


Figure 3.2: High molecular weight genomic DNA extraction from leaves of a 30-40-day old *A. tricolor* plant. (L) = O'Generuler™ 1 kb Plus DNA Ladder (Thermo Scientific™).

DNA Library Preparation and Sequencing

One microgram of genomic DNA was used to prepare sequencing libraries with the Nextera DNA Library Preparation Kit v3 (Illumina, San Diego, USA) according to the manufacturer's protocol. After enzymatic digestion of the genomic DNA, a sample specific index was added to facilitate multiplexing during the sequencing run using the Nextera Index Kit (Illumina). Sequencing libraries were size selected by excising an approximately 300 bp band from a 1% agarose gel stained with 0.5 µg/ml ethidium bromide. Gel purification

was performed using the Qiagen Gel Elution kit (Qiagen) according to the manufacturer's protocol. DNA quantification was performed using Qubit fluorometry (Invitrogen). The barcoded paired end (insert size 300 bp) and long mate pair (insert size 3 kb and 5 kb) libraries were subjected to cluster generation on the cBot (Illumina) and sequenced on the HiScanSQ Illumina sequencer (Illumina, San Diego, USA), at the Biotechnology Platform Sequencing Facility (Agricultural Research Council (ARC-BTP), Pretoria, South Africa). Sequencing reads of 2 x 100 bp were generated for each library.

Extraction of RNA

Approximately 100 mg of fresh leaf, root and shoot tissues were collected from a 30-40-day old *A. tricolor* plant, and immediately washed with RNase-free water. The tissues were snap frozen in liquid nitrogen and stored at -80°C until use. Total RNA was extracted using the Nucleospin RNA Plant Kit (Macherey-Nagel, Düren, Germany) according to manufacturer's instructions. The RNA was quantified using the Qubit® 2.0 Fluorometer RNA Broad Range quantification assay (Invitrogen) and visualized on a 1% agarose gel stained with 0.5 µg/ml ethidium bromide, as well as on a 2100 Bioanalyzer (Agilent Genomics) using the Agilent RNA 6000 Pico Kit (Figure 3.3). The RNA integrity number (RIN) was calculated as 8.4 for the extraction from leaf material.

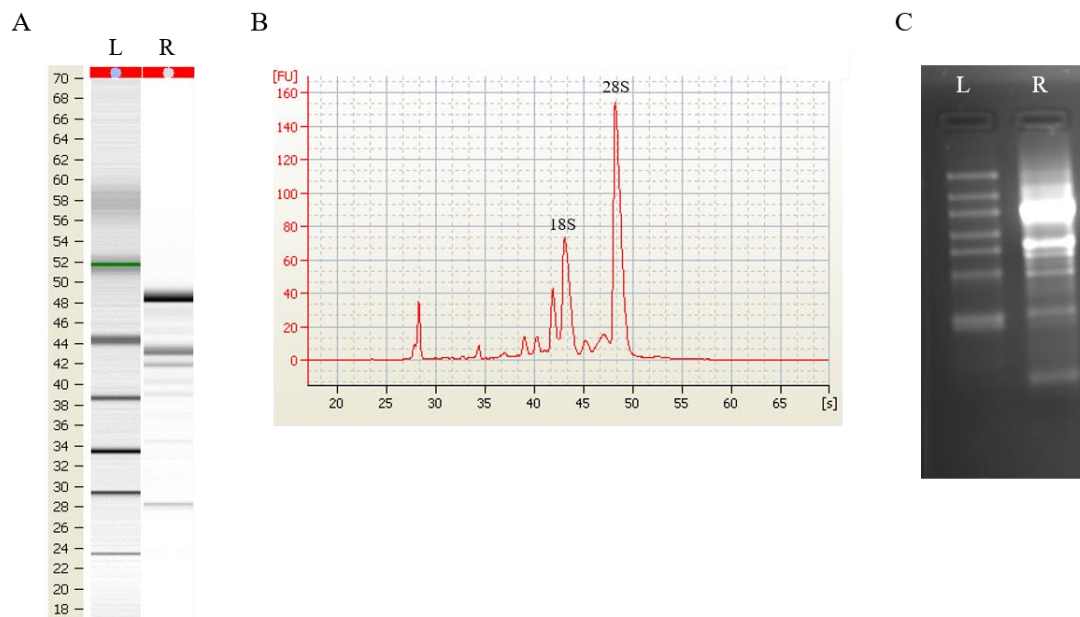


Figure 3.3: Total RNA extracted from 30-40-day old *A. tricolor* leaves. RNA quality and quantity were measured on a 2100 Bioanalyzer (A) L = RNA 6000 Pico Ladder, R = Total RNA and (B) Electropherogram representing ribosomal peaks (18S and 28S), as well as a 1% agarose gel (C) L = O'Generuler™ 1 kb Plus DNA Ladder (Thermo Scientific™), R = Total RNA.

RNA Library Preparation and Sequencing

Two micrograms of RNA for each tissue type (leaves/roots/shoots) were individually subjected to library preparation using the TruSeq RNA Library Preparation Kit (Illumina) according to the manufacturer's guidelines. The poly-A mRNA was subjected to enzymatic fragmentation, after which the fragments were size selected to 300 bp on a 1% agarose gel stained with 0.5 µg/ml ethidium bromide. Library quality and quantity was measured with a Bioanalyzer (Agilent). Three unique barcodes were added to each cDNA library and pooled in equimolar quantities for cluster generation on the cBot (Illumina) and sequencing on the HiScanSQ Illumina Sequencer (Illumina) at the ARC-BTP (Pretoria, South Africa). Sequencing reads of 2 x 100 bp were generated for each library.

Read Quality Control

The genomic and transcriptomic reads were de-multiplexed using the Casava 1.8.2 pipeline (Illumina) and adapters/barcodes (together with the first three low quality base pairs of each read) were removed using Trimmomatic version 2.33 (www.trimmomatic.com). Trimmomatic also removed portions of poor quality reads using a sliding window approach, with settings: LEADING:3, TRAILING:3, SLIDING-WINDOW:4:20 and MINLEN:36. Quality control analysis was performed using FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>), where read pairs with (i) a quality PHRED score of ≥ 20 , (ii) a total of less than 20 ambiguous base pairs per fragment and (iii) a read length of above 40 bp were considered sufficient quality and retained for further analysis. Filtered reads were mapped to the chloroplast sequence of a close relative, *Beta vulgaris*, using CLC Bio Genomics Workbench v8 (Aarhus, Denmark) with default parameters to identify chloroplast specific fragments. Reads that mapped to the chloroplast were removed from further genomic and transcriptomic assemblies.

De novo Genome Assembly

A total of 135.3 Gb reads (mean length 99 bp) were obtained across the paired end (93.8 Gb) and long mate pair (41.5 Gb) libraries after the Trimmomatic software removed barcoding/indexing primers and FastQC filtering was applied (Table 3.1). Approximately 90.7 Gb (67%) of reads were incorporated during *de novo* contig assembly (paired end reads) and subsequent scaffolding (mate pair reads) using the ALLPATHS-LG assembler (Broad Institute, Cambridge, MA; (Gnerre *et al.* 2011) with default parameters (Butler *et al.* 2008). The quality of the gene space assembled was further assessed using the Core Eukaryotic Genes Mapping Approach (CEGMA; (Parra *et al.* 2007), as well as a Benchmarking Universal Single-Copy Orthologue (BUSCO) analysis (Simão *et al.*, 2015, Waterhouse *et al.*, 2017). CEGMA was initially developed to identify orthologues within a newly assembled genome, for a set of 248 conserved, core eukaryotic genes/proteins from six well described species genomes (*Homo sapiens*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*). BUSCO was subsequently developed on the same principle but expanding the core gene sets (or BUSCO orthologous groups) to include additional genes from bacteria (6 653 BUSCO groups), eukaryotes (66 391 BUSCO groups), embrophyta (1 440 BUSCO groups) and protists (449 BUSCO groups) to more globally represent

different species clades (Simão *et al.*, 2015). These conserved orthologs aid in the quantitative measurement of the completeness of a newly assembled genome (in terms of expected gene content), by providing a comprehensive gene predictor training set to use during the genome annotation. For the current assembly, BUSCO analysis was performed against the embryophyte BUSCO groups, which specifically represent different plant families.

***De novo* Transcriptome Assembly**

The raw sequence read data (mean length 99 bp) from root (84.5 Gb), shoot (81.7 Gb) and leaf (78.4 Gb) material were pooled into a single dataset for downstream analysis (244.6 Gb). A total of 143.2 Gb (58%) reads were retained after the Trimmomatic software removed barcoding/indexing primers and FastQC filtering was applied (Table 3.1). Trinity version 2.0.6 (Grabherr *et al.* 2011) was used for *de novo* transcriptome assembly with default parameters (contig length $\geq$ 300 bp, k-mer size 25). The software package, TopHat, uses the high throughput short read mapping aligner Bowtie2 (Burrows-Wheeler indexing) to map transcriptomic reads back to the *de novo* assembled transcriptome, and analyses the mapping results to validate the initial assembly and to identify exon splice junctions (Langmead and Salzberg 2012; Trapnell *et al.* 2012). To determine the rate of transcriptomic fragment mapping during the current analysis, TopHat was executed with the following parameters: (i) read gap length of 5 bp (ii) intron size $\geq$ 20 bp (iii) minimum of 20x read coverage for any given intron. RNA-Seq by Expectation-Maximization (RSEM) was employed to estimate transcript abundance through alignment-based quantification methods, using Bowtie2 to align reads to the initial transcript assembly. The RSEM alignment files were further filtered to retain only fragments with a Fragments Per Kilobase of transcript per Million (FPKM) value of above two, to remove possible artefacts and lowly expressed transcripts.

Genome Annotation

The Maker-P(lant) annotation pipeline was used to predict gene structures within the *A. tricolor* genomic scaffolds. Due to the computationally intensive nature of Maker-P, the analysis was performed on a high-performance computing cluster at the Texas Advanced Computing Centre through the iPlant Collaborative

Cyberstructure (www.iplantcollaborative.org) with Message Passing Interface (MPI) across 16 computer nodes/processors for increased scalability.

RepeatMasker is a built-in add-on to the Maker-P pipeline and was used to identify both simple (stretches of homopolymer regions) and complex (transposons/retrotransposons) repeats that match entries in the RepBase repeat library (Bao *et al.* 2015). The fraction of the assembly represented by repeat sequences was determined based on the Repeatmasker output and the repeats were subsequently hard (complex repeats) or soft masked (simple repeats).

Initially, RNA-Seq transcript data for *A. tricolor* and the SwissProt protein database (www.ncbi.nlm.nih.org, release 2011-2012) were aligned to the genomic assembly as evidence, to identify potential protein models and their splice sites. Considering the evidence, the *ab initio* gene predictor Semi-HMM-based Nucleic Acid Parser - SNAP, (Korf 2004) is trained for the best-supported gene models based on the quality metric Annotation Edit Distance (AED) scores. Subsequently, Maker-P was initiated a second time, incorporating the trained SNAP hidden Markov Models (HMMs) for gene prediction, the assembled genomic and transcriptomic scaffolds, as well as the SwissProt protein database. Output files from the analysis included a GFF3 file (with all gene predictions, evidence data and quality metrics), as well as transcript and protein FASTA files for the predictions. Functional analysis of the protein sequences was performed using InterProScan version 68, with default predictive models and parameters. InterProScan utilizes the latest releases of the UniProtKB (TrEMBL and Swiss-Prot) to construct a holistic database of representative protein signatures, which were searched for within the *A. tricolor* protein dataset. In addition to predicting the protein functional domains, the Gene Ontology Annotation group (consisting of 33 944 GO terms mapped to InterPro entries) was used to assign Gene Ontology (GO) terms to the protein sequences (Jones *et al.*, 2014; Finn *et al.*, 2017).

BUSCO analysis, using the Embryophyta lineage dataset (1 440 BUSCO groups representing 31 plant species) as reference, was performed on the *A. tricolor* predicted proteins, to quantitatively assess the completeness of the genome/transcriptome assembly.

Results and Discussion

De novo Genome Assembly

The whole genome of a traditional leafy vegetable, *Amaranthus tricolor*, was sequenced from fresh leaf tissue using the NGS Illumina HiSeqSQ platform. A total of 135.5 Gb reads were generated across the paired end (93.8 Gb, mean length 99 bp) and long mate pair (41.5 Gb, mean lengths of 4 306 bp (5 Kb library) and 2 308 bp (3 Kb library)) (Table 3.1). Approximately 1.6% of the reads were removed from downstream analysis due to low frequency k-mers. A further 7% of reads were removed that had mapped to the chloroplast genome of *Beta vulgaris*, a close relative of *A. tricolor* within the family Amaranthaceae. The ALLPATHS-LG assembler estimated the genome size at 402 Mb, which had an average of 44x coverage by all the read fragments. During previous studies, the genome sizes of several diploid *Amaranthus* species (*A. caudatus*, *A. cruentus*, *A. hybridus*, *A. hypochondriacus*, *A. powellii*, *A. quitensis*, *A. retroflexus*, *A. spinosus* and *A. viridis*, n=16 or n=17) were estimated at approximately 500 Mb (Radwan *et al.* ; Bennett and Smith 1991; Clouse *et al.* 2016; Stetter *et al.* 2016; Stetter and Schmid 2017). However, diploid *A. tricolor* (n=17) was estimated to have a genome size of approximately 780-800 Mb (Bennett and Smith 1991; Stetter and Schmid 2017).

It has been hypothesized that species within *Amaranthus* subgenus *Albersia* (*A. tricolor*, *A. blitum*) tend to have larger genome sizes, compared to *Amaranthus* subgenus *Amaranthus* (*A. caudatus*, *A. cruentus*, *A. hypochondriacus*, *A. hybridus*, *A. quitensis*, *A. retroflexus*, *A. powellii*), possibly due to historical whole genome duplication (WGD) events, a higher percentage of repetitive regions, hybridization, fusion events and/or chromosome losses (Clouse *et al.* 2016; Stetter and Schmid 2017). The ancestral genome for the *Amaranthus* genus has been estimated at 569 Mb, but with a large standard deviation (± 200 Mb), indicating that the genome sizes of all extant species still fall within these measurements (Stetter and Schmid 2017). The estimated genome size for *A. tricolor* in current study was 402 Mb, representing approximately 45%-52% of the total genome size. The total contig size (based on paired end data) was 275.7 Mb, assembled into 68 656 contigs with approximately 212 contigs per Mb and an N50 of 6 kb (Table 3.1). Eighty percent of the 3 kb and 78.6% of the 5 kb mate pair libraries mapped to their respective length intervals of 2-4 kb and 4-8 kb.

Table 3.1: Read quality control (QC) for genomic and transcriptomic data

Library type	Raw reads	Reads after QC	Percentage retained after QC	Mean read length
Genomic (Gb)	242.1	135.5	56%	
Paired End (Gb)	158.6	93.8	59.1%	99 bp
Long Mate Pair (Gb)	83.5	41.5	49.7%	2 308 bp (3 Kb library) 4 306 bp (5 Kb library)
Transcriptomic (Gb)	244.6	143.4	58.60%	
Leaves Paired End (Gb)	78.4	47.1	60.10%	98-99 bp
Roots Paired End (Gb)	84.5	45.1	53.40%	98-99 bp
Shoot Paired End (Gb)	81.7	51.2	62.70%	98-99 bp

Further scaffolding of the contigs using the long mate pair data resulted in 323 Mb (80.3% of the estimated genome size) captured in the entire scaffold assembly (largest scaffold 279 Kb) and approximately 42 scaffolds per Mb. Gapped regions represented 15.72% of the assembly, with an average gap size of 616 bp (standard deviation = 115 bp). Seven ambiguous bases were identified in every 10 kb of sequence data (comprising a total of 32% in the entire scaffold assembly). The N50 of the scaffold assembly was approximately 51 kb, with 1 831 scaffolds being of this size and larger. The frequencies of the different scaffold sizes are shown in Figure 3.4. The amount of read data generated during the present assembly (234 Mb) was less than for the two previously assembled *A. hypochondriacus* genomes (869 Mb - (Sunil *et al.* 2014); 106 Gb - (Clouse *et al.* 2016), with 44x coverage compared to 106x and 229x respectively (Table 3.2). However, for detecting large genome mutations, SNP's and genome rearrangements, generally a genome coverage of 10x to 30x is sufficient for most studies (www.illumina.com).

Table 3.2: Summary of whole genome sequencing assemblies of *A. tricolor* compared to previously assembled *A. hypochondriacus* and *Beta vulgaris*

	<i>A. tricolor</i>	<i>A. hypochondriacus</i> assembly	<i>A. hypochondriacus</i>	<i>Beta vulgaris</i> assembly
Total number of reads (bp)	234 727 730	869 113 408	106 689 078 000	568 984 731
Estimated genome size	402	-	431.8	731
Coverage	44	106	229	30.2
GC Content	37.3	34	34.7	33.71
Contigs				
Total Contig Size (bp)	275 699 291	-	357 417 249	491 778 857
N50 Contig Length (Kb)	6	1.8	32	17
Scaffolds				
Total Scaffold Size (bp)	323 827 803	645 211 960	376 726 029	568 984 731
N50 Scaffold Length (Kb)	51	35	371	168
% Estimated to be	54	13.76	47.7	42.3

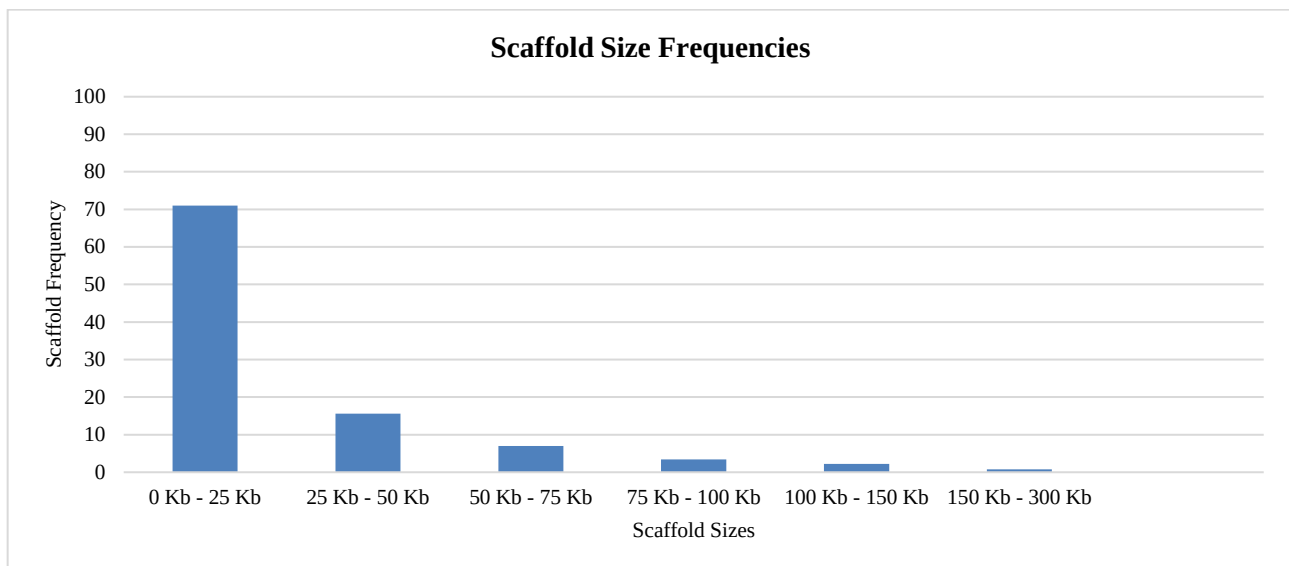


Figure 3.4: *Amaranthus tricolor* scaffold size frequencies. Approximately 70% of the scaffolds were of sizes between 0 Kb and 25 Kb, while only 3% of scaffolds were 100Kb and larger.

The GC content was highly comparable between the three assemblies of *A. tricolor* (37.3%), and *A. hypochondriacus* (34% and 34.7%) (Table 3.1) and was highly similar to the genome GC content of other angiosperm species (Šmarda *et al.* 2012). The ALLPATHS-LG and RepeatMasker software estimated that approximately 54% of the *A. tricolor* genome is composed of repetitive elements, which is significantly higher than for *A. hypochondriacus* (36%) and could likely be an additional explanation for the larger genome size of *A. tricolor* (880 Mb, 2n=34) vs *A. hypochondriacus* (466 Mb, 2n=32) (Clouse *et al.* 2016). A total of 72 631 (39% of the total repeat fraction) simple sequence repeat (SSR) loci were identified across the scaffold assembly. The most common repeats were trinucleotides (37.1%), followed by tetranucleotide (25.3%), dinucleotide (22.1%), pentanucleotides (9.5%) and hexanucleotides (6.3%). The most common SSR motifs were (AT/TA; 41.5%) followed by TC (7.3%), corroborating previous reports that the most common motifs across the amaranth genomes are AT-repeats (Mallory *et al.* 2008). The SSR loci can be further investigated for their inter- and intra-species polymorphic content within the *Amaranthus* genus, to facilitate the development of informative markers for in depth population studies and to determine genetic relatedness between the different species (Chapter 4).

Of the remaining repeat fraction within the genome assembly, 4% could be assigned to Long Interspersed Nuclear Elements (LINE's), 0.3% to Short Interspersed Nuclear Elements (SINE's) 10% to low complexity homo-polymer regions and 23% to Long Terminal Repeat (LTR) regions (Table 3.3). This leaves approximately 23.8% unclassified repetitive regions with no representation in the RepBase database of RepeatMasker and is comparable to the number of unclassified repeat regions identified during the previous genome assembly of *A. hypochondriacus* (Clouse *et al.* 2016).

Table 3.3: Summary of repetitive regions within the *A. tricolor* genomic scaffolds

	Number of	Percentage of total Repeats
Long Interspersed Nuclear Elements (LINE)		
LINE/L1	2 586	
LINE/RTE	3 888	
Total	7 575	4%
Long Terminal Repeats (LTR)		
LTR/Copia	24 317	
LTR/Gypsy	10 798	
Total	43 117	23%
Short Interspersed Nuclear Elements (SINE)	566	0.3%
Low Complexity Repeats	18 357	10%
Simple Repeats	72 631	39%

The alignment of the protein sequences in a predefined set of 248 core eukaryotic genes (CEGs) within the CEGMA database, to candidate protein regions in the current assembly (identified through tBLASTn analysis) was investigated to determine the completeness of the gene space assembled. In total, 191 (77%) complete CEGs (with an alignment score of $\geq 70\%$) and 237 (95%) incomplete CEGs (with an alignment score of $\leq 70\%$) could be identified within the current genome assembly. In addition, 51% of the complete CEGs and 64.98% of the partial CEGs were found to have more than one orthologue in the assembly, with an average of 2 orthologues per CEG. The CEGMA output compares favorably to the output from the *A. hypochondriacus* assembly where 217 (87.5%) complete CEGs were identified (Clouse *et al.* 2016). As in the present study, it was found that an average of 2.3 orthologues is present per CEG. The presence of these orthologues likely represents a WGD event in the history of the *Amaranthus* genus. Clouse *et al.* (2016) estimated the duplication to have occurred between 36.7 and 67.9 Ma, possibly followed by stages of chromosomal loss or fusion events to yield the present day diploid chromosomal numbers of $2n=32$ (*Amaranthus* subgenus *Amaranthus*) or $2n=34$ (mostly *Amaranthus* subgenus *Albersia*). Since CEGMA only identifies genes inside a contig/scaffold, the chance of “missing” a gene due to a split across contigs/scaffold may lead to an inaccurate

representation of CEG alignments. Additional sequencing data will enhance the contig assembly, and subsequently a higher number of CEG's could be identified.

In addition, a BUSCO analysis was performed which incorporates a larger set of lineage-specific orthologues for comparisons compared to CEGMA. The genomic scaffolds of *A. tricolor* were searched against the Embryophyta lineage dataset of BUSCO (1 440 BUSCO groups representing 31 plant species). The analysis revealed the presence of 1 272 (88.4%) complete orthologue genes in the genome scaffolds, while 43 (3%) and 125 (8.6%) were fragmented or missing, respectively (Figure 3.5). Of the complete genes identified, 1 222 (96%) were present as single copy and 50 (3.9%) were duplicated within the assembly. Duplicated copies of the genes are expected to have a low percentage (<5%), since the genes within the BUSCO database were chosen based on their single-copy presence within a genome (Waterhouse *et al.*, 2017).

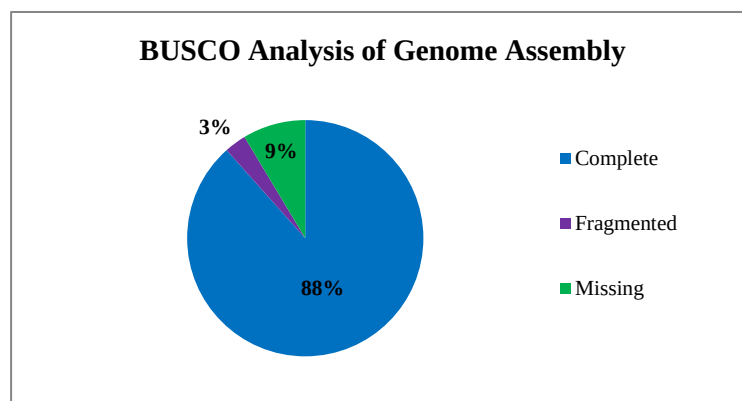


Figure 3.5: BUSCO analysis of the genome assembly. Approximately 88% of the orthologue genes from the BUSCO database were present as full-length genes in the assembly. Only 9% of orthologues were missing in the assembly, while 3% were fragmented.

***De novo* Transcriptome Assembly**

Transcriptomic sequencing libraries were prepared from leaf, shoot and root tissues of *Amaranthus tricolor* plants. Approximately 244.6 Gb sequencing data was generated across the three tissues (leaves - 78.4 Gb, roots - 84.5 Gb, shoots – 81.7 Gb, Table 3.2). After Trimmomatic was applied to remove barcoding indexes, and FastQC further discarded reads of low quality (PHRED<20), approximately 143.4 (58%) Gb data (leaves

– 47.1 Gb, roots – 45.1 Gb, shoots – 51.2 Gb) remained for downstream analysis. TopHat (utilizing the high-throughput short read aligner Bowtie2) was used to map the transcriptomic reads back to the genome assembly, to assess the splice junctions between exons. In total, 48.3% and 39.8% of the left and right sequence reads mapped to the genome assembly, with a 44.1% overall mapping rate. Of these reads, 2.8% had multiple alignment sites, 1.9% were discordant alignments and 36% had a concordant pair alignment rate. The lower concordant pair alignment could have been influenced by the sequence trimming parameters, since shorter reads have lower specificity during mapping. Increasing the minimum read length parameter for trimming, together with an increase in the number of sequences generated should increase the number of concordant pair alignments. In contrast, the discordant alignments (reads for which the relative orientation of forward/reverse or insert size is not correct) has a low percentage, indicating that the size selection of fragments during library preparation was optimal. Multiple alignments of reads are considered normal, due to the presence of gene orthologues, artefacts of genome multiplication events or the presence of pseudogenes contained within a genome.

The pooled transcriptomic reads from different tissue types were *de novo* assembled using the Trinity software program. The initial assembly yielded 553 585 transcripts, with a contig N50 size of 790 bp and a median contig length of 529 bp. After determining the relative abundance of each fragment, the data was further filtered on FPKM >1 and FPKM >2 (Figure 3.6). Since the transcriptome assembly for *A. tricolor* was only used for preliminary genome annotations (and not specific transcript differential expression analysis), an FPKM filter was applied to remove any transcript artefacts or low abundance/lowly expressed genes from the assembly. After filtering on FPKM >1, the number of genes and transcripts were reduced by 36% and 46% respectively, while further filtering with FPKM >2 reduced the number of genes and transcripts by an additional 36% and 30% respectively (Figure 3.6). However, the removal of these low transcript fragments does not significantly influence the contig N50 size, median contig length and average contig size. These results indicate that a large portion of the transcriptome dataset is composed of reads that are lowly represented, either because the sequencing was too shallow to discern these genes in significant quantities, or because these transcripts are preferentially expressed in different tissue types or under different environmental conditions. The addition of alternative tissue types (such as seed, flowers) as well as different time points

during growth of these plants could reveal an increase in low abundance transcripts identified during the current study.

After FPKM>2 filtering, the number of contigs (79 380), contig N50 (1 036 bp) and average contig length (833 bp) of *A. tricolor* compares favorably with the results for two previously assembled transcriptomes of *A. hypochondriacus*, where 57 658 contigs with an N50 of 3 279 bp (Sunil *et al.* 2014) and 66 370 contigs, an N50 of 1 491 bp and an average contig length of 864 bp (Clouse *et al.* 2016) were obtained (Table 3.4).

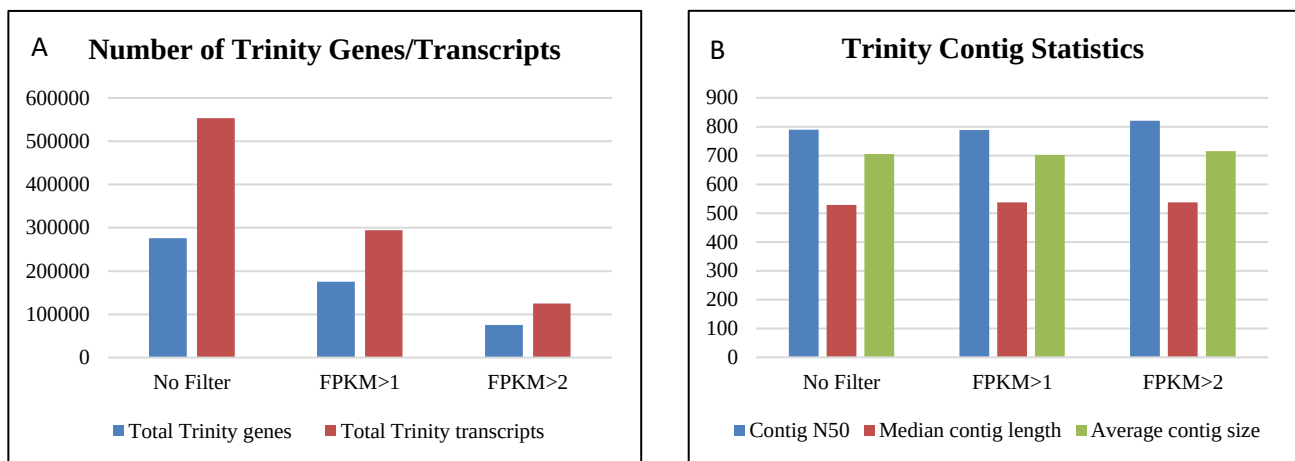


Figure 3.6: The number of trinity genes/transcripts (A) and contig statistics (B) of unfiltered trinity statistics compared to filtering of FPKM >1 and FPKM >2. The filtering removes a significant amount of low abundance transcripts (A), while at the same time keeping the contig N50, median contig length and average contig size relatively constant (B).

Table 3.4: Summary of *de novo* transcriptome sequencing assemblies of *A. tricolor* compared to previously assembled *A. hypochondriacus*

	<i>A. tricolor</i> assembly (FPKM > 2)	<i>A. hypochondriacus</i> assembly (Sunil <i>et al.</i> 2014)	<i>A. hypochondriacus</i> assembly (Clouse <i>et al.</i> 2016)
Number of reads (Gb)	9.1	8.6	57.3
Contig N50	1 036	3 279	1 491
Maximum contig size (bp)	14 446	17 471	12 410
Average contig size (bp)	833	-	864
% GC Content	37.91	38	39

Genome Annotation

The genomic and transcriptomic assemblies, in conjunction with the SwissProt protein database, were used as evidence during genomic scaffold annotation of *A. tricolor* using the MAKER annotation pipeline. Using quality values based on protein evidence, the MAKER software aligned expressed sequence tags (EST's) and

proteins from the GenBank NCBI database to the genome and created *ab-initio* gene predictions using SNAP and Augustus (Stanke and Waack 2003). Once the SNAP program has been trained for the best-supported gene models, MAKER was re-initiated using these gene predictions as further evidence during annotation. A total of 16 495 gene models (genes/CDSs) were identified, with a mean length of 4 461 bp. Of these, 94.3% correspond to the transcript sequences identified during the *de novo* transcript assembly. The extra non-genic transcripts that do not contain gene/CDS regions (23% of total filtered transcripts) could be ascribed to (i) being non-coding RNA transcripts that function at the transcriptional or post-transcriptional level to regulate gene expression or (ii) could indicate the presence of incorrectly assembled transcript fragments. Approximately 24% of the genome is covered by gene regions, with a mean of 5 exons and 4 introns per gene (Table 3.5). To further assess the quality of the assembled genome, the annotation edit distance (AED) score of each gene model was calculated by evaluating the quality of the annotation given the RNAseq evidence. A value of 0 indicated perfect agreement with the evidence, while a value of 1 indicated no evidence to support the annotation. An average AED score of 0.24 was obtained for the current annotation, where 63% of the gene predictions had an AED value of less than 0.25 (Figure 3.7). The AED score can subsequently identify scaffolds with problematic annotations or measure the changes to the quality of annotations across annotation optimization protocols.

Table 3.5: Genome annotation statistics of *A. tricolor*

Total sequence length (Mb)	306
Number of genes	16 495
Number of exons	84 907
Number of introns	68 413
Number of CDS	16 495
CDS: Complete	14 178
CDS: Start codon, no stop codon	1 158
CDS: Stop codon, no start codon	703
Total gene length (Mb)	73
Total exon length (Mb)	20
Total intron length (Mb)	53
Total CDS length (Mb)	17
Shortest gene (bp)	180
Longest gene (bp)	32 878
Mean gene length (bp)	4 461
Mean exon length (bp)	240
Mean intron length (bp)	780
Mean CDS length (bp)	1 077
% of genome covered by genes	24
% of genome covered by CDS	5.8
Mean mRNA's per gene	1
Mean exons per gene	5
Mean introns per gene	4

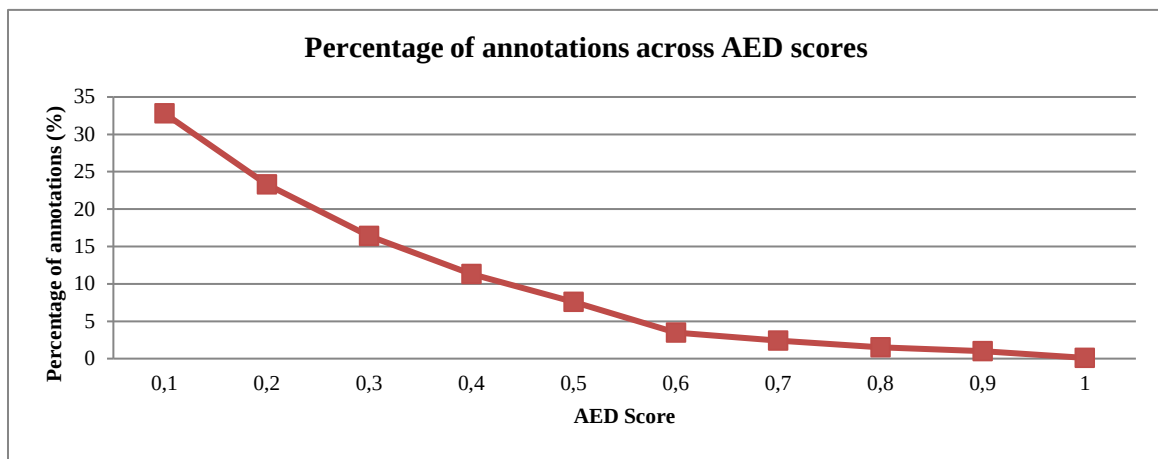


Figure 3.7: Percentage of scaffold annotations with correlating AED scores. Approximately 63.3% of genome scaffold annotations have an AED score of < 0.25 , which indicates annotations that are highly supported by homology-based or transcript evidence.

BUSCO analysis of the protein dataset revealed the presence of 763 (53%) complete matches, while 169 (12%) and 508 (35%) were fragmented and missing, respectively (Figure 3.8). Of the complete BUSCO genes recovered, 95% are single copy and 4.9% are duplicated within the dataset.

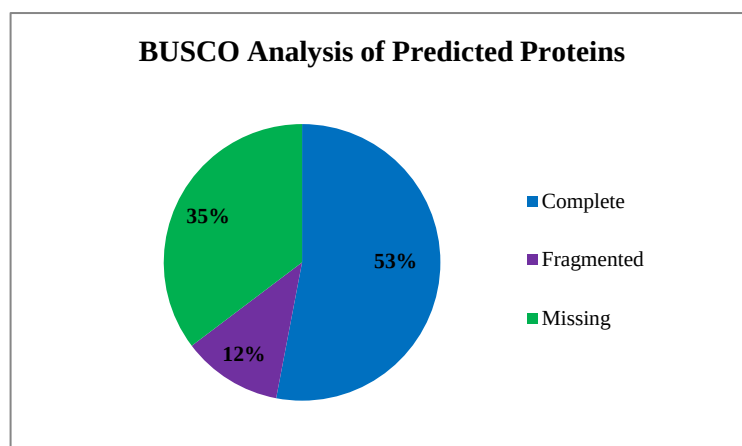


Figure 3.8: BUSCO analysis of *A. tricolor* protein predictions. Approximately 53% of the orthologue genes from the BUSCO database were present as full-length genes in the amaranth dataset. Additionally, 35% of orthologues were missing in the assembly, while 12% were fragmented.

The embryophyta lineage dataset of BUSCO encompasses representative core genes from 31 different plant species from 11 families, including Poaceae (51.6%), Brassicaceae (9.7%), Solanaceae (6.5%), Fabaceae (6.5%), Vitaceae, Selaginellaceae, Rosaceae, Salicaceae, Funariaceae, Musaceae and Amborellaceae (each with 3.2%). The lower number of complete orthologues that were found from the *A. tricolor* protein dataset (53%) compared to the genome assembly (88%) could possibly be ascribed to (a) the lack of Amaranthaceae orthologue/parologue genes in the Embryophyta lineage database, (b) low quality protein sequence alignments due to initial low-quality sequence read assembly of the *A. tricolor* genome or (c) the representative sequences for *A. tricolor* is too divergent or complex to be identified by BUSCO. The most likely scenario would probably be insufficient sequencing data for the assembly of rare, low copy gene regions. Additional deep sequencing (as well as long range sequencing) would improve the assembly to such an extent that complete orthologues from the genome could be assembled and annotated. Alternative splicing of transcript fragments between plant species groups could lead to the formation of different protein morphologies than what is represented in the BUSCO database. With the addition of phylogenetically diverse, species-level coverage for plant genes, the BUSCO analysis will likely indicate a higher percentage of completeness of the protein dataset.

Approximately 94% of the 16 495 potential coding regions had significant alignments to the NCBI SwissProt protein database (average alignment length of 713 bp, e-value < 0.0001). Nucleotide BLAST analysis of the genes/CDS regions revealed a high percentage of homology to genes of *A. hypochondriacus* (79%), *B. vulgaris* (68%) and *A. palmeri* (36%) (Figure 3.9).

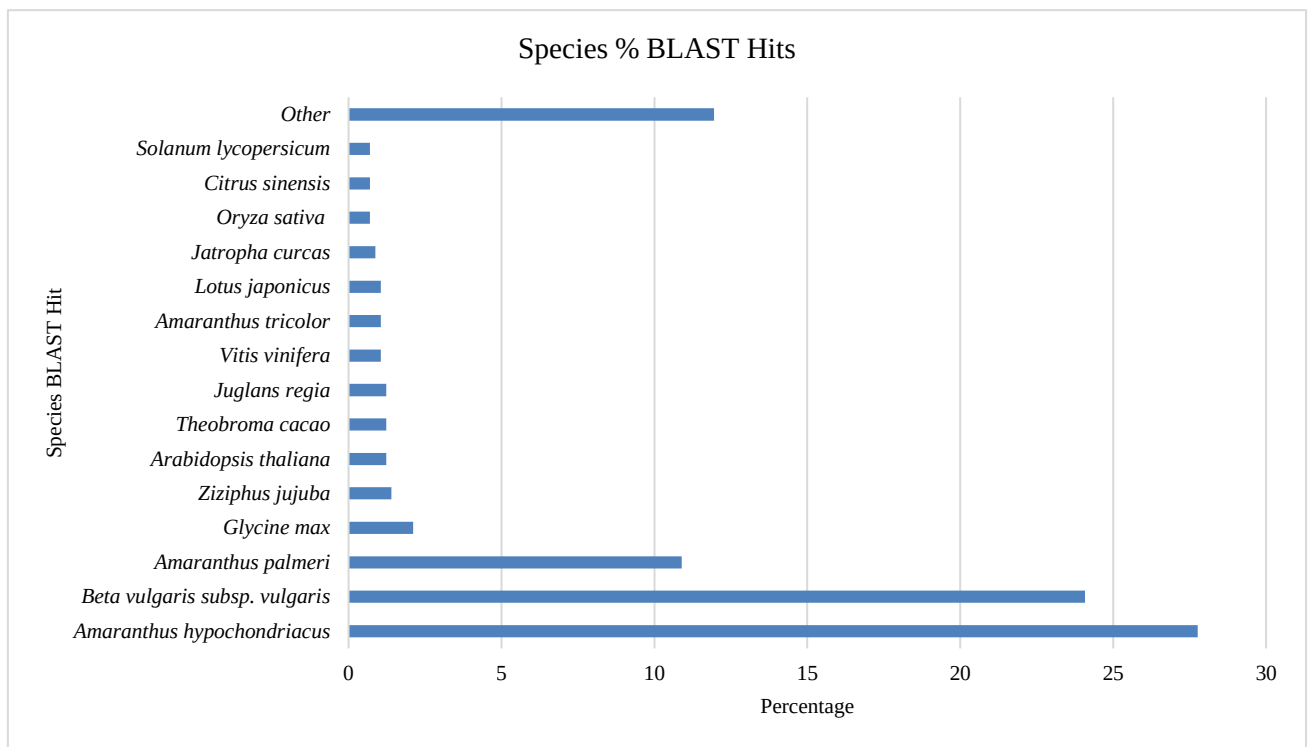


Figure 3.9: Top % species nucleotide BLAST hits against the SwissProt database. The highest number of BLAST hits were to *A. hypochondriacus* (79%), *B. vulgaris* (68%) and *A. palmeri* (36%).

In total, 13 895 (84%) CDS regions could be assigned 5 401 unique InterProScan annotations and approximately 5% could be mapped to Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathways. Uncharacterized proteins or proteins that could not be assigned a known function or domain made up 4% of the total protein set. The top InterProScan annotations assigned to the annotated proteins is presented in Figure 3.10.

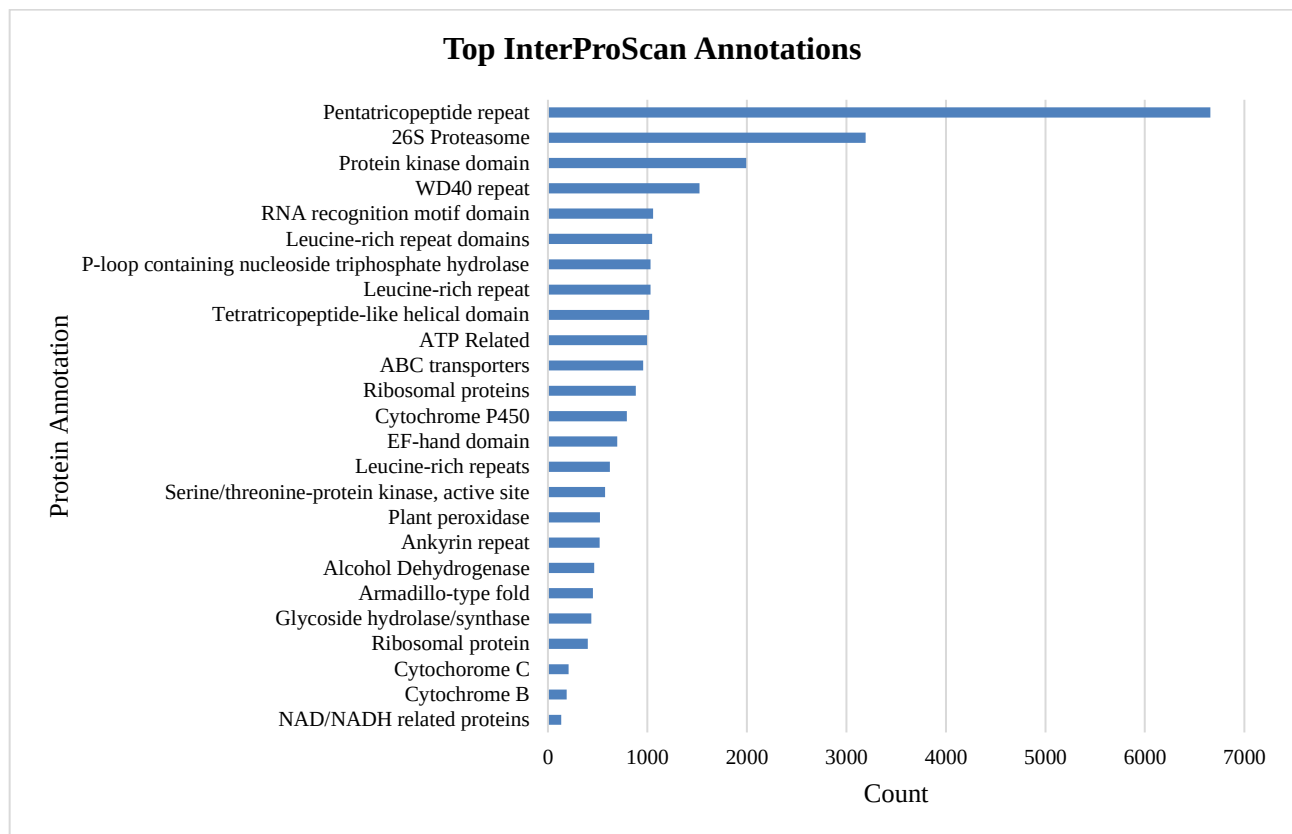


Figure 3.10: Representation of the top 25 protein annotations found within the *A. tricolor* genome. InterProScan analysis revealed that 84% of the predicted CDS regions could be assigned a protein signature/domain from the UniProtKB (TrEMBL and Swiss-Prot) database. Approximately 4% of the protein sequences could not be assigned a known functional domain.

The most common functional domain identified within the *A. tricolor* assembly is pentatricopeptide repeats (PPR), which is known to play a role during the post transcriptional regulation in plant organelles, but their exact function remains unknown (McCormick, 2018). The top InterProScan annotations represent mostly common genome housekeeping genes such as proteasomes (19.9%), protein kinases (12.4%), ATP related proteins (6.2%) and ABC transporters (6%). Several interesting protein targets for future enhanced breeding was identified from the protein dataset. Approximately 10% of the protein annotations were related to plant defence, including plant peroxidases, RPW8 (powdery mildew resistance genes) and EDR2 (enhanced disease resistance 2 genes), which are all known to play a significant role during the initiation of plant defence mechanisms (cell death or the hypersensitive response) against microbial or fungal pathogens in various plant species (Meyers *et al.*, 2005; Vorwerk *et al.*, 2005; Jalaja, 2011; Hu *et al.*, 2018).

Since resistance to abiotic stresses are of particular importance when growing crops in a changing climatic environment, the identification of a stress response protein (NST1) provides a putative initiation point for research in the way *A. tricolor* responds to environmental stresses. The NST1 gene has been shown to increase the secondary cell wall thickness in *Arabidopsis thaliana* plants, which increased their tolerance to drought and high salt conditions (Mitsuda *et al.*, 2005). In addition, the NAC gene family (which includes the NST1) gene was identified to be highly expressed in leaf material of cotton (*Gossypium hirsutum*), which also increased the plants' resistance to drought and high salt conditions (Meng *et al.*, 2009). Protein homology studies to *A. tricolor* could enable researchers to develop a genotype which displays a higher tolerance to drought and salt stress than native genotypes.

Gene ontology terms could be assigned to 55.4% of the proteins with InterProScan annotations. Of these, 11% were biological process, 6% to cellular component and 83% to molecular functions (Figure 3.11).

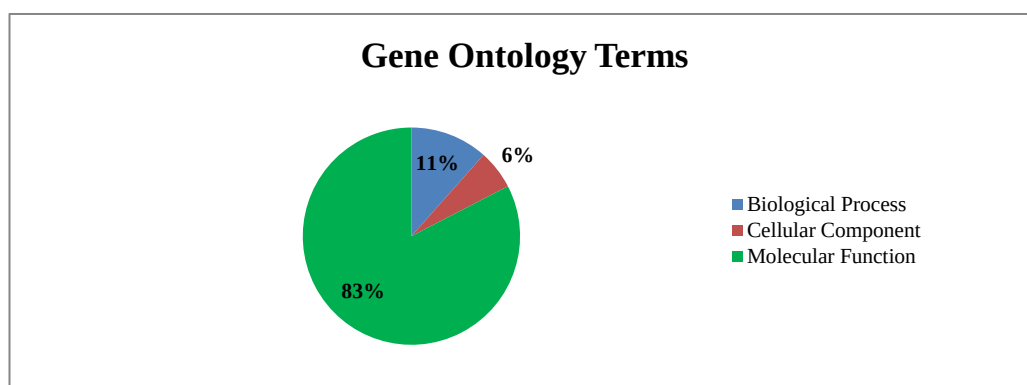


Figure 3.11: Gene ontology terms could be assigned to 55.4% of the identified proteins from the *A. tricolor* genome. Approximately 83% belonged to molecular functions, 11% to biological process and 6% to cellular components.

The most dominant molecular functions included ATP binding (17%), protein binding (16%), oxidoreductase activity (7%), heme binding (6%) and DNA binding (5%). A high percentage of biological processes were assigned to oxidation-reduction processes (9%), protein phosphorylation (8%), protein kinases (6%), regulation of transcription (6%) and response to oxidative stresses (5%). The most common cellular

components were membrane (3%), intracellular (3%), ribosome (3%), nucleus (3%) and cytoplasm (1.3%) (Figure 3.12).

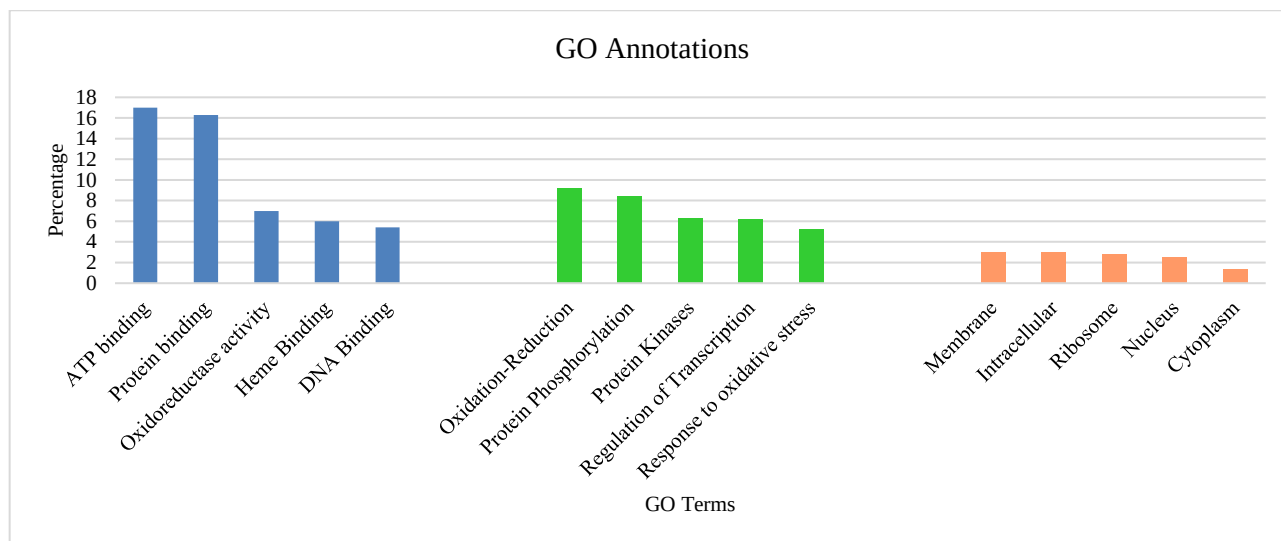


Figure 3.12: Most dominant GO terms identified for molecular function (blue), biological process (green) and cellular components (orange).

The GO term assignments were highly comparable to the results obtained for the previously published genome of *A. hypochondriacus*, where protein binding (molecular function), oxidation-reduction (biological process) and membrane (cellular components) were most dominant (Clouse *et al.*, 2016).

Conclusion

The draft genome and transcriptome assembly of *A. tricolor* reported here is a first step to the development of a full-length, completely annotated genomic assembly for a leafy amaranth accession, which can be used to develop different genomic tools for future amaranth research. These include (i) facilitating the availability of annotated gene regions for targeted genetic manipulation or exploitation (ii) the identification of potential SSR loci to investigate genetic diversity within the genus (Chapter 4) and (iii) the extraction of chloroplast sequences for phylogenetic analysis (Chapter 5). The results from this chapter supports the hypothesis that a good-quality draft genome assembly could be produced using Illumina sequencing data.

The genome can simplify fine mapping of quantitative trait loci, especially if more information on the chromosome physical maps become available. The genome and transcriptome are also a first step toward the identification of genetic variants to characterize gene functions with respect to unique leafy amaranth profiles of nutritional composition, yield, biotic and abiotic stresses, as well as for species genotyping and phylogenetic analysis. Speciation barriers such as chromosomal rearrangements and genome duplication events could also shed light in the future breeding programs associated with *Amaranthus* species to produce superior lines for commercial breeding.

Further improvements to the *A. tricolor* assembly could be made by utilizing single molecule, long read technology (Pacific Biosciences), as was done for the *A. hypochondriacus* draft genome assembly (Clouse *et al.* 2016; Lightfoot *et al.* 2017). Furthermore, chromatin interaction mapping (Hi-C) can be employed to consolidate scaffolds and contigs, to produce an eventual chromosome scale assembly. Physical maps together with the genetic maps would facilitate the efficient identification of the precise genomic location of candidate genes involved in agriculturally driven targets such as yield, lysine production or drought resistance pathways.

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
- 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 Sciences* 201:321-329
- Bansah KJ, Addo WK (2016) Phytoremediation potential of plants grown on reclaimed spoil lands. *Ghana Mining Journal* 16:68-75
- Bao W, Kojima KK, Kohany O (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* 6:11
- Bennett M, Smith J (1991) Nuclear DNA amounts in angiosperms. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 334:309-345
- Butler J *et al.* (2008) ALLPATHS: *de novo* assembly of whole-genome shotgun microreads. *Genome Research* 18:810-820
- Clouse J *et al.* (2016) The amaranth genome: genome, transcriptome, and physical map assembly. *The Plant Genome* 9
- Conesa A, Götz S (2008) Blast2GO: A comprehensive suite for functional analysis in plant genomics. *International Journal of Plant Genomics* 2008
- Deschamps S, Campbell MA (2009) Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. *Molecular Breeding* 25:553-570
- Dohm JC *et al.* (2014) The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature* 505:546-549
- Finn RD *et al.* (2016) InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Research* 45: D190-D199.
- Gnerre S *et al.* (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences* 108:1513-1518
- Goecks J, Nekrutenko A, Taylor J (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* 11:R86

- Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* 17:333-351
- Grabherr MG *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 29:644-652
- Grubben G (1980) Germplasm of vegetable and grain amaranths. *Plant Genetic Resources Newsletter*: 33-35
- Gupta S, Jyothi Lakshmi A, Manjunath MN, Prakash J (2005) Analysis of nutrient and antinutrient content of underutilized green leafy vegetables LWT. *Food Science and Technology* 38:339-345
- Hashmi U *et al.* (2015) Plant exomics: concepts, applications and methodologies in crop improvement. *Plant Signaling & Behavior* 10:e976152
- Hu Y *et al.* (2018) Ectopic expression of *Arabidopsis* broad-spectrum resistance gene RPW8.2 improves the resistance to powdery mildew in grapevine (*Vitis vinifera*). *Plant Science* 267: 20-31
- Jalaja, N (2011) Development of transgenics for fungal resistance and discovery of chemically induced mutations in Pearl Millet (*Pennisetum glaucum* L.) population by TILLING. PhD dissertation, Osmania University.
- Jansen P, Grubben G, Denton O, Messiaen C, Schippers R, Lemmens R, Oyen L (2004) *Amaranthus hypochondriacus* L. *Plant Resources of Tropical Africa* 2:78-80
- Jones P *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30.9: 1236-1240.
- Kaur P, Gaikwad K (2017) From Genomes to GENE-omes: Exome sequencing concept and applications in crop improvement. *Frontiers in Plant Science* 8:2164
- Korf I (2004) Gene finding in novel genomes. *BMC Bioinformatics* 5:59
- Kunyanga CN, Imungi JK, Vellingiri V (2013) Nutritional evaluation of indigenous foods with potential food-based solution to alleviate hunger and malnutrition in Kenya. *Journal of Applied Biosciences* 67:5277-5288
- Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9:357-359
- Levy SE, Myers RM (2016) Advancements in next-generation sequencing. *Annual Review of Genomics and Human Genetics* 17:95-115
- Lightfoot D, Jarvis D, Ramaraj T, Lee R, Jellen E, Maughan P (2017) Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (*Amaranthus hypochondriacus*) chromosomes provide insights into genome evolution. *BMC Biology* 15:74

- Lushbough C, Bergman MK, Lawrence CJ, Jennewein D, Brendel V (2010) BioExtract server—an integrated workflow-enabling system to access and analyze heterogeneous, distributed biomolecular data IEEE/ACM. Transactions on Computational Biology and Bioinformatics 7:12-24
- Mallory MA, Hall RV, McNabb AR, Pratt DB, Jellen EN, Maughan PJ (2008) Development and characterization of microsatellite markers for the grain amaranths. Crop Sciences 48:1098-1106
- McCormick S (2018) Binding sites for pentatricopeptide repeat proteins differentially activate chloroplast transgenes. The Plant Journal 94.1: 6-7.
- Meng C, Caiping C, Tianzhen Z, Wangzhen G (2009) Characterization of six novel NAC genes and their responses to abiotic stresses in *Gossypium hirsutum* L. Plant Science 176: 352-359.
- Mitsuda N, Motoaki S, Kazuo S, Masaru O (2005) The NAC transcription factors NST1 and NST2 of *Arabidopsis* regulate secondary wall thickenings and are required for anther dehiscence. The Plant Cell 17: 2993-3006.
- Morozova O, Marra MA (2008) Applications of next-generation sequencing technologies in functional genomics. Genomics 92:255-264
- Meyers BC, Shail K, Raja SN (2005) Evolving disease resistance genes. Current Opinion in Plant Biology 8: 129-134.
- Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE (2011) Landscape of next-generation sequencing technologies. Analytical Chemistry 83:4327-4341
- Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23:1061-1067
- Radwan SA-A, Badr S, Mira MM, Helmy R (2014) Genetic variability and phylogenetic relationships in *Amaranthus* using karyotype. Journal of Biotechnology Research 14:53
- Raza K, Ahmad S (2016) Principle, analysis, application and challenges of next-generation sequencing: a review arXiv preprint arXiv:160605254
- Ribeiro JE, Pieterse PJ, Famba SI (2017) Vegetative growth of *Amaranthus hybridus* and *Amaranthus tricolor* under different watering regimes in different seasons in southern Mozambique. South African Journal of Plant and Soil 34:201-210
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210-3212.

- Sitkey V, Gaduš J, Kliský L, Dudák A (2013) Biogas production from amaranth biomass. *Acta Regionalia et Environmentalica* 10
- Šmarda P, Bureš P, Šmerda J, Horová L (2012) Measurements of genomic GC content in plant genomes with flow cytometry: a test for reliability. *New Phytologist* 193:513-521
- Stanke M, Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19:ii215-ii225
- Stetter MG, Müller T, Schmid K (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
- Sunil M *et al.* (2014) The draft genome and transcriptome of *Amaranthus hypochondriacus*: A C4 dicot producing high-lysine edible pseudo-cereal. *DNA Research* 21:585-602
- Tomé D, Bos C (2007) Lysine requirement through the human life cycle. *The Journal of Nutrition* 137:1642S-1645S
- Trapnell C *et al.* (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* 7:562-578
- 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
- Vorwerk S *et al.* (2007) EDR2 negatively regulates salicylic acid-based defenses and cell death during powdery mildew infections of *Arabidopsis thaliana*. *BMC Plant Biology* 7: 35.
- Waterhouse, RM *et al.* (2017) BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution* 35:543-548

CHAPTER 4

SPECIES RELATIONSHIPS WITHIN THE *AMARANTHUS* GENUS REVEALED BY ILLUMINA MISEQ SEQUENCING OF MICROSATELLITE MARKERS

Abstract

The *Amaranthus* genus includes several important leafy vegetables, which are considered an emerging, sustainable source of nutritious food in most developing countries. Inadequate species classifications have hampered the identification of parental lines to use in breeding programs, to develop superior genotypes incorporating a multitude of important traits. The advent of next generation sequencing (NGS) has revealed immense potential for high throughput genotyping techniques at a fraction of the time and cost of more conventional capillary electrophoresis approaches. The aim of this study was to discover potentially polymorphic microsatellite markers from a draft genome sequence of *Amaranthus tricolor*, which could be used to genotype previously identified and unknown amaranth accessions on the Illumina NGS platform. The results of this study revealed that (i) a 570 Mb draft genome assembly (13 606 scaffolds) of *Amaranthus tricolor* enabled the identification of 671 microsatellites with potential as SSR markers, (ii) further validation of a marker subset revealed 6 robust SSR markers for inter- and intraspecies genotyping, (iii) the allele scores obtained from both gel based capillary electrophoresis (ABI) and next generation (Illumina MiSeq) platforms were highly comparable and (iv) downstream analysis of SSR markers screened with NGS revealed robust genetic relationships between known and unknown *Amaranthus* accessions, enabling the identification of unknown accessions. In conclusion, NGS is a highly feasible approach for SSR marker genotyping.

Introduction

The rapid growth in global human populations (estimated to increase from 7.5 billion in 2017 to 9.7 billion by 2050) has placed increasing pressures on plant science researchers to intensify their efforts in breeding of enhanced or alternative plant varieties to sustain global food demand (www.worldometers.info). Although several conventional plant breeding strategies exist for trait development and varietal enhancement, the popular approach of using molecular markers to identify and follow desirable traits, as well as to study genetic variation through a breeding population remains most effective in selecting enhanced varieties (Zalapa *et al.* 2012). The development of next generation sequencing technologies has not only increased sequence yields for genome studies but has also revealed the potential of high throughput marker screening (Vartia *et al.*, 2016).

Microsatellite markers are simple sequence repeats (SSRs) of di- tri- tetra- penta- or hexanucleotides, occurring mostly within an allele in non-coding genomic regions, and to a lesser extent in coding regions (Kalia *et al.* 2011). Tandem arrays of microsatellite motifs can mutate at a rate of 10^{-7} to 10^{-3} mutations per locus per generation (Zalapa *et al.* 2012). Natural slippage mutations (slipped strand miss-pairing) during DNA synthesis will lead to a gain or loss in the number of repeats in different species or populations, resulting in a unique genetic identity (Utomo *et al.* 2012). Microsatellite markers are considered highly polymorphic, informative, multi-allelic, locus specific, co-dominant and highly reproducible (Mallory *et al.* 2008) and can occur as multiple alleles per locus. It is this inherent hyper-variability that makes SSR markers more informative per site than the generally bi-allelic single nucleotide polymorphism (SNP) markers. As such, they are valuable in genetic diversity studies, conservation genetics, trait and genome mapping through marker assisted selection, linkage map development, cultivar fingerprinting, quantitative trait loci (QTL) mapping and forensic analysis (Santana *et al.* 2009; Zalapa *et al.* 2012; Omboki *et al.* 2015).

A significant drawback to the use of SSR markers is that the low-throughput process of marker discovery and testing, which inhibits the production of timely large, genome-wide marker sets. Additionally, PCR-amplified microsatellite products are traditionally separated by gel-slab electrophoresis or limited multicolor fluorescence capillary gel electrophoresis in an automated sequencer, but due to the limited number of dye sets available, markers that share overlapping size ranges will have to either be multiplexed separately or

labeled with a different dye, leading to a moderate throughput of samples and a high financial input (Kalia *et al.* 2011; Scheible *et al.* 2011).

The advent of high-throughput, next generation sequencing (NGS) offers a powerful alternative for comprehensive genomic screening and testing of molecular markers (Ferrandiz-Rovira *et al.* 2015). Using NGS together with emerging bioinformatics techniques, it is now possible to investigate genetic diversity across large populations at a fraction of the time and cost of conventional methods (Onda and Mochida 2016). Limited sequence data for a species is no longer a limitation to microsatellite discovery, since ample flanking sequence is available for PCR primer design even if the genome is sequenced at relatively low coverage and without expensive and complicated SSR enrichment protocols. The large datasets produced by NGS can encompass millions of reads, which increases the likelihood of SSR discovery exponentially. Markers can be detected using the raw read data, but sequence assembly into contigs can lend higher confidence to the SSR marker identified, since the original reads are joined together and are supported by the total number of reads in a specific contig. Furthermore, a relatively simple genome mapping algorithm can eliminate the possibility of multiple primer site binding, largely eliminating the possibility of cross-locus amplification (Zalapa *et al.* 2012). Prioritizing SSR loci for downstream analysis is an important step and largely depends on specific search parameters tailored for a research project. In general, higher polymorphism rates are associated with shorter motifs with higher repeat numbers, since they have a higher mutation rate compared to longer motifs with a low repeat rate. Perfect SSR markers (a pure set of repeats without interruption) are also to be prioritized, since they have been found to have greater allelic variability than imperfect or compound repeat sequences (Zalapa *et al.* 2012).

The *de novo* discovery of SSR markers from NGS datasets has been performed using Roche 454/GS-FLX sequencing in fungus, wasp, nematodes, plants and fish (Santana *et al.* 2009; Carlsson *et al.* 2013; Vartia *et al.* 2016), Illumina sequencing in plants, aquaculture species, fungi and insects (Nock *et al.* 2011; López-Urbe *et al.* 2013; Zhou *et al.* 2015; Patnaik *et al.* 2016; Thai *et al.* 2016) and Pacific Biosciences (PacBio) technology in birds and plants (Grohme *et al.* 2013; Wöhrmann *et al.* 2016). However, once these markers have been discovered and appropriate primers flanking the SSR of interest have been designed, screening still proceeds with mostly gel-based laboratory methods and complicated data analysis to determine polymorphic content (Muller *et al.*, 2016). Marker genotyping by NGS can be done on a substantial multiplexed scale, without the

need for complicated fluorescently labeled primer design or experimental layout as with capillary based sequencers (Vartia *et al.* 2016). In addition, a significant advantage of the NGS approach to SSR marker screening is the base pair sequence output, instead of estimated allele lengths. Fordyce *et al.* (2011) has compared the analysis of human SSR markers using both conventional capillary electrophoresis sequencing and the Roche Genome Sequencer (GS) FLX. The results indicated that the two technologies produced highly comparable datasets, but with the added advantage in NGS of important base variation discovery in marker flanking regions (homoplasmy) which are not observed through ABI capillary electrophoresis sequencing (Fordyce *et al.*, 2011). Vartia *et al.* (2015) also report the advantage of homoplasmy detection in a study where marker screening between capillary electrophoresis and NGS was compared in fish species (*Gadus morhua*). The authors emphasize that the scored allele sizes with traditional capillary electrophoresis sequencing does not reveal the presence of homoplasmy, but that NGS reveals the underlying sequence information to detect the single nucleotide differences. As a result, extra information on the analogous features arising through convergent evolution was revealed (Vartia *et al.*, 2015).

An essential component of accurate microsatellite scoring is effective bioinformatics tools. PCR amplification and sequencing can introduce erroneous insertions and deletions, leading to “alleles” with varying sequence lengths. Inferring genotypes for a population involves finding the optimal parameters to call a true microsatellite, including perfect and non-perfect repeats, at a reasonably low computational cost. Two methods have been developed and tested on NGS data; MiCNeSs on 454 pyrosequencing data of flies (*Drosophila melanogaster* and *Cervus elaphus*) and MEGASAT, tested on Illumina MiSeq data for fish (*Poeciliareticulata*) (Suez *et al.* 2016; Zhan 2016). Both methods showed sufficient inferences of microsatellite genotypes from NGS data.

The Amaranthaceae family (Order Caryophyllales) comprises numerous economically important plant species such as spinach, sugar beet and quinoa. The genus *Amaranthus* also resides within this family and is home to several important leafy and pseudo-grain vegetables, collectively known as the amaranths. These species are highly nutritious, with significant levels of β -carotene and iron in the leaves; and essential oils, minerals and rare amino acids (such as lysine) in the grains (Mnkeni *et al.* 2007; Raju *et al.* 2007). Amaranth crops have the potential to combat the effects of hidden hunger in rural and urban human populations, as well as play a large role in sustainable food production in future, due to their superior nutritional profile and easy cultivation.

Historically, morphological characterizations have been the most popular method to describe different amaranth genotypes and infer their relatedness (Mandal and Das 2002). Though amaranth is mostly a self-pollinating crop, some out-crossing between different species has been attempted in laboratory conditions with varying degrees of success (Pal and Khoshoo 1973; Trucco and Tranel 2011; Rastogi and Shukla 2013; Stetter *et al.* 2017). The successful hybridization experiments have thus far also been limited to between and within the grain amaranth species (*A. caudatus*, *A. cruentus* and *A. hypochondriacus*) (Gupta and Gudu 1991; Stetter and Schmid 2017). Sporadic, unconventional hybridizations that form without human interference have led to intermediate morphological features and in turn made conventional identifications of specifically the grain species groups through morphology or taxonomy extremely difficult. Additional complications arise where environmental fluctuations can lead to an increased diversity in morphological types. Therefore, a molecular approach could be a solution in the delineation of species identities, as the genetic makeup is not normally influenced by external, environmental factors.

Several molecular studies have been performed for amaranth, where genetic diversity was determined through different types of marker analysis technologies, including amplified fragment length polymorphism (AFLP), random amplified polymorphic DNA (RAPD) and microsatellite (SSR) markers (Costea *et al.* 2006; Mallory *et al.* 2008; Ray and Roy 2009). Most of the diversity studies to date have been performed on exclusively the three-grain species (*A. cruentus*, *A. caudatus* and *A. hypochondriacus*). The general conclusion across all studies was that *A. caudatus* is more closely related to *A. hypochondriacus* than to *A. cruentus* (Xu and Sun 2001; Mandal and Das 2002; Costea *et al.* 2006; Mallory *et al.* 2008, Stetter and Schmid, 2017). Only a few studies have included other members of the species complex, including *A. powellii*, *A. quitensis*, *A. spinosus*, *A. tricolor* and *A. dubius* (Chan and Sun, 1997, Sun, 1999; Xu and Sun, 2001; Mallory *et al.*, 2008), but have mostly offered conflicting and confusing results.

In this study, we demonstrate that by using high throughput, next generation sequencing, the (i) *de novo* identification of markers from a draft genome (ii) multiplexed amplification (iii) direct sequencing of fragments on the Illumina MiSeq platform and (iv) strict data filtering and analysis will reveal an adequate number of informative markers for genetic diversity studies in a much timelier and cost-effective manner compared to conventional capillary based methods. Furthermore, we demonstrate that the application of marker data generated through NGS reveals inter-specific genetic diversity in the plant genus *Amaranthus*.

Materials and Methods

Amaranthus accessions and growth

Thirteen different *Amaranthus* species, represented by 45 accessions, were provided by Dr David Brenner, North Central Regional Plant Introduction Station (NCRPIS, USA) (Table 4.1). Accessions were chosen from the Germplasm Resources Information Network (GRIN) computer database of the USDA-ARS National plant germplasm system (hereafter referred to as GRIN accessions). In addition, 14 unknown *Amaranthus* accessions were obtained from the germplasm bank at the Agricultural Research Council Vegetable and Ornamental Plant Institute (ARC-VOPI), Pretoria, South Africa (hereafter referred to as SAG accessions) (Table 4.1). Seeds from 59 accessions were germinated in irrigated glasshouses at ARC-VOPI with day temperatures of 25°C-35°C, night temperatures of 20°C-25°C and natural light intensity. Germination occurred 3-15 days post planting and leaves were harvested at approximately 2 weeks old. Samples were stored at -80°C until use.

Table 4.1: *Amaranthus* genus germplasm sets

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

*Identified based on preliminary morphological analysis (Data not shown)

Identify microsatellite loci from draft genome assembly

Total genomic DNA was extracted from leaf material (approximately 100 mg) of *Amaranthus tricolor*, using a DNeasy Plant Mini DNA Isolation kit (Qiagen, Valencia CA, USA) following the protocol outlined by the manufacturer. The genomic DNA was quantified using the Qubit® 2.0 Fluorometer Broad Range dsDNA quantification assay (Invitrogen, Life Technologies, CA, USA) and visualized on a 1% agarose gel stained with 0.5 µg/ml ethidium bromide. One microgram of DNA was used to prepare sequencing libraries with TruSeq®DNA Sample Prep Kit (Illumina, San Diego, USA) according to manufacturer's protocol. Paired end (100-300 bp) and long mate pair (3kb and 5kb) library data was sequenced on the HiScanSQ Illumina sequencer (Illumina, San Diego, USA). The sequencing was performed at the Biotechnology Platform Sequencing Facility (Agricultural Research Council (ARC-BTP), South Africa). Quality checks and filtering of all reads was done using FastQC (<http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc>).

CLC Bio Genomics Workbench v8 (CLCBio Genomics Workbench, Aarhus, Denmark, <https://qiagenbioinformatics.com>) was used to trim the Illumina adapter sequences, as well as removing low quality reads from the dataset. Sequence contig assembly and scaffolding was performed using the software ALLPATHS-LG with default parameters (Butler *et al.*, 2008). Forty randomly chosen scaffolds (length>10 kb) were investigated for microsatellite repeat regions using MSATCOMMANDER (Rozen and Skaletsky 1999; Faircloth 2008), identifying 671 loci containing microsatellite repeat regions. Only microsatellites (and flanking sequence) that mapped to a single region during a local BLAST analysis (CLC Bio) against the scaffolds were reserved for analysis to exclude any cross-hybridization during amplification. In total, 49 microsatellite loci were retained. The repeat region as well as approximately 1000 bp up- and downstream of the repeat was subjected to nucleotide BLAST analysis and potential functional annotation using Blast2GO analysis with default parameters (Conesa *et al.* 2005; Conesa and Götz 2008).

Marker validation

Forty-nine primer pair sequences were obtained using the software BatchPrimer3 (<https://probes.pw.usda.gov/batchprimer3/>) (You *et al.* 2008) and were designed to flank approximately 150 bp up- and downstream of the target repeat region. Microsatellite loci for which both a forward and reverse primer could not be designed were discarded. Primers were specifically designed to have (i) a melting

temperature of 57°C-60°C; (ii) primer sequence lengths at 18-25 bp or more to ensure adequate specificity; (iii) a GC content of 40%-60%; and (iv) no self-complementarities or secondary structures within the primers. The fragments were filtered to only represent di- (≥ 6), tri- (≥ 5), tetra- (≥ 3), penta- (≥ 3) and hexanucleotide (≥ 3) repeat classes. The primer oligos were synthesized by Integrated DNA Technologies (IDT, Whitehead Scientific, Cape Town, South Africa). All 49 SSR microsatellite loci were individually amplified from *A. tricolor* genomic DNA in a final concentration of 5 mM and DreamTaq PCR Mastermix (2x, DreamTaq DNA Polymerase, 2x DreamTaq Buffer, dNTPs and 4 mM MgCl₂) (Thermo Fisher Scientific, Waltham, MA, USA) in a final volume of 25 μ l. The PCR was performed on a G-Storm GS1 thermal cycler (Vacutec, South Africa) with the following conditions: 95°C (3 min); 35 cycles of 94°C (30 s), 58°C (30 s), 72°C (30 s) and a final extension of 72°C (10 min). PCR amplification and quality were assessed by visualization on a 3% (w/v) agarose gel stained with 0.5 μ g/ml ethidium bromide. Loci that showed low or no amplification were discarded from further analysis.

Twenty-three microsatellite loci that yielded a single amplicon for *A. tricolor* were further assessed on three additional amaranth species (*A. caudatus* – GRIN7, *A. cruentus* – GRIN10 and *A. graecizans* – GRIN15) to assess the inter-species amplification ability and polymorphic content of the primer sets. PCR reactions, conditions and template visualizations were performed as stated above. In total, six markers were uniformly amplifiable and polymorphic across the three additional amaranth species. Each of the six markers amplified (across four accessions) was individually purified using the MinElute PCR Cleanup kit (Qiagen) using manufacturers' instructions and eluted in a final volume of 12 μ l Elution Buffer (EB). The alleles were Sanger sequenced at Inqaba Biotechnical Industries Pty. Ltd. (Pretoria, South Africa), to confirm the presence of repeat elements.

Marker screening via ABI capillary gel electrophoresis technology

Forty-five amaranth accessions (Table 4.1 – GRIN germplasm set) were screened with six SSR markers using ABI capillary gel electrophoresis technology. Genomic DNA was prepared from leaf tissues of the accessions using the DNeasy Plant Mini DNA Isolation kit (Qiagen) and quantification done on the Qubit® 2.0 Fluorometer (Invitrogen) using the Broad Range dsDNA quantification solution. The six SSR primer pairs were obtained from IDT (Whitehead Scientific), where the forward primer was fluorescently labelled at the 5'

end, while the reverse primer remained unlabelled. Markers were divided between multiplexed reactions to avoid overlapping amplicon sizes and the combination of common fluorescently labelled primer sets. Two multiplexed reactions contained three labelled primer pairs each and were amplified using the Multiplex PCR kit (Qiagen), according to manufacturers' instructions. Each reaction contained 2x Qiagen Multiplex PCR Master Mix (HotStarTaq® DNA Polymerase, a mix of dNTPs and Multiplex PCR Buffer containing 6 mM MgCl₂), equimolar concentrations (1 uM) of each primer, 20 ng/μl template DNA and RNase Free water up to a final volume of 25 μl. The PCR was performed on a G-Storm thermal cycler (BioRad, CA, USA) with the following conditions: Initial activation step of 95°C (15 min); 30 cycles of 94°C (30 s), 58°C (90 s), 72°C (60 s) and a final extension of 72°C (45 min). Each sample was purified using the MinElute PCR Cleanup kit (Qiagen) and eluted in a final volume of 12 μl EB buffer. Samples were diluted and re-suspended in Hi-Di™ Formamide (Thermo Fisher Scientific) together with the GeneScan™ 600 LIZ® Size Standard (Thermo Fisher Scientific). The reactions were separated on an ABI 3730xl capillary sequencer (Applied Biosystems) at the DNA Sequencing Facility (University of Pretoria, Pretoria, South Africa). Genotyping of the samples was performed using the GeneMapper® v4 software (Applied Biosystems) to resolve allele sizes.

Marker screening via MiSeq NGS technology

The six primer pairs used in the ABI capillary electrophoresis analysis were re-designed to have the Illumina adapter primers (Forward: 5'TCGTCGGCAGCGTCAGATGTGTATAAGAGAC-AGCCTACGGGNGGCWGCAG3' and Reverse: 5'GTCTCGTGGGCTCGGAGATGTGTA-TAAGAGACAGGACTACHVGGGTATCTAATCC3') as overhangs on the 5' ends, according to the 16S Metagenomic Sequencing Library Preparation (MSLP) protocol (Illumina). Neither primer was fluorescently labeled. A multiplexed PCR reaction containing the six adapter-overhang primer pairs was prepared for each of the 59 amaranth accessions (GRIN and SAG germplasm sets) as described in section 2.3 of this chapter. The MSLP protocol was followed for the index PCR step (where unique Illumina dual indices/barcodes are ligated for each sample/accession) and subsequent PCR clean-up reactions using Agencourt AMPure XP beads (Beckman Coulter Inc, CA, USA). Library quantification was performed with the Qubit® (Invitrogen) system and all 59 samples were multiplexed in equal molar quantities for a single lane run on a MiSeq Illumina sequencer (Illumina) at the ARC-BTP (Supplementary Figure S4.1).

The raw reads (250 bp – 300 bp) were de-multiplexed into individual data bins for each of the 59 amaranth accessions. The Illumina adapter sequences were removed, as well as reads with low quality scores (default 0.05, corresponding to a Phred score of approximately 15) using CLC Bio. Overlapping read pairs were merged, while unmerged/singleton reads (variants represented by a single allele) were discarded from further analysis. Data analysis was performed on all accessions using an in-house developed python script – Microsatellite Analysis using NGS Data (MAND, Appendix A) (Figure 4.1). Software prerequisites for running the script included BioPython (www.biopython.org) and the source file of Tandem Repeats Finder (TRF, <https://tandem.bu.edu/trf/trf.download.html>). The sequence read files, together with a text file of the different primer sequences, formed the input of the MAND script. For each accession, the reads were assigned to a specific locus based on the perfect matching of locus specific primer pair sequences on each fragment end. Sequences generated from non-specific binding (e.g. due to a miss-match in the primer sequence) or fragments that do not contain matching primer sequences on both ends were discarded. The software creates individual folders for each amaranth accession, corresponding to each microsatellite locus. Within each folder, the fragments were grouped according to their length, henceforth referred to as Fragment Length Groups (FLGs). One representative sequence from each FLG was subjected to SSR discovery by tandem repeats finder and output a text file with SSR length, repeat size, repeat copy number and repeat pattern. Furthermore, each representative fragment of a FLG was visually inspected to verify the microsatellite repeat, as well as any insertions and deletions that might have been present in the flanking regions. Lastly, a filtering step was applied to remove FLGs considered to be a result of stuttering or inaccurate sequencing. If the number of fragments in a FLG was less than 40% of the number of fragments within the largest FLG, it was discarded from further analysis. The FLGs that remained after filtering were assumed to represent true allele sizes.

Prerequisites	•BioPython •Tandem Repeats Finder source file
Input	•Trimmed/Merged NGS data (.fa) •Primer sequence file (.txt)
Fragment Groups	•For each accession, the fragments are grouped based on matching terminal primer sequences
Fragment Counts	•Within each fragment group, the number of fragments/alleles with equal length are further grouped into Fragment Length Groups (FLGs)
FLG Filtering	•Removes low count FLGs, regarded as errors or stuttered sequencing
SSR Discovery	•One sequence from each FLG is subjected to SSR discovery using Tandem Repeats Finder

Figure 4.1: Steps followed by the Microsatellite Analysis from NGS Data (MAND) script. Using BioPython and the Tandem Repeats Finder (TRF) source file, it was possible to accurately detect and describe targeted microsatellite repeats from trimmed/merged NGS data. Fragments of a specific marker were grouped according to terminal primer sequences. Alleles were further grouped according to fragment lengths, referred to as Fragment Length Groups (FLGs), and a filtering step removed all FLGs that contain less than 40% of the total fragments of the largest FLG, therefore discarding allele amplicons containing stuttering or inaccurate sequencing. One sequence from each FLG was subjected to SSR discovery using TRF.

Genetic diversity of ABI capillary electrophoresis compared to MiSeq NGS technologies

Allelic data generated by the ABI capillary electrophoresis and NGS technologies were analyzed using GeneMapper software and the custom MAND script respectively. The resulting data was directly compared for the 45 GRIN accessions to determine the level of similarity in allele calling between the two techniques. Allele frequencies observed and expected levels of heterozygosity and fixation indices for both datasets were determined using GenAIEx (Peakall and Smouse, 2006). Sanger sequences obtained from ABI fragments (Section 2.3) were aligned with default parameters (CLC Bio) to the sequences obtained from MiSeq analysis.

Species relationships of unknown SAG genotypes revealed by MiSeq analysis of SSR loci

During the current study, the GRIN accessions were firstly assigned to a species group based on morphological information within the GRIN database. Subsequently, SAG accessions were assigned to a species group based on preliminary morphological comparisons to the GRIN germplasm accessions (Table 4.2). Downstream SSR analysis will be compared to these baseline identifications for confirmation.

Table 4.2: Species groupings of GRIN and SAG accessions

Species group	Accessions	Species
1	GRIN 1, GRIN 2	<i>A. blitum</i>
2	GRIN 3, GRIN 4, GRIN 5, GRIN 6, GRIN 7, SAG 3, SAG 30	<i>A. caudatus</i>
3	GRIN 8, GRIN 9, GRIN 10, SAG 4, SAG 7, SAG 9	<i>A. cruentus</i>
4	GRIN 11, GRIN12, GRIN13, GRIN14, SAG 14, SAG 34	<i>A. dubius</i>
5	GRIN 15, GRIN 16, SAG 10, SAG 12, SAG 17	<i>A. graecizans</i>
6	GRIN 17, GRIN 18, GRIN 19, GRIN 20, GRIN 21	<i>A. hybridus</i>
7	GRIN 22, GRIN 23, GRIN 24, GRIN 25, GRIN 26, GRIN 27	<i>A. hypochondriacus</i>
8	GRIN 28, GRIN 29, GRIN 30, SAG 1	<i>A. powellii</i>
9	GRIN 31, GRIN 32	<i>A. quitensis</i>
10	GRIN 33, GRIN 34, GRIN 35	<i>A. retroflexus</i>
11	GRIN 36, GRIN 37, GRIN 38	<i>A. spinosus</i>
12	GRIN 39, GRIN 40, GRIN 41, SAG 11, SAG 29, SAG 36	<i>A. tricolor</i>
13	GRIN 42, GRIN 43, GRIN 44, GRIN 45	<i>A. viridis</i>

Principle Component Analysis (PCoA) was performed for the entire dataset using the GenAlEx software with default parameters. An unrooted dendrogram was constructed using DendroUPGMA (online software <http://genomes.urv.cat/UPGMA/>) applying Pearson's coefficient with 1000 bootstrap replicates (Garcia-Vallvé and Puigbo 2009). The dendrogram was viewed with the software FigTree v1.4.2 (<http://tree.bio.ed.ac.uk/software/figtree/>).

Results

Identification of microsatellite loci from a draft genome assembly of *Amaranthus tricolor*

A draft genome was assembled from sequenced leaf tissue of *A. tricolor*. Approximately 1.3 Gb reads were obtained from both paired end and long mate pair sequencing libraries (mean read length of 98 bp) using the Illumina HiScan technology. The average GC content was estimated at 37.2%, slightly higher than the reported 35% for BAC-end sequencing in *A. caudatus* (Maughan *et al.*, 2009) and the previously sequenced *A. hypochondriacus* genome (34% and 34.7%) (Sunil *et al.*, 2014; Clouse *et al.*, 2015). After quality filtering and adapter trimming using FastQC; 569.5 Mb data was captured in the assembly using ALLPATHS-LG software. Contig assembly yielded 68 656 contigs with an N50 size of 6 kb. Further assembly of the contigs into scaffolds resulted in 13 606 scaffolds, representing 402 Mb (45.6%) of the estimated *Amaranthus tricolor* genome size of 880 Mb (Bennet and Smith, 2000) (Supplementary Table S4.1).

Forty randomly chosen scaffolds larger than 10 kb were analyzed using the MSATCOMMANDER 1.0.8 software to identify di- (≥ 6), tri- (≥ 5), tetra- (≥ 3), penta- (≥ 3) and hexanucleotide (≥ 3) microsatellites. Genomic scaffold repeat motifs are presented in Supplementary Table S4.2. A total of 671 microsatellite loci were identified, indicating an occurrence of one SSR per 6.8 kb of sequence. Trinucleotide motifs were found to be most common (37%), followed by tetranucleotide (22%), dinucleotide (25%), pentanucleotide (9%) and hexanucleotide (6%) motifs. Using a local BLAST analysis (CLCBio), 49 microsatellite containing loci were identified that do not occur at multiple locations within the entire scaffold assembly.

Each of the 49 SSR microsatellite repeat regions, together with approximately 1000 bp flanking regions, were subjected to nucleotide BLAST analysis and the top species hit distribution from the GenBank database are presented in Figure 4.2. Ten SSR regions returned no BLAST hits, while the remaining 39 microsatellite loci returned hits with e-values above $1e-5$. The top hits included 20% *Amaranthus palmeri* (Amaranthaceae) followed by the closely related *Beta vulgaris* (Amaranthaceae) and *Vitis vinifera* (Vitaceae) with 8% and 5% respectively. Approximately 1.7% of the total hits included *A. hypochondriacus*. Blast2GO analysis revealed that none of the microsatellite sequences could be assigned to genic regions with gene ontology terms.

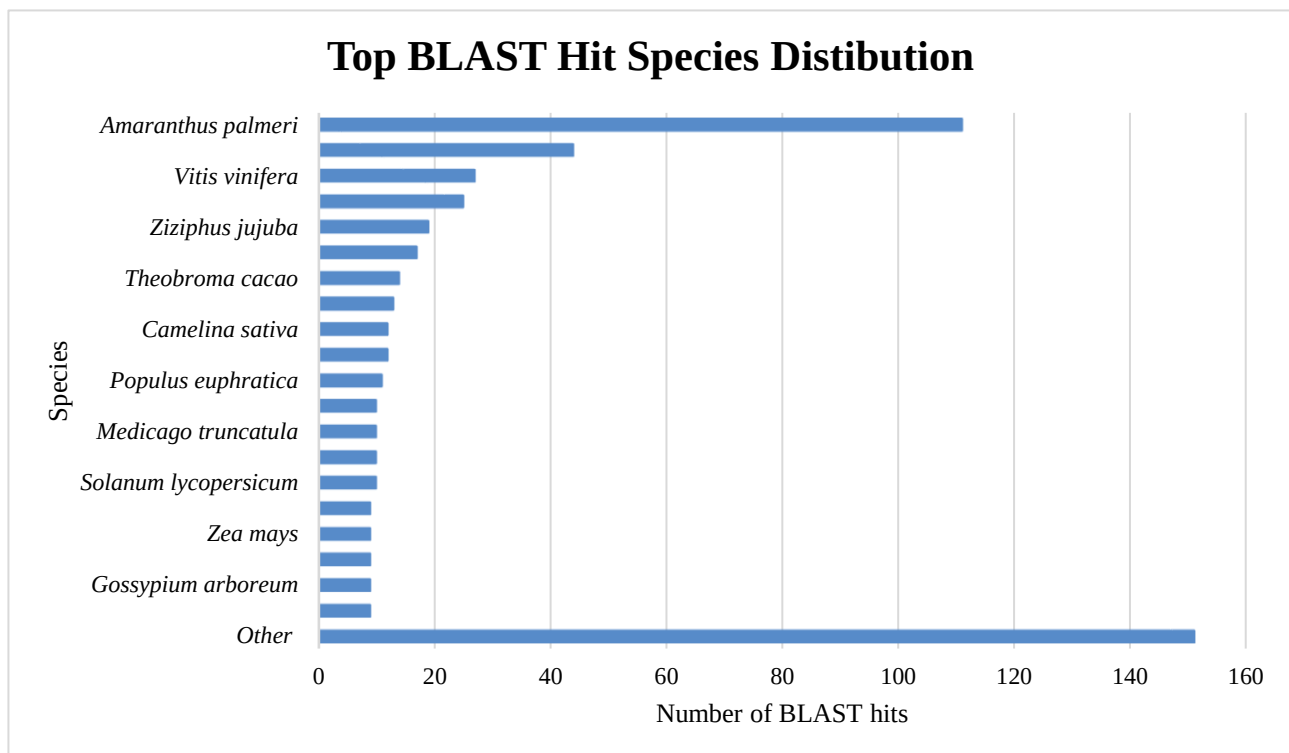


Figure 4.2: Top nucleotide BLAST hit results for individual microsatellite sequences. Each microsatellite locus, together with approximately 1000 bp sequence data up- and downstream of the site, was subjected to nucleotide BLAST analysis and the top hits were distributed between *A. palmeri* (20%), *B. vulgaris* (8%) and *V. vinifera* (5%). Blast2GO analysis revealed that none of the microsatellite regions mapped to genic regions.

Microsatellite validation

BatchPrimer3, which has a primer design function in the software, was used to design both forward and reverse primers for each of the 49 identified microsatellites, approximately 150 bp up- and downstream of the repeat region. Microsatellites, for which primers could not be designed to fit specified prerequisites (melting temperature of 57°C-60°C, primer length of 18-25 bp, GC content of 40%-60% and no self-complementarities or secondary structure formation) were discarded from further analysis. The amplification efficiency of the remaining 23 microsatellite loci (Table 4.3) was tested on *A. tricolor* genomic DNA, using one uniform PCR protocol. Reactions that did not result in an amplified product or produced multiple amplicons/smear after amplification were discarded from further analysis.

Table 4.3: Primer and repeat properties for 23 microsatellite loci identified from *A. tricolor* scaffolds

SSR	Genomic Scaffold	Scaffold Size (bp)	Forward Primer (5'-3')	Reverse Primer (5'-3')	Repeat
ATR1	510	90 709	GAAGCTTTTTAGCTTTGAAGAGACC	GTCACCTCTGCTGAACCTTTCTTG	(CAT)9
ATR 5	45	169 873	TTATACCATTCTTGCCCTTCAGC	TCATTCAGCAGTACAGTTGAACC	ATGGGG
ATR 8*	415	98 681	GAAACCAACAAAGTAGTGGGAGTT	AGAACCCTCTTGTCCTCTTTATC	(GATAAA)7
ATR 11	19 048	242 352	CGGCCCTCTATAGTCCTAATCAAT	GTGAGAGCCTGGATTCAACTTTTA	(TAT)10
ATR 12*	19 048	242 352	GGACTAACTGAATAAAGCCAAGTCA	TGTATGAGTACGTACATGTGATAGTGC	(ATT)12
ATR 17	2 074	46 363	GGAGAACAAGGTGAAAATAAGGAG	GACGTTCACTACAGTTTGTGATCC	(AG)14
ATR 18	2 074	46 363	GCAAATACCCACAAACATTTCC	GGACTATTGTTGGCTGCTTT	(AAG)23
ATR 19*	2 074	46 363	ATACGCAGAATCACATCTCTCTTG	GAAGTCGATAGCGTGTGTTTGAC	(TAT)34
ATR 20	6 443	57 098	GGGGAGTTTCGTATCTCAAGAGTAA	CGGTCTTTGTGAGAAGAATACGAT	(ATT)9
ATR 24	3 780	205 347	AAGGGGGTCTACTATCTTTACTAGG	CTCATGCTTTTTGAGAGGTACA	(TTA)15
ATR 25	975	70 485	CTATAATCATGAACCGCGTGAAC	CTTCCAAGCTTCAGGTTCTTT	(TAT)11
ATR 28*	1 427	57 103	TGAGGTCAATTGCCACAACCTAC	GATTGAGGAAAGAGAAAGCGAAAG	(AGA)10
ATR 31	372	99 585	TCGCTCTGATTGGGTATATAGCTT	GACAGTTTGGGTGATGGCTATAAT	(TA)11
ATR 32*	372	99 585	GAACGGATCTCTGCTTGCTAAATA	GTAACACACATCTGGGAGTTTGAG	(TTG)9
ATR 38	159	128 981	CATCAAAGTACTTCGCATCGTATC	GAATGCCACTTATTACCAAACCTC	(CAA)7
ATR 52	18 989	50 321	GAGTGTTCCAAGAAGGGTATGTTT	AGAGAGATCCAAGAGACAGAAGGA	(AGCAG)7
ATR 55	91	214 804	GTGTATTTTCATGATGAGACCAT	TGAATTTGGAGGTTACTTGAG	(TAA)14
ATR 56	1 214	61 432	ATCTCCCTCTTCTTGAAACAG	TGTATGTTGTTGTTTCGCATT	(GA)8
ATR 58	3 584	27 222	TAATGGTGGTGGTAGAGATCA	CAACCCCTACCATAACTATC	(CTT)7
ATR 60	1 110	65 443	TCACGAGAAAACATGCAAT	GCACATTAAGGGCTCAATTA	(TAT)21
ATR 61	288	111 565	CTTGAGAAGATAAAGCCAACA	CTAGATCCGTCTCCTCAAGT	(CTT)11
ATR 62*	8	225 737	TATGTAATGCCTGCACCTACT	CACACAAAGGAGCTACTCAAC	(AGA)18
ATR 63	8	225737	AACTGGCAAATCATGTTACAC	GTAGCTGTGCTGAACAAAAC	(CT)11

* Markers in bold were used in downstream analysis.

The amplification efficiency and polymorphic capacity of the 23 microsatellite loci were further assessed on three additional *A. tricolor* related *Amaranthus* species, namely *A. cruentus* (GRIN10), *A. caudatus* (GRIN7) and *A. graecizans* (GRIN15). Markers, which did not yield amplicons in all four accessions, were discarded from further analysis. In total, six alleles were identified which amplify with a uniform PCR protocol and show polymorphism between four different *Amaranthus* species. Sanger sequencing of the PCR products further validated the presence of the SSR repeats. For ATR32, *A. tricolor* revealed the expected repeat of (TTG)⁹, while in comparison the *A. graecizans* had an imperfect (TTG)¹³ repeat and both *A. caudatus* and *A. cruentus* had a (TTG)⁶ repeat sequences (Figure 4.3). The remaining microsatellite loci had similar polymorphic profiles across the different accessions sequenced (Supplementary Figure S4.2).

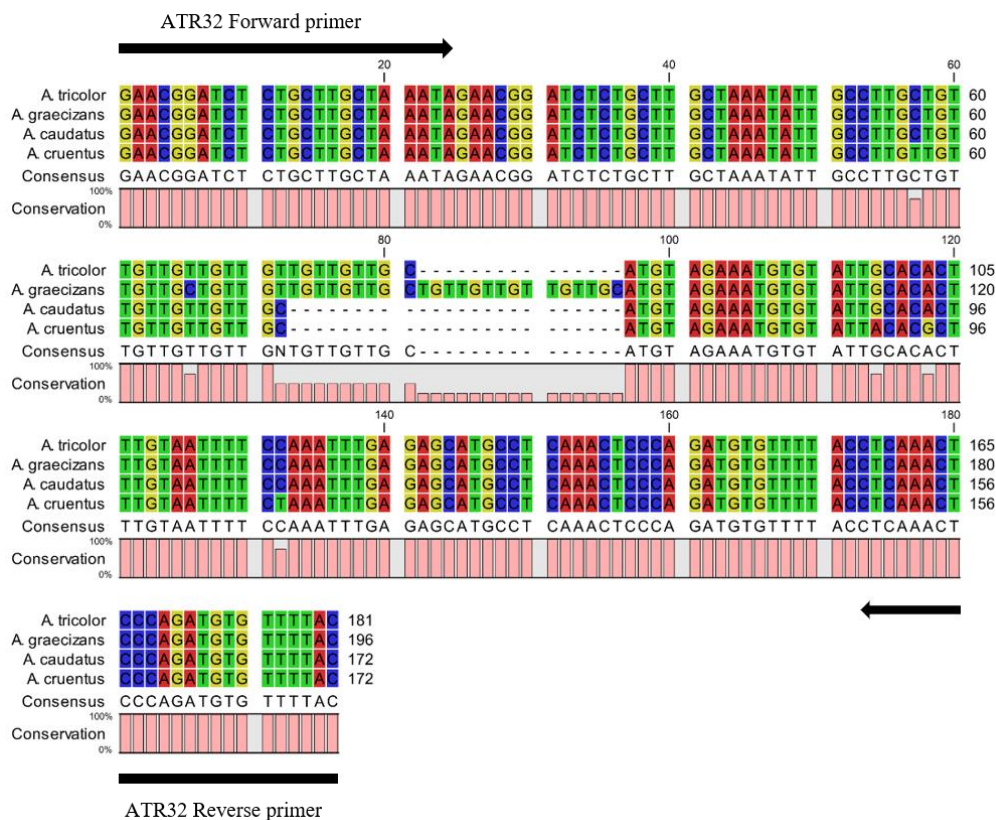


Figure 4.3: Sanger sequencing of ATR32 in *A. graecizans* (GRIN15), *A. tricolor* (SAG29), *A. caudatus* (GRIN3) and *A. cruentus* (GRIN8) reveals the inter-species polymorphic nature of the marker. The expected repeat of (TTG)⁹ was validated for *A. tricolor*, while in comparison the *A. graecizans* had an imperfect (TTG)¹³ repeat and both *A. caudatus* and *A. cruentus* had imperfect (TTG)⁶ repeat sequences. Forward and reverse primer sequences are indicated on the sequence.

Marker screening via ABI capillary gel electrophoresis technology

For each of the 45 GRIN *Amaranthus* accessions, two multiplexed PCR reactions were prepared. Each contained three of the six SSR markers (Table 4.4). PCR amplification was performed on each germplasm accession and the result revealed polymorphic banding patterns between the different species (Figure 4.4).

Table 4.4: Multiplexed PCR layout for each accession

Multiplex 1		Multiplex 2	
Marker	Fluorescent Label	Marker	Fluorescent Label
8	6-Fam	28	6-FAM
12	VIC	32	VIC
19	NED	62	NED

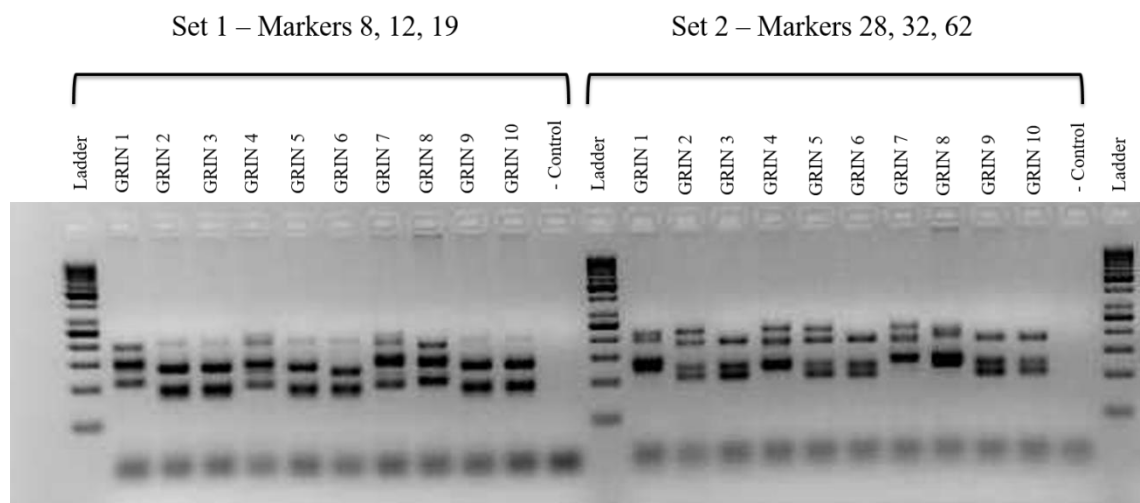


Figure 4.4: Multiplexed amplification of two marker sets on 10 diverse accessions (GRIN1-GRIN10) of *Amaranthus*. Each multiplexed reaction contained three sets of locus specific primers, which were simultaneously amplified for each accession. The patterned bands observed are indicative of the polymorphic nature of the markers. Ladder = 50bp O’Rangeruler (Thermo Fischer Scientific).

For each accession, GeneMapper v4 (Applied Biosystems) software was utilized to visually inspect and call allele sizes for each locus compared to the LIZ 600 size standard (Thermo Scientific). When peaks could not be accurately called due to stuttering or if the genotyping failed at a particular locus, the data was scored as missing. Two markers resulted in missing data (null alleles) for certain genotypes, namely ATR19 (11.2%) and ATR32 (6.7%). The genotypic information was captured in a data file for downstream analysis software (Table 4.5). All loci were homozygous for the accessions analyzed, except for GRIN 14 (*A. dubius*) at locus ATR62, which had a heterozygous allele call.

Table 4.5: Genotypes identified using ABI capillary gel electrophoresis technology and GeneMapper v4 software

	ATR8		ATR12		ATR19		ATR28		ATR32		ATR62	
	Allele 1	Allele 2	Allele 1	Allele 2	Allele 1	Allele 2	Allele 1	Allele 2	Allele 1	Allele 2	Allele 1	Allele 2
GRIN1	107	107	180	180	193	193	140	140	123	123	149	149
GRIN2	107	107	180	180	187	187	151	151	123	123	136	136
GRIN3	96	96	183	183	184	184	128	128	119	119	168	168
GRIN4	96	96	183	183	208	208	137	137	119	119	168	168
GRIN5	96	96	183	183	184	184	128	128	119	119	168	168
GRIN6	96	96	186	186	211	211	137	137	123	123	168	168
GRIN7	96	96	183	183	184	184	128	128	119	119	168	168
GRIN8	96	96	186	186	232	232	137	137	123	123	168	168
GRIN9	96	96	186	186	203	203	137	137	123	123	168	168
GRIN10	96	96	186	186	232	232	137	137	123	123	168	168
GRIN11	96	96	162	162	184	184	137	137	119	119	168	168
GRIN12	96	96	162	162	184	184	137	137	119	119	168	168
GRIN13	96	96	162	162	184	184	137	137	119	119	168	168
GRIN14	96	96	162	162	184	184	137	137	119	119	168	171
GRIN15	102	102	165	165	203	203	144	144	149	149	136	136
GRIN16	102	102	189	189	203	203	144	144	129	129	127	127
GRIN17	96	96	191	191	205	205	134	134	123	123	168	168
GRIN18	96	96	183	183	186	186	137	137	126	126	168	168
GRIN19	96	96	186	186	208	208	137	137	123	123	168	168
GRIN20	96	96	186	186	232	232	137	137	123	123	168	168
GRIN21	96	96	183	183	203	203	137	137	119	119	168	168
GRIN22	96	96	191	191	208	208	137	137	123	123	168	168
GRIN23	96	96	191	191	211	211	137	137	123	123	168	168
GRIN24	96	96	191	191	211	211	137	137	123	123	168	168
GRIN25	96	96	191	191	208	208	137	137	123	123	168	168
GRIN26	96	96	186	186	213	213	137	137	123	123	168	168
GRIN27	96	96	191	191	211	211	137	137	123	123	168	168
GRIN28	96	96	183	183	186	186	137	137	123	123	168	168
GRIN29	96	96	189	189	186	186	137	137	126	126	168	168
GRIN30	96	96	185	185	186	186	137	137	126	126	168	168
GRIN31	96	96	189	189	0	0	137	137	119	119	168	168
GRIN32	96	96	183	183	184	184	137	137	119	119	168	168
GRIN33	96	96	183	183	186	186	137	137	123	123	168	168
GRIN34	96	96	183	183	186	186	137	137	123	123	168	168
GRIN35	96	96	183	183	186	186	137	137	123	123	168	168
GRIN36	96	96	162	162	0	0	137	137	0	0	171	171
GRIN37	96	96	162	162	0	0	137	137	0	0	171	171
GRIN38	96	96	162	162	0	0	137	137	0	0	171	171
GRIN39	107	107	162	162	198	198	151	151	132	132	149	149
GRIN40	107	107	186	186	208	208	151	151	132	132	149	149
GRIN41	107	107	165	165	216	216	157	157	132	132	158	158
GRIN42	96	96	165	165	213	213	137	137	139	139	136	136
GRIN43	96	96	165	165	190	190	137	137	139	139	136	136
GRIN44	96	96	165	165	208	208	137	137	139	139	127	127
GRIN45	96	96	165	165	196	196	137	137	139	139	130	130

A value of “0” indicates a missing genotype/null allele.

Marker screening via NGS Illumina MiSeq technology

The multiplexed PCR reactions and amplification conditions were identical to those used for the ABI capillary electrophoresis analysis; except that all markers were multiplexed in a single reaction per individually barcoded accession and no fluorescently labeled primers were used. The multiplexed sample was run on the Illumina MiSeq platform and generated 11.7 Mb reads across the entire germplasm set (Figure 4.5). The number of reads generated per accession are highly comparable (an average of 261 kb per accession), except for accessions GRIN 1, GRIN 2, SAG 7, SAG 10, SAG 17, SAG 34 and SAG 36, which had an almost 10-fold increase in the number of reads produced. The discrepancy in the number of reads could be attributed to pipetting errors during library construction or MiSeq sample loading. The reads were quality filtered (Phred score > 20, removal of low-quality bases), trimmed and merged to a final dataset of 4.4 Mb reads across all individuals. On average, 32% of reads were removed by quality filtering (minimum 28% - GRIN30; maximum 38% - GRIN8). The trimmed and merged read length ranged from 20-570 bp, with an average read length of 141 bp. Approximately 12%-43% (mean 32%) of the total merged reads had both the forward and reverse primer sequences at the respective ends, as well as a microsatellite repeat within the fragment. These reads could therefore confidently be assigned to the respective marker loci for each individual.

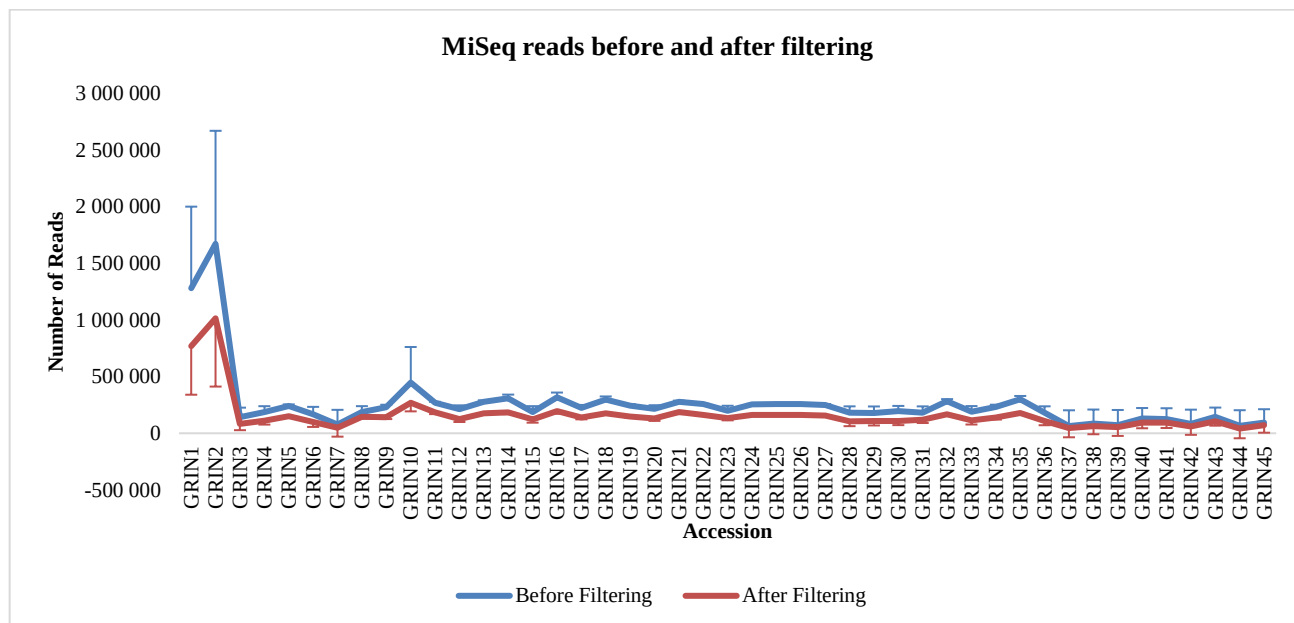


Figure 4.5: Number of raw reads (blue) and number of filtered reads (red) generated for each *Amaranthus* GRIN accession, on one lane of the Illumina MiSeq platform. Error bars indicate standard deviation from the mean number of reads before and after filtering, per MiSeq run. An approximately 10-fold increase in data was generated for GRIN1 and GRIN2 compared to the remaining accessions, possibly due to pipetting error during library construction or sample loading. An average of 32% of reads was removed by quality filtering (minimum 28% - GRIN30; maximum 38% - GRIN8) across the germplasm set.

In each accession, the MAND script separated marker loci first according to corresponding forward and reverse primers and secondly according to the different fragment length groups (FLG). If the total number of fragments within a FLG is less than 40% of the number of fragments within the largest FLG, the group was discarded as stutter or sequencing errors (Figure 4.6). In this way fragments that had size differences due to insertions or deletions in their flanking regions (not due to the repeat motif) were also eliminated. On average, the MAND script removed 85%-89% of FLG's per locus, resulting in approximately 11%-15% FLGs representing true alleles. The MAND script was applied to all markers for each accession and true alleles scored across the germplasm set. Two markers revealed missing genotypes across accessions (no FLGs could be observed); ATR19 (8.9%) and ATR32 (6.7%). All loci were homozygous across the species groups, except for GRIN14 (*A. dubius*) at locus ATR62, which had a heterozygous allele call. All genotypic information was captured in a data file for downstream analysis (Table 4.6).

The Sanger sequencing output generated from the ABI fragments were directly compared to the sequencing output from the MiSeq analysis (Supplementary Figure S4.3). For ATR32, the sequencing from both technologies for the four accessions analyzed (*A. tricolor* – SAG29, *A. graecizans* – GRIN15, *A. caudatus* (GRIN3) and *A. cruentus* (GRIN8) each reveal the predicted fragment size from the respective technologies (Table 4.5, Table 4.6). The two leafy (*A. tricolor* and *A. graecizans*) and one grain amaranth (*A. cruentus*), consistently shows a one base deletion in the ABI data compared to the MiSeq data and the grain *A. caudatus*, the ABI data shows a deletion of three base pairs. For ATR28, the ABI indicates a two and one base insertion for *A. tricolor/A. graecizans* and *A. caudatus/A. cruentus* respectively. ATR8 shows consistent deletions for the ABI data in all accessions, three for *A. tricolor* and two each for *A. graecizans/A. cruentus/A. caudatus*. The discrepancies between the data could be due to inability of the sequencing Taq polymerase enzyme to accurately sequence (i) the first 20-30 bases after primer binding and (ii) long homopolymer regions using the ABI capillary electrophoresis technology. Since all the insertions/deletions occur within the first 20-30 bases after primer binding, the inaccurate calling by the ABI could be an artifact of the enzyme used during sequencing. Therefore, the sequencing by synthesis technology employed by the MiSeq platform, coupled with the deep coverage of each fragment, adds the needed confidence to effectively assign the correct base to the correct sequence position (Niedringhaus *et al.*, 2011).

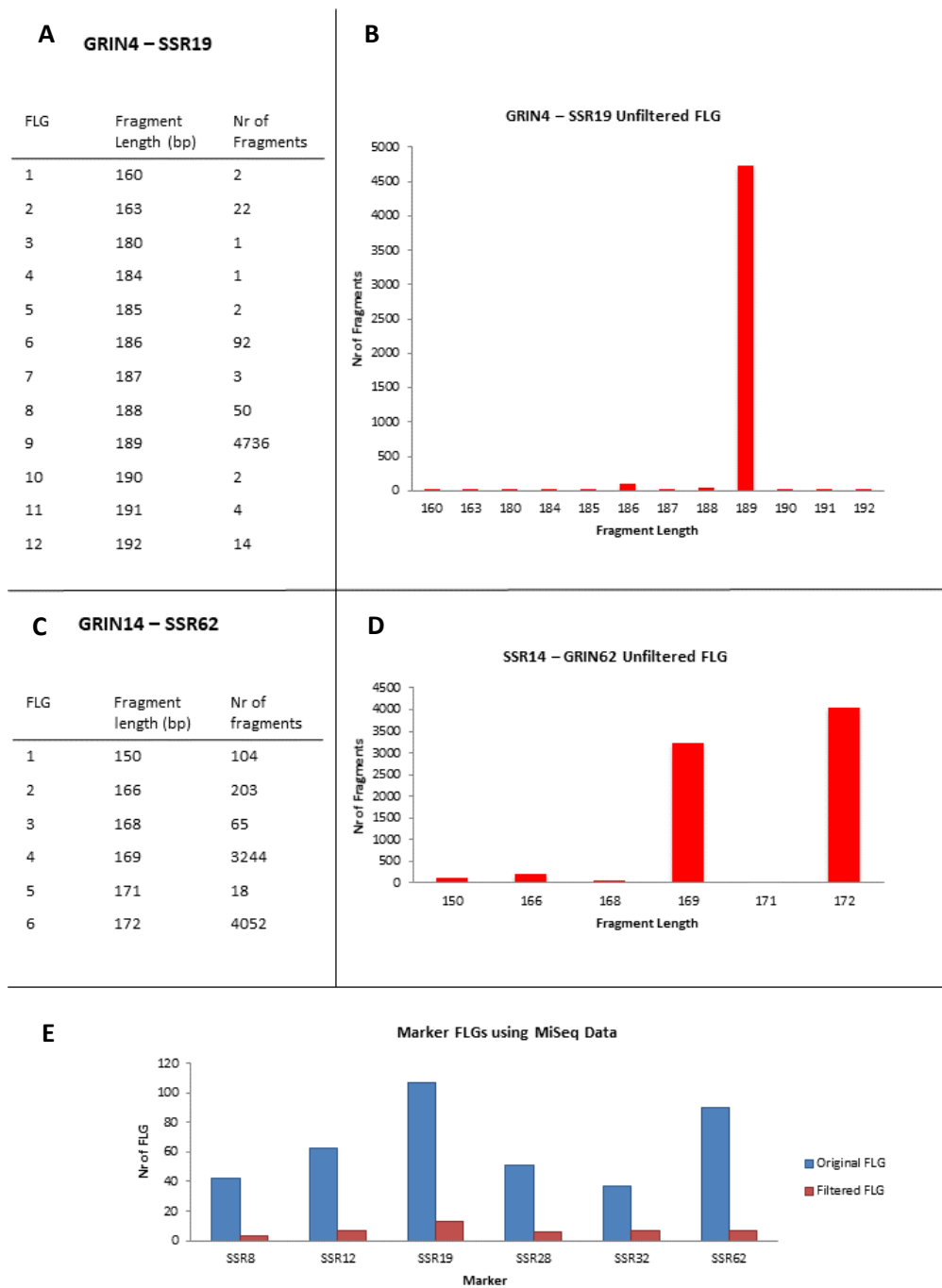


Figure 4.6: Representation of MAND script filtering steps for two marker loci. (A) A total of 12 Fragment Length Groups (FLGs) were identified for ATR19 in accession GRIN4 (B) Graphical representation of the number of fragments within each FLG. Since the number of fragments for FLG 189 is significantly higher than the remaining FLGs, it was assumed to be the true allele for ATR19 (C) A total of 6 FLGs were identified for ATR62 in accession GRIN14 (D) Graphical representation of the number of fragments within each FLG. Two FLGs show a significantly higher number of fragments within FLG 169 and 172; therefore, a heterozygote allele call was assumed (E) Original number of FLGs vs. number of filtered FLGs for each locus across the germplasm set. The FLGs were filtered based on whether a specific FLG has less than 40% of the number of fragments found in the largest FLG.

Table 4.6: Genotyping calls using Illumina NGS MiSeq technology and the MAND script

Accession	ATR8		ATR12		ATR19		ATR28		ATR32		ATR62	
	Allele 1	Allele 2	Allele 1	Allele 2	Allele 1	Allele 2	Allele 1	Allele 2	Allele 1	Allele 2	Allele 1	Allele 2
GRIN1	110	110	178	178	194	194	139	139	124	124	150	150
GRIN2	110	110	178	178	188	188	148	148	124	124	138	138
GRIN3	98	98	181	181	185	185	127	127	121	121	169	169
GRIN4	98	98	181	181	209	209	136	136	121	121	169	169
GRIN5	98	98	181	181	185	185	127	127	121	121	169	169
GRIN6	98	98	184	184	212	212	136	136	124	124	169	169
GRIN7	98	98	181	181	185	185	127	127	121	121	169	169
GRIN8	98	98	184	184	233	233	136	136	124	124	169	169
GRIN9	98	98	184	184	203	203	136	136	124	124	169	169
GRIN10	98	98	184	184	233	233	136	136	124	124	169	169
GRIN11	98	98	160	160	185	185	136	136	121	121	169	169
GRIN12	98	98	160	160	185	185	136	136	121	121	169	169
GRIN13	98	98	160	160	185	185	136	136	121	121	169	169
GRIN14	98	98	160	160	185	185	136	136	121	121	169	172
GRIN15	104	104	163	163	203	203	142	142	148	148	138	138
GRIN16	104	104	187	187	203	203	142	142	130	130	129	129
GRIN17	98	98	190	190	206	206	136	136	124	124	169	169
GRIN18	98	98	181	181	188	188	136	136	127	127	169	169
GRIN19	98	98	184	184	209	209	136	136	124	124	169	169
GRIN20	98	98	184	184	209	209	136	136	124	124	169	169
GRIN21	98	98	181	181	203	203	136	136	121	121	169	169
GRIN22	98	98	190	190	209	209	136	136	124	124	169	169
GRIN23	98	98	190	190	212	212	136	136	124	124	169	169
GRIN24	98	98	190	190	212	212	136	136	124	124	169	169
GRIN25	98	98	190	190	209	209	136	136	124	124	169	169
GRIN26	98	98	184	184	215	215	136	136	124	124	169	169
GRIN27	98	98	190	190	212	212	136	136	124	124	169	169
GRIN28	98	98	181	181	188	188	136	136	124	124	169	169
GRIN29	98	98	187	187	188	188	136	136	127	127	169	169
GRIN30	98	98	187	187	188	188	136	136	127	127	169	169
GRIN31	98	98	187	187	188	188	136	136	121	121	169	169
GRIN32	98	98	181	181	185	185	136	136	121	121	169	169
GRIN33	98	98	181	181	188	188	136	136	124	124	169	169
GRIN34	98	98	181	181	188	188	136	136	124	124	169	169
GRIN35	98	98	181	181	188	188	136	136	124	124	169	169
GRIN36	98	98	160	160	0	0	136	136	0	0	172	172
GRIN37	98	98	160	160	0	0	136	136	0	0	172	172
GRIN38	98	98	160	160	0	0	136	136	0	0	172	172
GRIN39	110	110	187	187	200	200	148	148	133	133	150	150
GRIN40	110	110	184	184	209	209	148	148	133	133	150	150
GRIN41	110	110	163	163	218	218	154	154	133	133	159	159
GRIN42	98	98	163	163	215	215	136	136	139	139	139	139
GRIN43	98	98	163	163	191	191	136	136	139	139	139	139
GRIN44	98	98	163	163	209	209	136	136	139	139	130	130
GRIN45	98	98	163	163	197	197	136	136	139	139	133	133

A value of "0" indicates a missing genotype/null allele.

Genetic diversity of *Amaranthus* as revealed by ABI capillary electrophoresis compared to MiSeq NGS technologies

Across the six SSR markers, 46 alleles were identified for the capillary electrophoresis analysis (an average of 7.67 alleles per locus), while 45 alleles were identified through MiSeq analysis (an average of 7.50 alleles per locus). The allele frequencies for each marker across the accessions are presented in Figure 4.7A (ABI capillary electrophoresis) and Figure 4.7B (MiSeq NGS). During the ABI analysis, the least number of alleles were identified for ATR8 (3), while ATR19 had the highest number of alleles (14). The allele frequencies within the dataset were reasonably uniform for markers ATR12 and ATR19, while the remainder (ATR8, ATR28, ATR32 and ATR62) each had at least one allele with a two- to four-fold higher occurrence than the remainder. In the MiSeq analysis, the least number of alleles were found for ATR8 (3) and the highest number for ATR19 (13). Only ATR12 indicated an even distribution of different allele frequencies across the different accessions, while the remainder of the loci revealed one or more common alleles that have a consistent higher frequency across the species groups. Five alleles were identified during chromatogram analysis using the ABI capillary electrophoresis technology, which were not found in the MiSeq analysis (Figure 4.7, indicated by red stars) namely ATR12 (allele 185), ATR19 (alleles 186, 198), ATR28 (allele 134) and ATR62 (alleles 130).

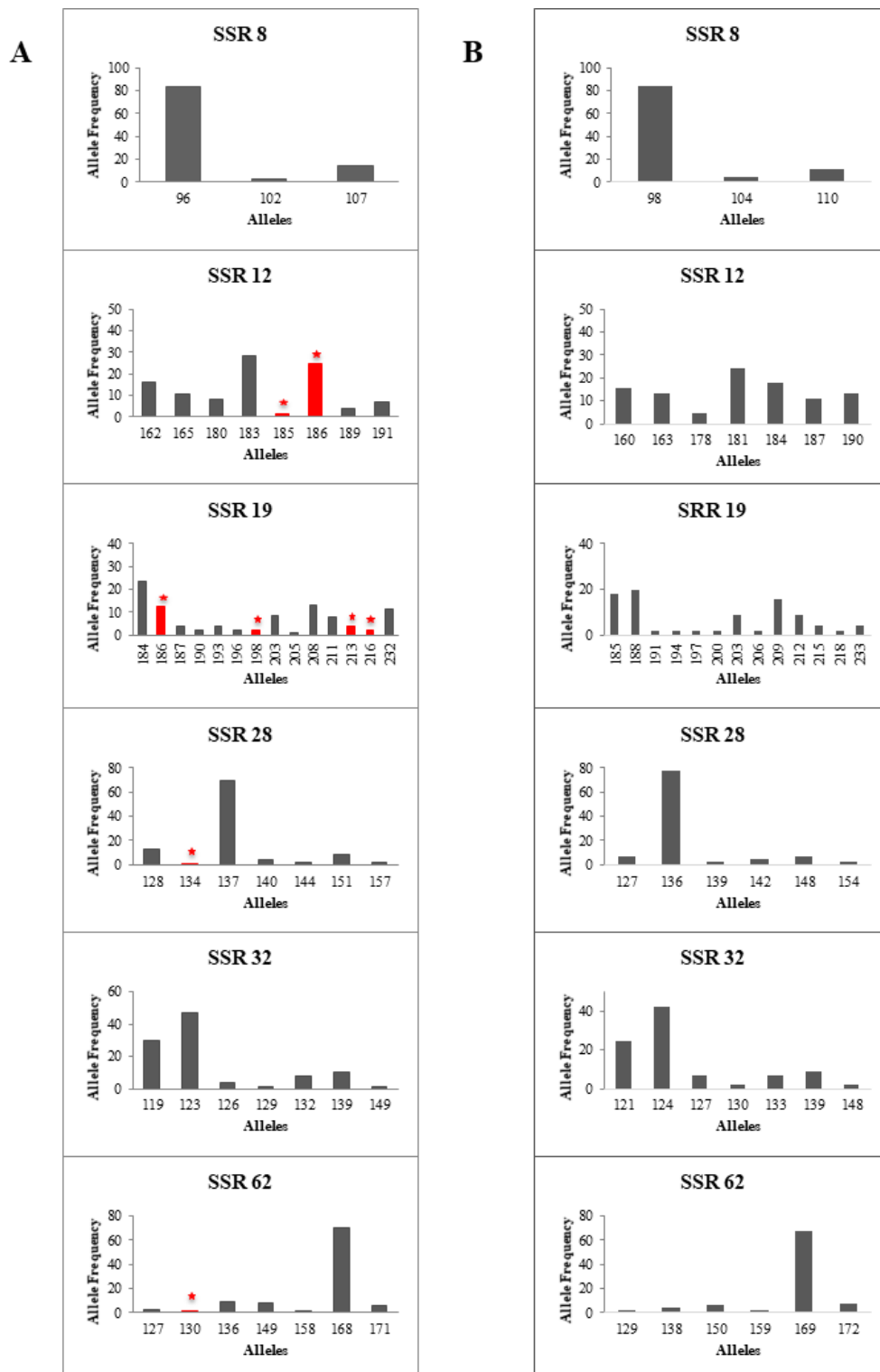


Figure 4.7: Allele frequencies observed during (A) ABI capillary electrophoresis analysis and (B) NGS MiSeq analysis. In both cases, the least number of alleles were observed for ATR8 (3) and the highest number of alleles for ATR19 (14 – ABI; 13 - NGS). Allele frequencies were mostly evenly distributed for ATR12 (ABI and NGS) and ATR19 (ABI), but one or more dominant alleles could be observed for the remainder of the markers. Starred red bars represent alleles that were identified during the ABI capillary electrophoresis analysis, but not during the NGS MiSeq analysis.

Globally, the observed and expected heterozygosity, PIC values and fixation indices were highly comparable between the ABI capillary electrophoresis and NGS MiSeq technologies. For both analyses, the observed heterozygosity remains close to zero across all loci, while the expected heterozygosity ranges from 0.276 to 0.891 (mean 0.612). The PIC values ranged from 0.252 to 0.869 (mean 0.584), indicating intermediate to high genetic diversity at each locus. The highest PIC value was observed for ATR19, which was to be expected since it produced the highest number of alleles. The fixation indices were calculated as one, for all loci except ATR62, which was 0.958, indicating highly inbred/homozygous species accessions (Table 4.7).

Table 4.7: Genetic Diversity, PIC values and Fixation indices for ABI vs. MiSeq data					
ABI					
Locus	Number of Alleles	Observed Heterozygosity	Expected Heterozygosity	PIC Value	Fixation Index
ATR8	3	0	0.276	0.252	1
ATR12	8	0	0.844	0.813	1
ATR19	14	0	0.891	0.869	1
ATR28	7	0	0.421	0.403	1
ATR32	7	0	0.715	0.667	1
ATR62	7	0.022	0.525	0.499	0.957
Mean	7.667	0.004	0.612	0.584	0.993
MiSeq					
Locus	Number of Alleles	Observed Heterozygosity	Expected Heterozygosity	PIC Value	Fixation Index
ATR8	3	0	0.276	0.252	1
ATR12	7	0	0.844	0.813	1
ATR19	13	0	0.874	0.85	1
ATR 28	6	0	0.388	0.369	1
ATR32	7	0	0.715	0.667	1
ATR62	9	0.022	0.53	0.509	0.958
Mean	7.50	0.00	0.60	0.58	0.99

Genetic diversity of the genus *Amaranthus* revealed by MiSeq analysis of polymorphic SSR loci

Fifty-nine amaranth accessions (GRIN and SAG germplasm sets) were genotyped using six polymorphic SSR markers. Species groups were defined according to the species name assigned to the accession by the GRIN database (Table 4.1, Table 4.2), resulting in 13 different groups. The unknown SAG accessions were hypothesized to belong to a specific species group based on preliminary morphological analysis (data not shown).

Based on the calculated genetic distance between all amaranth accessions, a principle component analysis plot was constructed where principle component 1 (x-axis) accounted for 21.90% of the variance, while principle component 2 (y-axis) accounted 13.56% of the variance (Figure 4.8). The genetic relatedness of the accessions within the species groups largely verified the existing taxonomic designations initially assigned. In total, nine of the 13 species groups originally described could be observed during PCA analysis (Figure 4.8).

Principal Coordinates (PCoA)

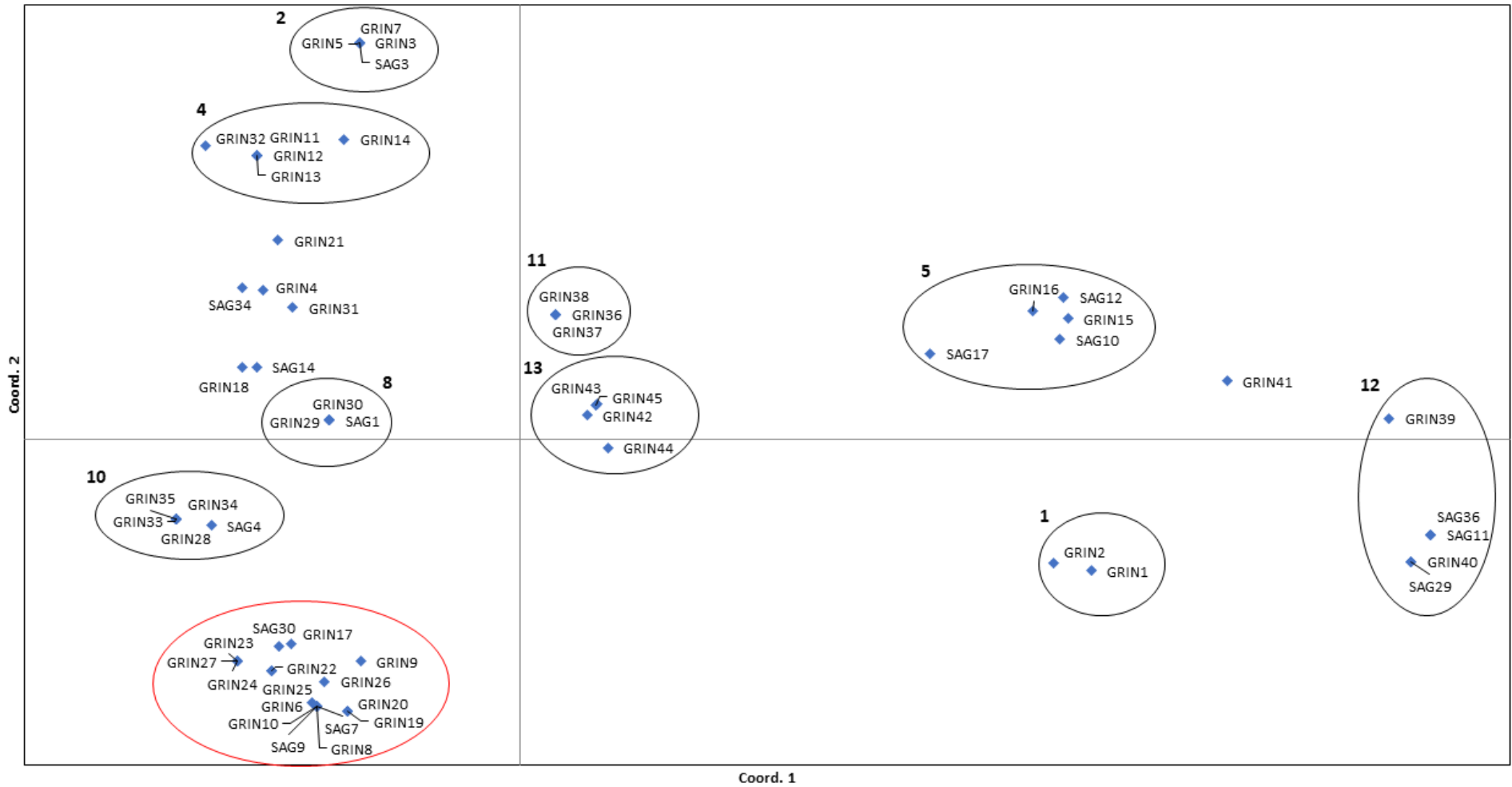


Figure 4.8: Principle component analysis of 59 *Amaranthus* germplasm accessions. Nine of the 13 initially described species groups were validated through clustering PCoA analysis. These include *A. blitum* (1), *A. caudatus* (2), *A. dubius* (4), *A. graecizans* (5), *A. powellii* (8), *A. retroflexus* (10), *A. spinosus* (11), *A. tricolor* (12) and *A. viridis* (13). The cluster indicated by a red circle indicates a tight cluster formation of accessions from different species groups, namely *A. hypochondriacus* (7), *A. cruentus* (3) and *A. hybridus* (6). The *A. quitensis* (9) accessions were divided, with one accession un-clustered (GRIN 31) and the other (GRIN 32) grouping with *A. dubius* accessions. Approximately 57% of the unknown SAG individuals could be assigned a putative identity based on their grouping with known accessions.

Accessions of *A. blitum* (1), *A. caudatus* (2), *A. dubius* (4), *A. graecizans* (5), *A. powellii* (8), *A. retroflexus* (10), *A. spinosus* (11), *A. tricolor* (12) and *A. viridis* (13) form individual clusters as expected based on the morphological identifications. Accessions from three particular species groups; *A. hypochondriacus* (7), *A. cruentus* (3) and *A. hybridus* (6) are clustered tightly together and cannot be accurately distinguished from one another. The two *A. quitensis* accessions are split, with one accession (GRIN 32) showing a high correlation with the *A. dubius* accessions and the other (GRIN 31) not allocated to an individual group. There is a high correlation/relationship and low genetic diversity between species groups 2, 4, 8 and 10. An intermediary cluster is formed by groups 11 and 13. Groups 1, 5 and 12 are more closely related to one another, than to the aforementioned groups. Eight (57%) of the unknown SAG amaranth accessions clustered with known GRIN accessions (Figure 4.8). In particular, SAG 3 clustered with the *A. caudatus* species (2); SAG 1 with *A. powellii* (8), SAG 4 with *A. retroflexus* (10), SAG 10/SAG 12/SAG 17 with *A. graecizans* (5) and SAG 11/SAG 29/SAG 36 with *A. tricolor* (12). Three SAG accessions (SAG 7/SAG 9/SAG 30) grouped with the mixed *A. hypochondriacus*/*A. cruentus*/*A. hybridus* cluster (Figure 4.8, red circle), while two accessions were not assigned to a specific cluster (SAG 14/ SAG 34).

Dendrogram analysis of all accessions revealed 11 individual clades (Figure 4.9). Bootstrap (BS) confidence levels of above 70% supported approximately 60% of the total nodes. *Amaranthus blitum* (1) forms a sister clade to *A. tricolor* (12) (BS<70%). Within the *A. tricolor* cluster, three SAG accessions are found that group with BS=87% (SAG29) and 98% (SAG 11/SAG 36) and therefore strongly hypothesized to be *A. tricolor*. A close sister group to *A. blitum*/*A. tricolor* was composed of accessions from *A. graecizans* (5) and *A. viridis* (13), where GRIN 15 (*A. graecizans*) and GRIN 43/GRIN 45 (*A. viridis*) form a sister clade to GRIN 16 (*A.*

graecizans) and GRIN 42/ GRIN 44 (*A. viridis*) with BS=80%. The aforementioned species are separated from a large sister clade by BS=87%. The sister clade is composed of *A. caudatus* (2) and SAG 3 (BS=99%, hypothesized to be *A. caudatus*), which shares a common ancestor with sub-clades *A. retroflexus* (10), *A. powellii* (8) and *A. quitensis* (9). The *A. powellii* (GRIN 31/GRIN 32) accessions group strongly with SAG 1 (BS=77%), which implies a classification as *A. powellii*. *Amaranthus quitensis* and *A. powellii* form sister clades (BS<70%), which in turn are closely related to *A. retroflexus* (BS<70%). Interestingly, GRIN 28 (formerly identified as *A. powellii*) now groups strongly with the *A. retroflexus* (GRIN 33) accession (BS=85%) and this clustering was also observed during the PCoA analysis. It is therefore proposed that the GRIN 28 accession had been miss-classified in the germplasm database and should rather have the identity of *A. retroflexus*. Sharing a common ancestor with *A. quitensis*/*A. retroflexus*/ *A. powellii* is *A. caudatus* (BS=76%), together with the unknown SAG 3 (possibly *A. caudatus*) accession (BS=99%). The *A. hypochondriacus* clade is not highly supported (BS=<70%), although terminal nodes do show an average of BS=78%. The GRIN 17 (previously identified as *A. hybridus*) accession also groups within this clade (albeit with low bootstrap support – BS<70%) but could putatively be re-classified to *A. hypochondriacus*. The unknown SAG 30 accession also groups together with *A. hypochondriacus* accessions (BS<70%). *Amaranthus hypochondriacus* (7) forms a sister clade to a mixed group (MP1), which is composed of different accessions including one *A. caudatus*, one *A. cruentus*, two *A. hybridus*, one *A. hypochondriacus* and unknown SAG 4. Due to the grouping of multiple species in one clade, the identification of SAG 4 remains elusive. The closest related sister clade to *A. hypochondriacus*/MP1 is composed of *A. dubius* (BS<70%), followed by an *A. cruentus* clade. Unknown SAG 34 was re-classified to *A. dubius* based on the clustering with *A. dubius* (GRIN 11/GRIN 12/ GRIN 13/ GRIN 14) accessions (BS=88%), while accessions SAG 7 and SAG 9 were re-classified as *A. cruentus* (BS=87%). Separated from all the aforementioned species (BS=100%) is *A. spinosus*, which forms a unique, solitary clade (Figure 4.9).

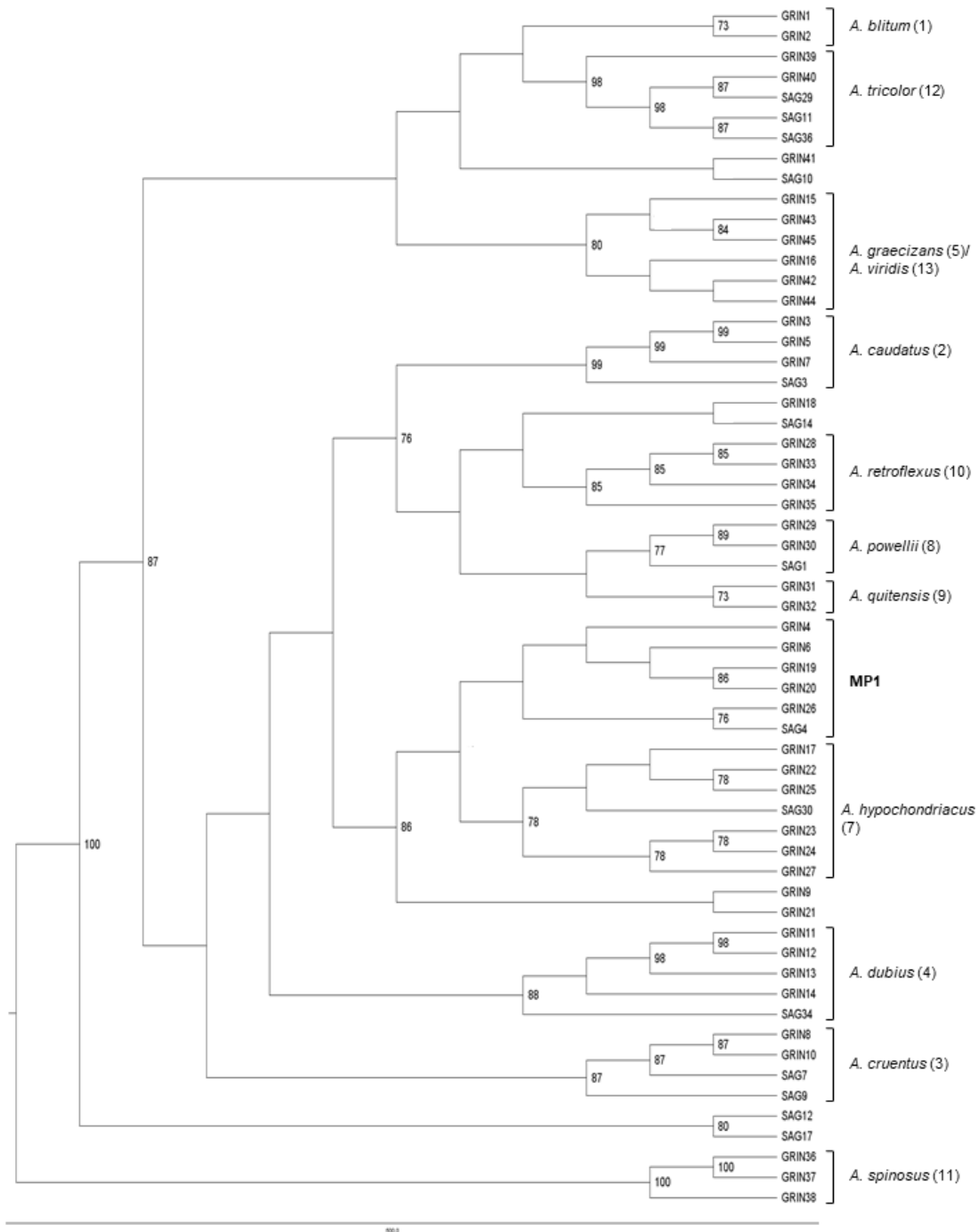


Figure 4.9: Dendrogram analysis of 59 diverse *Amaranthus* accessions based on MiSeq SSR data. The dendrogram was generated using DendroUPGMA (online software <http://genomes.urv.cat/UPGMA/>) applying Pearson's coefficient with 1000 bootstrap replicates. Eleven clades could be identified, which correspond closely to the initially described species groupings (Table 4.8). One clade was considered a mixed group (MP1); composed of accessions from different species including *A. caudatus*, *A. cruentus*, *A. hybridus*, *A. hypochondriacus* and unknown SAG 4. Species group number is indicated in brackets after species name.

Allele frequency analysis revealed the addition of four unique alleles, only found among the SAG accessions. Two additional loci were observed for ATR12 (166 – SAG10 and SAG12; 172 – SAG17). One additional locus was identified for ATR28 (145 – SAG10 and SAG12) and ATR62 (153 – SAG12). The total number alleles observed across the different species groups rose from 45 to 49 with the addition of SAG accessions. The observed heterozygosity ranged from zero to 0.037 (mean 0.018), while the expected heterozygosity ranged from 0.370 to 0.885 (mean 0.640). The PIC value had a mean of 0.611, while the fixation index had a mean of 0.949 (Table 4.8). Eighteen private alleles were identified over seven species groups (Figure 8). The highest number of private alleles were observed for *A. tricolor* (5), followed by *A. blitum* (3), *A. powellii* (3), *A. retroflexus* (3), *A. quitensis* (2), *A. dubius* (1) and *A. cruentus* (1). In total, ATR62 contributed 33% of the private alleles across all species groups, followed by ATR19 (28%), ATR12/ATR28 (17%), ATR32 (6%). No private alleles were contributed by ATR8 (Figure 4.10).

Table 4.8: Genetic diversity, PIC values and fixation indices for *Amaranthus* germplasm set based on MiSeq SSR data after the addition of

Locus	Number of Alleles	Observed Heterozygosity	Expected Heterozygosity	PIC Value	Fixation Index
ATR8	3	0	0.370	0.335	0.993
ATR12	9	0	0.847	0.820	1.000
ATR19	14	0.036	0.885	0.865	0.921
ATR28	7	0	0.476	0.450	0.998
ATR32	7	0.037	0.733	0.691	0.903
ATR62	10	0.034	0.527	0.506	0.877
Mean	8.333	0.018	0.640	0.611	0.949

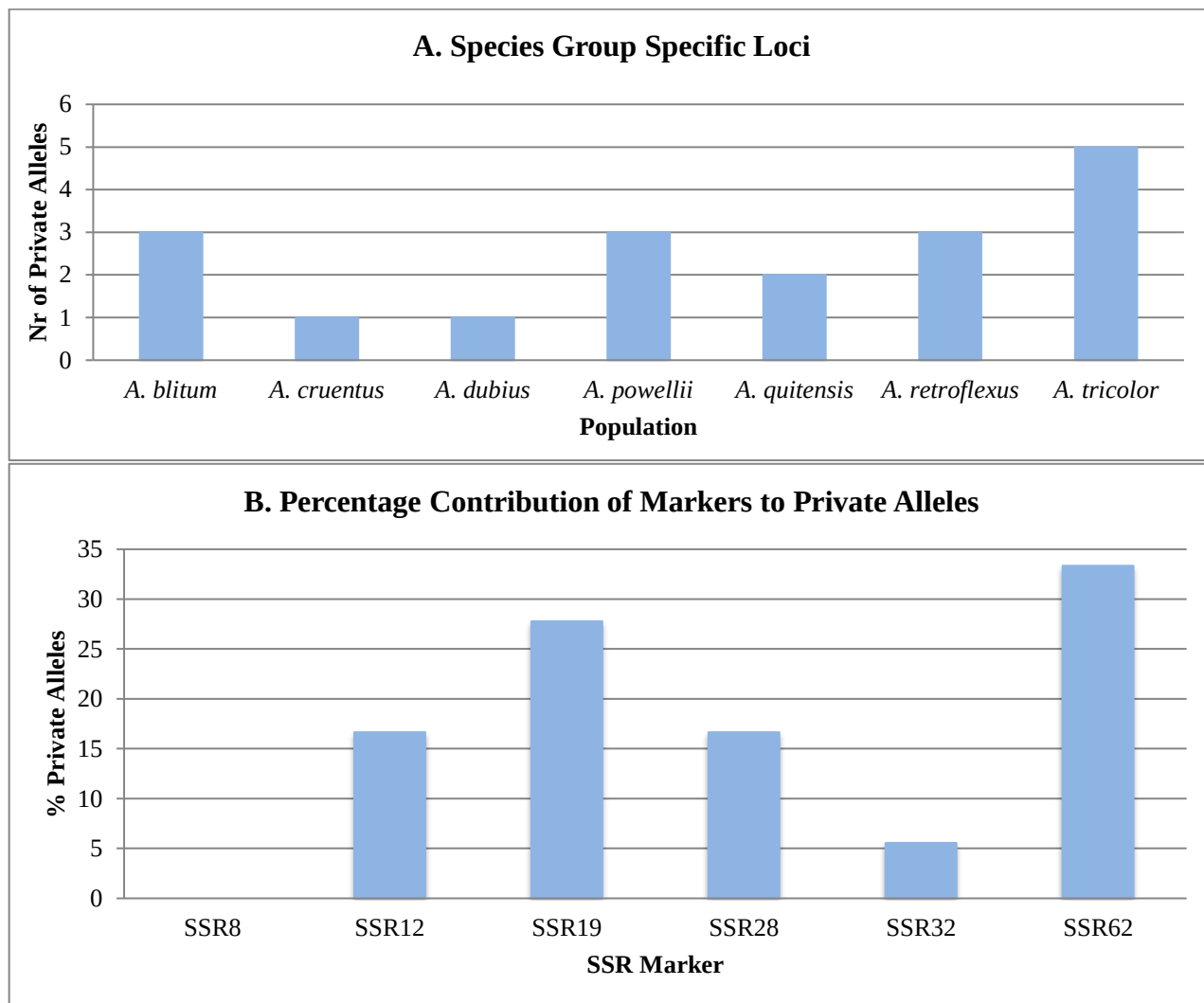


Figure 4.10: Loci unique to a species group (A), and the percentage contribution of each marker to the private allele set (B) across the entire germplasm set (GRIN and SAG). The *A. tricolor* grouping had five unique alleles, followed by *A. blitum*/*A. powellii*/*A. retroflexus* with three unique alleles each; *A. quitensis* with 2 unique alleles and *A. cruentus*/*A. dubius* with 1 unique allele each. The remainder of the species groups did not reveal the presence of private alleles. In total, ATR62 contributed 33% of the private alleles, while ATR19; ATR12/ATR28; ATR32 and ATR8 contributed 28%, 17%, 6% and 0% respectively.

Discussion

A draft genome sequence was assembled using the Illumina MiSeq technology for an important leafy vegetable, *Amaranthus tricolor*. The assembly spans approximately 402 Mb (45.6%) of the entire genome with 44x coverage, based on a comparison to an estimated genome size of 880 Mb published previously for *A. tricolor* (Sunil *et al.* 2014). The remaining 14% of the genome not covered by this assembly is thought to represent highly repetitive regions along the centromeric and telomeric regions of the genome, which are notoriously difficult to assemble. These results compare favorably to another published genome of *A. hypochondriacus*, where the final assembly was composed of 13 569 scaffolds, estimated to span 431.8 (92.6%) of the estimate *A. hypochondriacus* genome size of 466 Mb (Clouse *et al.*, 2016). Recently, an even more comprehensive genome assembly of *A. hypochondriacus* was published using the previously published sequencing data (Sunil *et al.*, 2014, Clouse *et al.*, 2016), and supplementing it with PacBio sequencing and chromatin interaction mapping (Hi-C) to further assemble scaffolds. The resulting assembly was composed of 16 chromosome-sized scaffolds (403.9 Mb), assumed to correspond to the haploid chromosomes ($n=16$) of the genome (Lightfoot *et al.*, 2017). *Amaranthus hypochondriacus* is a close relative to *A. tricolor* (sequenced in the present study), and the availability of both genomes will enable researchers to perform in-depth syntenic studies for future crop enhancement.

Microsatellite loci, which conformed to the following criteria, were retained for further testing and analysis; (i) primer sequences could be obtained up-and downstream of the particular locus (ii) internal BLAST analysis to the draft genome does not indicate alternative binding sites of the primers (iii) only a single band could be observed after PCR amplification (iv) the loci could be amplified with a standard PCR protocol across different species groups and (v) Sanger sequencing of the allele regions confirmed the predicted number of repeats. During the present study, 6 loci fulfilled the requirements of the abovementioned validation checks and should therefore be robust for future studies across the *Amaranthus* genus.

For each marker, the allele frequencies, observed and expected heterozygosities, presence of null alleles and PIC values are highly comparable (Table 4.7, Table 4.8). However, upon closer inspection it was found that the MiSeq data consistently scores alleles in sizes related to the repeat size (e.g. multiples of three for a trinucleotide repeat) whereas the ABI data often showed “in-between” markers, that were called probably due

to shift mutations or sequence stuttering. Consequently, five alleles called during the ABI analysis were not observed during the MiSeq analysis. Since the final allele calls/frequencies differ slightly between the techniques, it is proposed that the MiSeq analysis more accurately binned the “extra” markers to their correct allele size. No significant difference could be observed for the observed/expected heterozygosities and PIC values between the two techniques; both indicate that the germplasm set is highly inbred, and the presence of heterozygous alleles is extremely low (only found for one accession in marker ATR62). The PIC values were relatively high across the marker loci (except ATR8 – PIC 0.252), indicating that the markers have a high polymorphic content within the amaranth species groups. Null alleles were consistently scored for two markers (ATR19 and ATR32) for the same accessions (GRIN 36, GRIN 37 and GRIN 38 – *A. spinosus*). However, one null allele in the ABI analysis was confidently scored during the MiSeq analysis (ATR19 – GRIN 31).

The GRIN germplasm set was found to be highly inbred, which was to be expected from lines that have been self-fertilized over many generations in a germplasm bank. Interestingly, even though these accessions were collected from different parts of the world, their genetic information remains stable. Based on the MiSeq data, little to no variation was detected within an individual (1%), and only moderate variation between accessions within a species group (38%). For *A. tricolor* specifically, 4 of the 6 markers show variation between the three different accessions (GRIN39, GRIN40 and GRIN41). The within species variation fluctuates between the different species groups, very little to no variation was observed for accessions of *A. cruentus*, *A. dubius*, *A. retroflexus* and *A. spinosus*. Approximately 50%-60% variation was observed for accessions of *A. blitum*, *A. caudatus*, *A. graecizans*, *A. hypochondriacus*, *A. hybridus*, *A. powellii*, *A. quitensis* and *A. viridis*. These results indicate that the markers currently could be used for diversity analysis in some of the species groups, but additional markers will have to be developed for intra-species variation studies for the remainder.

The largest variance was observed between the different species groups (61%). However, the addition of “wild” SAG accessions collected from the field reveals the presence of novel private alleles across the accessions. The largest number of private alleles was found for *A. tricolor* (5 alleles), followed by *A. powellii*/*A. blitum* (3 alleles each), *A. quitensis* (2 alleles) and *A. cruentus*/*A. dubius* (1 allele each). Interestingly, two species belonging to the grain amaranth group (*A. caudatus* and *A. hypochondriacus*) did not reveal the presence of private alleles in their “wild” SAG counterparts, indicating that the genetic makeup of these

species have become stable during their domestication. However, the mostly undomesticated leafy/wild amaranth species show the addition of several unique, novel alleles, possibly indicating cross hybridization with other local accessions. With the addition of the novel alleles present in the SAG species, the within individual variation increased slightly to 2%, while the variation between accessions of a species group rose significantly to 75%. The variation among groups fell steeply to only 23%, possibly substantiating the claim that there has been inter-species cross hybridization and that wild species are now sharing more of their genetic content. The variation can also be observed in the number of heterozygous individuals (four SAG accession vs. one GRIN accession).

The basic relationships between *Amaranthus* species, as revealed by dendrogram analysis can be described as follows (Figure 4.9): *A. blitum* and *A. tricolor* (leafy amaranths) are closely related (BS<70%), followed by a sister group composed of *A. graecizans* and *A. viridis* (leafy amaranths) (BS<70%). This relationship has also been described by Chan and Sun (1997) using RAPD analysis, as well as by Stetter and Schmid (2017) during whole genome SNP analysis. *Amaranthus caudatus/A. retroflexus/A. powellii/A. quitensis* share a common ancestor with *A. hypochondriacus*/Clade MP1, while the *A. blitum/A. tricolor/A. viridis/ A. graecizans* clades are grouped separately to the aforementioned clades. Most of the uncultivated (wild) amaranths are clustered together (*A. powellii/A. retroflexus/A. quitensis*) and forms a sister clade to grain cultivated species, *A. caudatus* (BS=76%). The result is relatively unique, since previous studies all show the separate clustering of the wild amaranths to grain types (Chan and Sun 1997; Sun *et al.* 1999; Xu and Sun 2001; Mallory *et al.* 2008; Stetter and Schmid 2017). The discrepancies could possibly be explained by the poor reproducibility and inherent user screening bias of previously used RAPD and AFLP analysis, as well as smaller and less diverse samples sets utilized during previous studies. Interestingly, one clade (MP1) is composed of several different species (one *A. caudatus* – GRIN 4, one *A. cruentus* – GRIN 6, two *A. hybridus* – GRIN 19/GRIN 20, one *A. hypochondriacus* – GRIN 26 and an unknown SAG 4). The grouping of these species is often referred to the *A. hybridus* complex (composed of *A. hypochondriacus*, *A. caudatus*, *A. cruentus*, *A. hybridus* and *A. quitensis*) and their exact genetic relatedness has eluded researchers for some time (Chan and Sun 1997; Sun *et al.* 1999; Xu and Sun 2001; Mallory *et al.* 2008; Kietlinski *et al.*, 2014; Stetter and Schmid 2017). As in Kietlinski *et al* (2014), *A. quitensis* forms a separate sub-clade to the remaining members of the *A. hybridus* complex, while the *A. hybridus* accessions are scattered across clades containing *A. caudatus*, *A.*

cruentus and *A. hypochondriacus*. Their widespread groupings could indicate that different varieties/sub-species of *A. hybridus* exists, or that multiple morphologically cryptic species have evolved over time (Kietlinski *et al.*, 2014). Two accessions (GRIN 19 and GRIN 26), as well as SAG 4, were collected in central Mexico, adding confidence to the assumption that they are related. *Amaranthus dubius* and *A. cruentus* formed sister sub-clades but are not highly supported (BS<70%). During previous studies, the *A. dubius* sub-clade consistently groups with the grain (*A. hypochondriacus*/*A. cruentus*/*A. caudatus*) and wild (*A. retroflexus*/*A. powellii*/*A. quitensis*) accessions (Chan and Sun 1997; Stetter and Schmid 2017). Lastly, *A. spinosus* clustered completely separately from the other species indicating that these accessions are least related to the remainder of the germplasm set. Of the 14 unknown SAG accessions, 70% could be assigned a putative identity based on their clustering with known GRIN accessions. The identity of *A. tricolor* could be assigned to SAG 29, SAG 11 and SAG 36 with high confidence (BS=98%), while SAG 3, SAG 1, SAG 30, SAG 34 and SAG7/SAG 9 could be assigned identities of *A. caudatus* (BS=99%), *A. powellii* (BS=77%), *A. hypochondriacus* (BS<70%), *A. dubius* (BS=88%) and *A. cruentus* (BS=87%) respectively. Identities could not be assigned to SAG 10, SAG 4, SAG 12 and SAG 17, due to non-specific clustering. In addition, two GRIN accessions were re-classified based on their clade groupings; GRIN 28 (previously *A. powellii*) groups strongly with *A. retroflexus* accessions (BS=85%), and GRIN 17 (previously *A. hybridus*) groups with *A. hypochondriacus* accessions (BS<70%) (Figure 4.9).

The PCoA analysis revealed nine stable groupings of species, corresponding to their morphological descriptions (Figure 4.8). A higher variance exists between *A. caudatus* (2)/ *A. dubius* (4)/ *A. powellii* (8)/ *A. retroflexus* (10) and *A. blitum* (1)/ *A. graecizans* (5)/ *A. tricolor* (12) than either groups to *A. spinosus* (11) or *A. viridis* (13). The relationships suggested by the PCoA analysis were closely mirrored during the dendrogram analysis, indicating that the tree topology within this genus is highly robust when applying microsatellite markers.

Conclusion

The advent of high throughput sequencing technologies has consistently yielded highly informative, large datasets in a fraction of the time and at a significantly reduced cost than Sanger sequencing platforms used in the past. The construction of whole genomes has virtually become a routine starting point for several downstream applications such as gene discovery, marker analysis and selective breeding. During the present study, the hypothesis that sequencing and assembly of a draft genome for *A. tricolor* would enable the efficient identification of a high number of potential SSR loci, and through a set of validation techniques; a subset of markers would be found to be robust across several different species groups was accepted. A direct comparison of allele information between the routinely used capillary electrophoresis ABI and next generation MiSeq technologies revealed a highly significantly correlated dataset. Minor improvements to the SSR allele calls were observed during allele calling via the MiSeq technology. Coupled with the reduced time spent on sample preparation, decreased cost of screening highly multiplexed datasets and the highly informative data output; it supports the hypothesis that screening SSR markers using NGS would be a decidedly feasible route for future researchers. A general cost breakdown of the current experiment is given in Supplementary Table S4.3. The price tag associated with the screening of polymorphic SSR loci across a population is approximately 20% more when using the NGS route, compared to more conventional ABI fragment analysis. However, the screening of larger population sizes will lead to a decrease in associated costs, especially with the NGS route. However, the sample preparation and data analysis of NGS is significantly simplified and incorporates little to no human error. In addition, the quality and quantity of data output from NGS is far superior to ABI methods, since the actual sequence data is also available for downstream analysis.

In addition, the initial microsatellite locus dataset obtained from the draft genome sequence can be further investigated and exploited to add informative markers (SSR or SNP) to the current dataset, without incurring additional experimental costs. Downstream analysis of the *Amaranthus* germplasm set revealed highly robust PCoA and dendrogram results, which demonstrated the complex relationships between the different *Amaranthus* species groups. Even though only 6 markers were used for the genetic variation study, the hypothesis/objectives of identifying species relationships between previously described species and unknown accessions was supported. In addition, within species variation (based on three SSR markers) was observed for the different accessions of *A. tricolor*. If, in future, different varieties of *A. tricolor* need to be hybridized

to exploit beneficial traits, these specific SSRs could be utilized to follow QTL linked to favorable traits from one or the other parental line. It will, however, be necessary to identify additional SSR markers to ensure that the marker and QTLs are segregated together in the progeny.

The information on the species relatedness can be used during future marker selected breeding programs to breed superior genotypes.

References

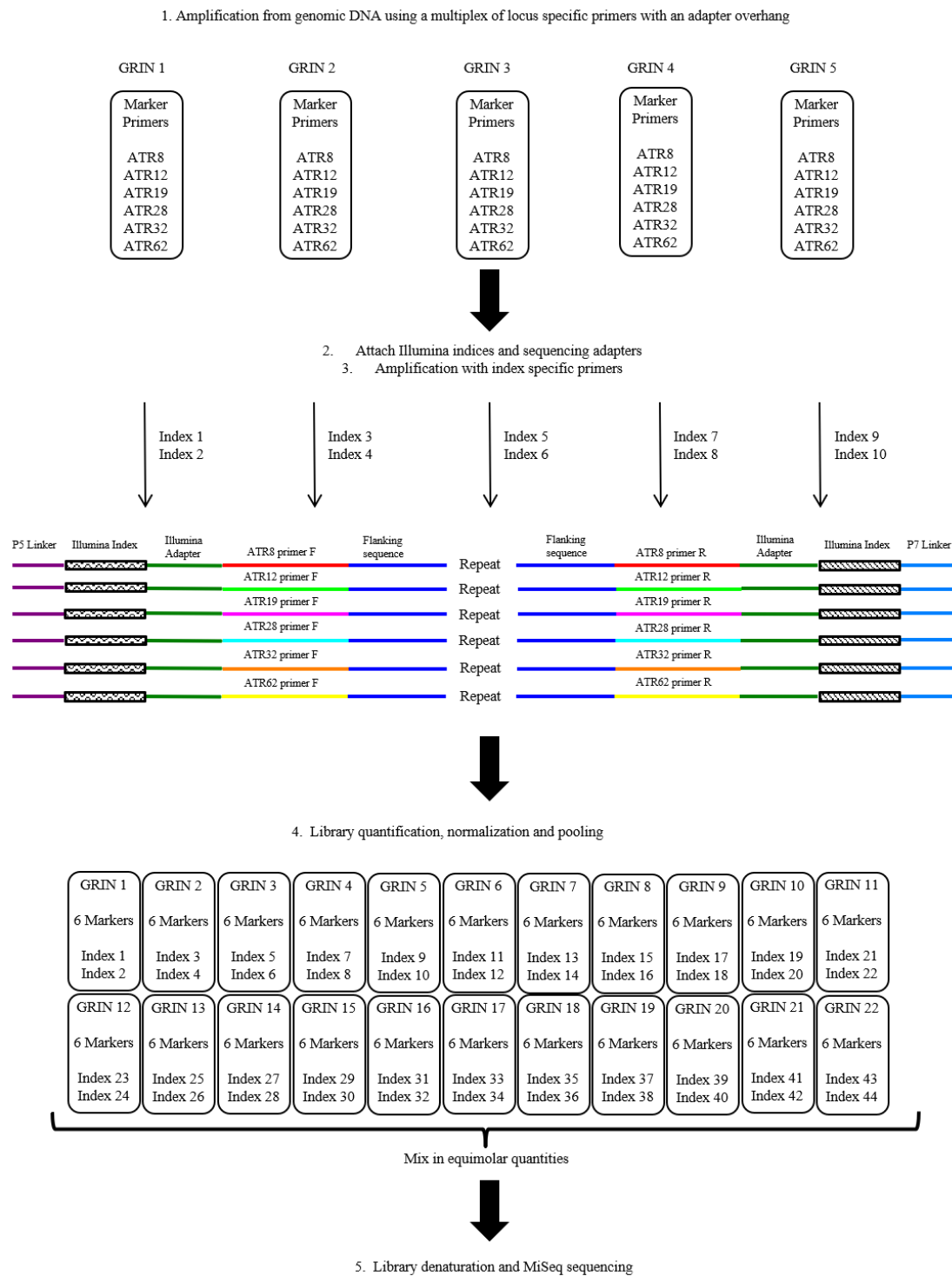
- Carlsson J *et al.* (2013) Rapid, economical single - nucleotide polymorphism and microsatellite discovery based on *de novo* assembly of a reduced representation genome in a non - model organism: a case study of Atlantic cod *Gadus morhua*. *Journal of Fish Biology* 82:944-958
- Chan K, Sun M (1997) Genetic diversity and relationships detected by isozyme and RAPD analysis of crop and wild species of *Amaranthus*. *Theoretical and Applied Genetics* 95:865-873
- Clouse J, Adhikary D, Page JT *et al.*, (2015) The Amaranth Genome: Genome, transcriptome, and physical map assembly. *The Plant Genome* 9
- Conesa A, Götz S (2008) Blast2GO: A comprehensive suite for functional analysis in plant genomics. *International Journal of Plant Genomics* 2008
- Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21:3674-3676
- 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
- Faircloth BC (2008) Msatcommander: detection of microsatellite repeat arrays and automated, locus - specific primer design. *Molecular Ecology Resources* 8:92-94
- Ferrandiz-Rovira M, Bigot T, Allainé D, Callait-Cardinal M, Cohas A (2015) Large-scale genotyping of highly polymorphic loci by next-generation sequencing: how to overcome the challenges to reliably genotype individuals? *Heredity* 114:485
- Garcia-Vallvé S, Puigbo P (2009) DendroUPGMA: A dendrogram construction utility. *Universitat Rovira i Virgili*
- Grohme MA, Soler RF, Wink M, Frohme M (2013) Microsatellite marker discovery using single molecule real-time circular consensus sequencing on the Pacific Biosciences RS. *BioTechniques* 55:253-256
- Gupta V, Gudu S (1991) Interspecific hybrids and possible phylogenetic relations in grain amaranths. *Euphytica* 52:33-38
- Kalia RK, Rai MK, Kalia S, Singh R, Dhawan A (2011) Microsatellite markers: an overview of the recent progress in plants. *Euphytica* 177:309-334

- Kietlinski KD, Jiminez F, Jellen EN, Maughan PJ, Smith SM, Pratt DB (2014) Relationships between the weedy *Amaranthus hybridus* (Amaranthaceae) and the grain amaranths. *Crop Science* 54:220-228
- Lightfoot D, Jarvis D, Ramaraj T, Lee R, Jellen E, Maughan P (2017) Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (*Amaranthus hypochondriacus*) chromosomes provide insights into genome evolution. *BMC Biology* 15:74
- López-Urbe MM, Santiago CK, Bogdanowicz SM, Danforth BN (2013) Discovery and characterization of microsatellites for the solitary bee *Colletes inaequalis* using Sanger and 454 pyrosequencing. *Apidologie* 44:163-172
- 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
- Mnkeni A, Masika P, Maphaha M (2007) Nutritional quality of vegetable and seed from different accessions of *Amaranthus* in South Africa. *Water SA* 33:377-380
- 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
- Omboki RB, Wu W, Xie X, Mamadou G (2015) Applications of genomics in Plant Breeding. *International Journal of Agriculture and Crop Sciences* 8:560
- Onda Y, Mochida K (2016) Exploring genetic diversity in plants using high-throughput sequencing techniques. *Current Genomics* 17:358-367
- Pal M, Khoshoo T (1973) Evolution and improvement of cultivated amaranths: VI. Cytogenetic relationships in grain types. *Theoretical and Applied Genetics* 43:242-251
- Patnaik BB et al. (2016) Sequencing, *de novo* assembly, and annotation of the transcriptome of the endangered freshwater pearl bivalve, *Cristaria plicata*, provides novel insights into functional genes and marker discovery. *PLoS One* 11:e0148622
- 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

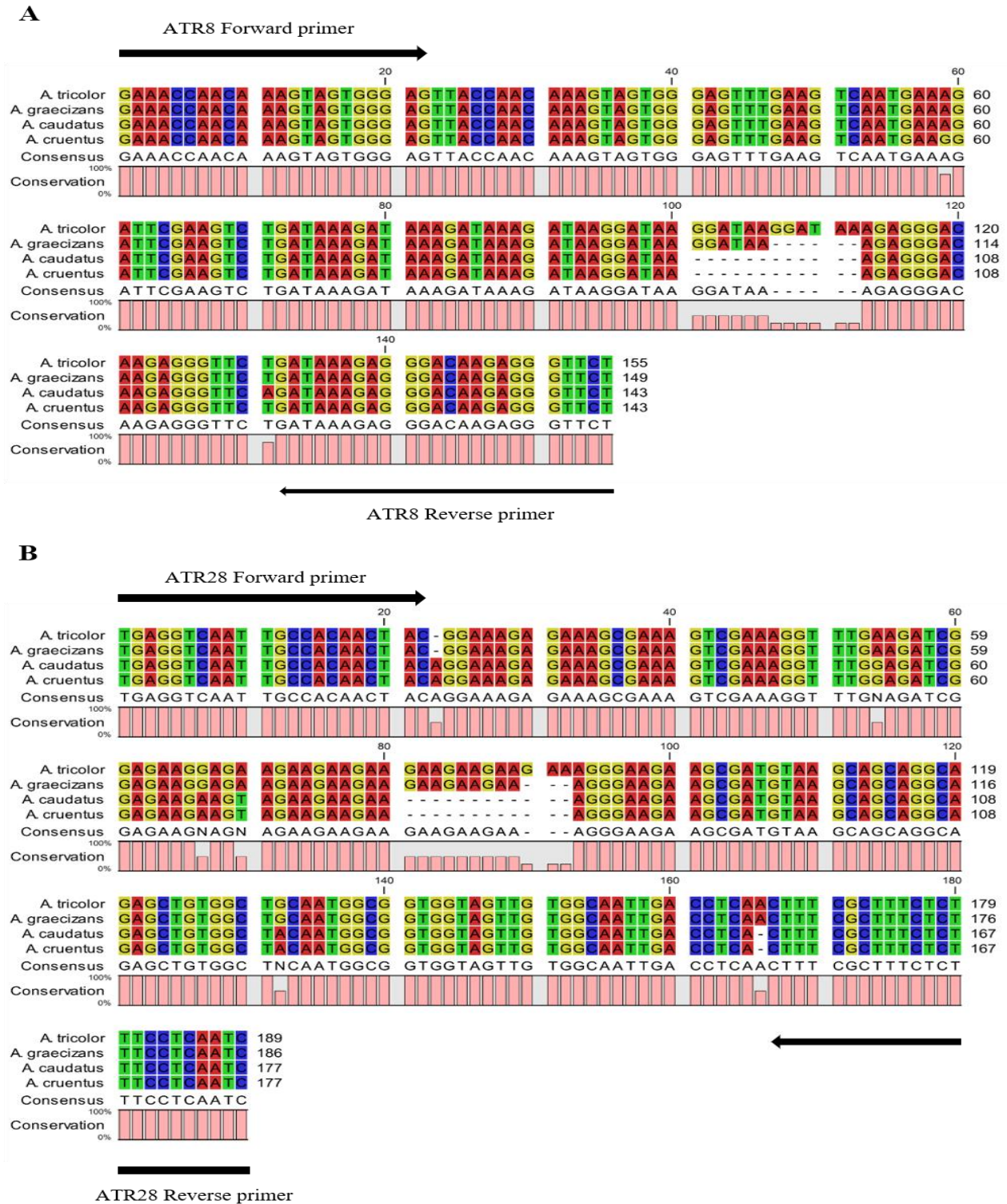
- Rastogi A, Shukla S (2013) Amaranth: a new millennium crop of nutraceutical values. *Critical Reviews in Food Science and Nutrition* 53:109-125
- Ray T, Roy SC (2009) Genetic diversity of *Amaranthus* species from the Indo-Gangetic plains revealed by RAPD analysis leading to the development of ecotype-specific SCAR marker. *Journal of Heredity* 100:338-347
- Rozen S, Skaletsky H (1999) Primer3 on the WWW for general users and for biologist programmers. In: *Bioinformatics Methods and Protocols*. Humana Press, Totowa, NJ. pp. 365-386
- Santana QC, Coetzee MP, Steenkamp ET, Mlonyeni OX, Hammond GN, Wingfield MJ, Wingfield BD (2009) Microsatellite discovery by deep sequencing of enriched genomic libraries. *BioTechniques* 46:217
- Scheible M, Loreille O, Just R, Irwin J (2011) Short tandem repeat sequencing on the 454 platform. *Forensic Science International: Genetics Supplement Series* 3:e357-e358
- Stetter MG, Müller T, Schmid K (2017) 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
- Suez M, Behdenna A, Brouillet S, Graça P, Higuët D, Achaz G (2016) MicNeSs: genotyping microsatellite loci from a collection of (NGS) reads. *Molecular Ecology Resources* 16:524-533
- Sun M, Chen H, Leung F (1999) Low-Cot DNA sequences for fingerprinting analysis of germplasm diversity and relationships in *Amaranthus*. *Theoretical and Applied Genetics* 99:464-472
- Sunil M *et al.* (2014) The draft genome and transcriptome of *Amaranthus hypochondriacus*: A C4 dicot producing high-lysine edible pseudo-cereal. *DNA Research*. 21:585-602
- Thai BT, Tan MH, Lee YP, Gan HM, Tran TT, Austin CM (2016) Characterisation of 12 microsatellite loci in the Vietnamese commercial clam *Lutraria rhynchaena* Jonas 1844 (Heterodonta: Bivalvia: Mactridae) through next-generation sequencing. *Molecular Biology Reports* 43:391-396
- Trucco F, Tranel PJ (2011) *Amaranthus*. In: C. Kole (ed) *Wild Crop Relatives: Genomic and Breeding Resources: Vegetables*. Springer, New York pp. 11-21
- Utomo HS, Wenefrida I, Linscombe SD (2012) Progression of DNA marker and the next generation of crop development. In: *Crop Plant InTech* Open Access Publisher

- Vartia S *et al.* (2016) A novel method of microsatellite genotyping-by-sequencing using individual combinatorial barcoding. *Royal Society Open Science* 3:150565
- Wöhrmann T, Huettel B, Wagner N, Weising K (2016) Microsatellites from *Fosterella christophii* (Bromeliaceae) by *de novo* transcriptome sequencing on the Pacific Biosciences RS platform. *Applications in Plant Sciences* 4
- 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 Evolution* 21:372-387
- You FM *et al.* (2008) BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics* 9:253
- Zalapa JE *et al.* (2012) Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. *American Journal of Botany* 99:193-208
- Zhan L (2016) Inferring ecological population structure and environmental associations through automated analysis of repeat-containing and polymorphic DNA sequences.
- Zhou Z *et al.* (2015) Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature Biotechnology* 33:408-414

Supplementary Data

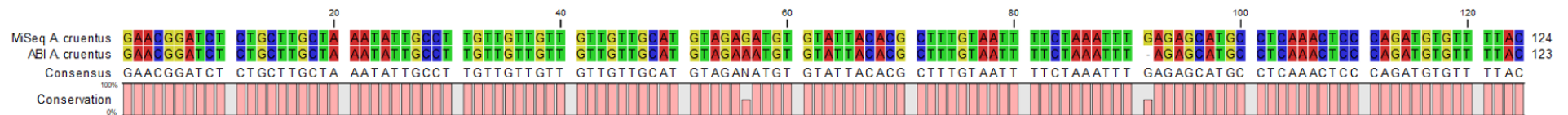
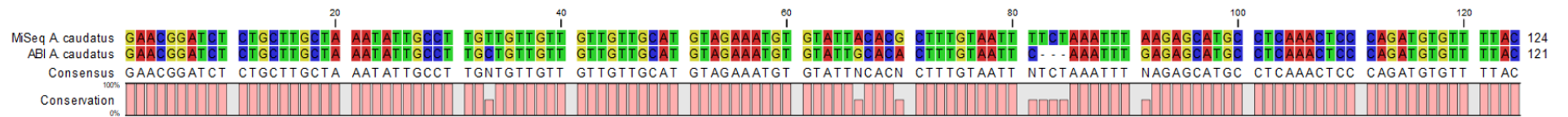
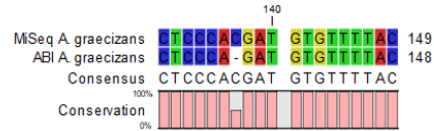
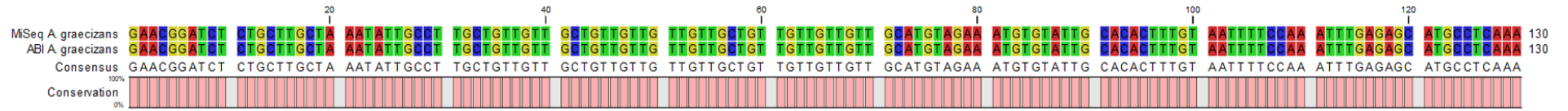
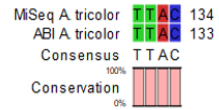
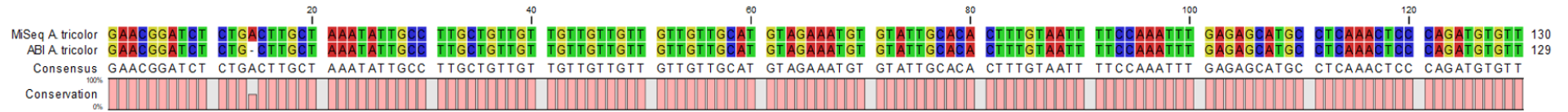


Supplementary Figure S4.1: Illumina MiSeq library preparation for amaranth accessions. (1) Illumina overhang adapter sequences were joined to the six locus specific primers and a multiplexed PCR amplification performed to amplify allelic regions from the genomic DNA of individual amaranth accessions (2) and (3) A subsequent amplification step is performed to add Illumina indices and the P5/P7 sequences (for binding the MiSeq flowcell) for each individual amaranth accession. (4) Equimolar quantities of the accession libraries were pooled and (5) sequenced on the Illumina MiSeq Platform.

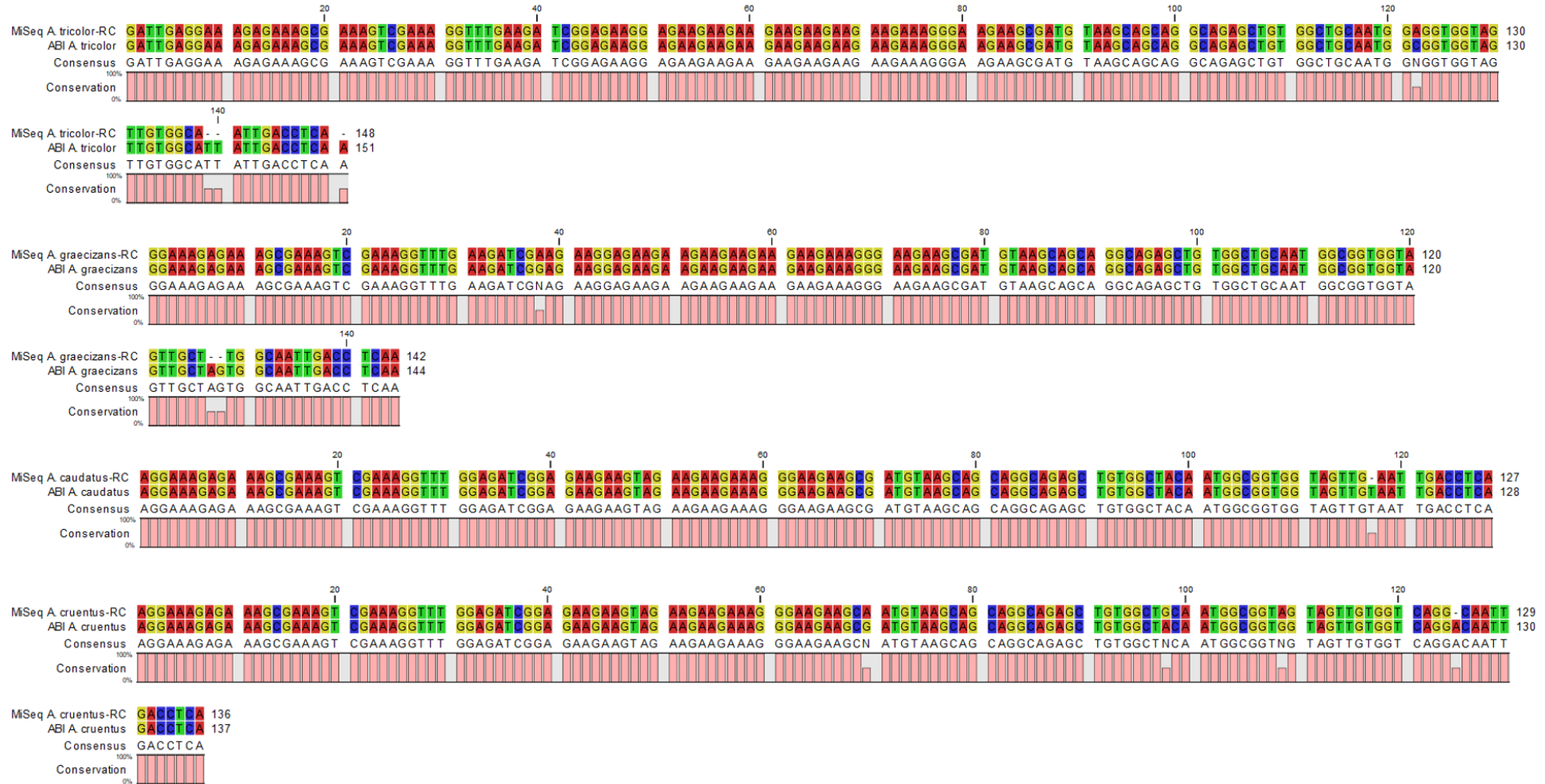


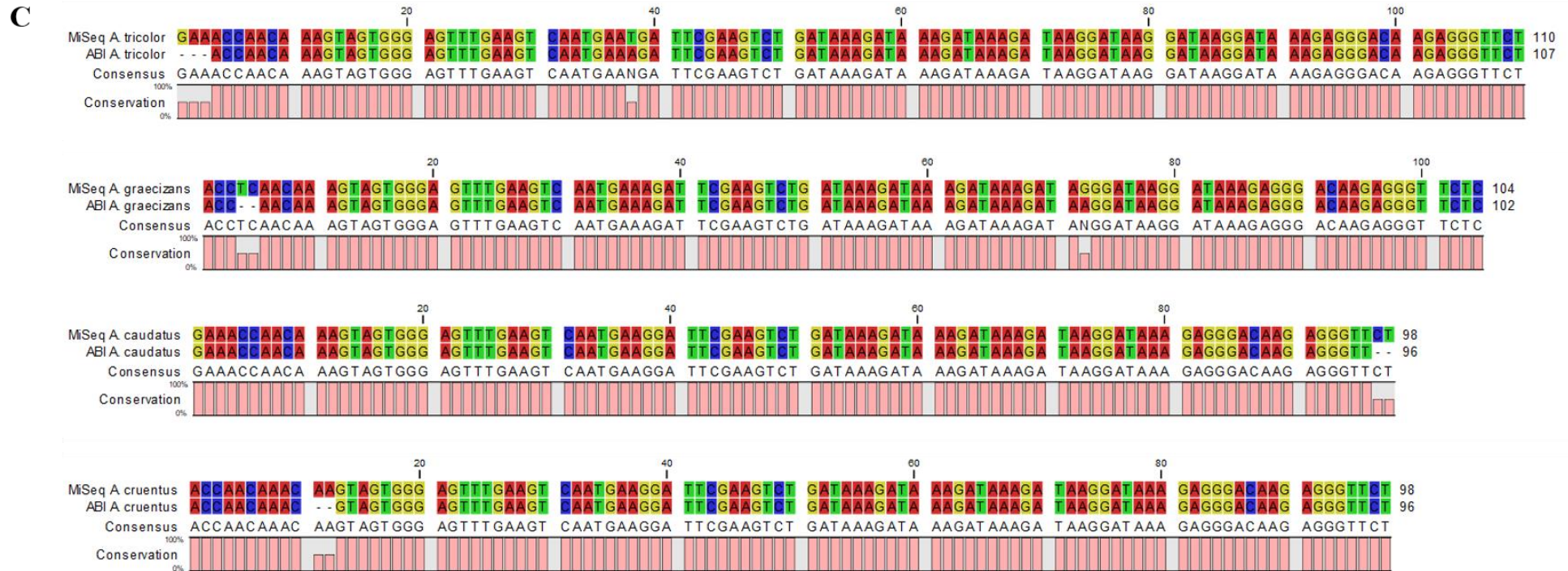
Supplementary Figure S4.2: Sanger sequencing of SSR markers. (A) ATR8 and (B) ATR28 in *A. graecizans* (GRIN15), *A. tricolor* (SAG29), *A. caudatus* (GRIN3) and *A. cruentus* (GRIN8), which indicates the inter-species polymorphic nature of the marker. Forward and reverse primer sequences are indicated on the sequence.

A



B





Supplementary Figure S4.3: Sequence alignments of Sanger sequenced ABI fragments and MiSeq sequenced fragments, for (A) ATR8, (B) ATR28 and (C) ATR32. In most cases, deletions of one to three bases occur during Sanger sequencing of the fragments. The location of these deletions could be ascribed to the low accuracy of the Taq polymerase at the beginning of a sequencing reaction, directly after the primer. MiSeq sequencing offers more accurate sequencing due to the sequencing-by-synthesis technology, and the high coverage of each fragment.

Supplementary Table S4.1: *Amaranthus tricolor* genome assembly statistics

Amount of raw read data	1.374 Gb
Amount of data after FastQC filtering	569.5 Mb
Number of reads used in assembly	234,727,730
Estimated genome size	402 Mb
GC Content	37.30%
Number of contigs	68 656
N50 Contig size	6 kb
Number of scaffolds	13 606
N50 scaffold size	51Kb

Supplementary Table S4.2: SSR Discovery on *A. tricolor* scaffolds using the MSATCOMMANDER 1.0.8 software package.

Scaffold Number	Scaffold Size (bp)	Dinucleotides	Trinucleotides	Tetranucleotides	Pentanucleotides	Hexanucleotides
0	279,404	4	12	9	2	2
8	225,737	6	9	4	2	3
24	201,813	1	8	3	4	4
45	169,873	5	5	6	1	1
62	164,531	0	8	6	3	1
91	214,804	13	12	7	3	3
143	134,631	6	8	6	2	1
159	128,981	2	7	2	3	0
190	127,823	3	7	3	1	1
229	120,405	2	7	5	2	0
235	188,774	2	8	1	1	1
288	111,565	4	4	2	1	2
372	99,585	4	3	5	4	0
415	98,681	0	2	2	0	7
492	94,408	1	3	2	0	0
498	97,066	1	6	3	0	0
510	90,709	3	3	2	3	0
544	88,247	1	3	4	1	0
733	82,638	2	3	2	2	1
832	74,203	0	7	4	1	1
839	73,700	2	4	0	0	0
927	71,600	5	4	7	0	0
975	70,485	0	0	2	2	0
1,110	65,443	3	1	2	1	0
1,214	61,432	3	1	0	0	0
1,427	57,103	0	1	2	1	0
1514	55,711	1	7	1	1	0
2,074	46,363	1	6	2	0	0
3,183	189,084	16	15	18	4	2
3,404	28,957	1	1	2	0	0
3,584	27,222	0	1	0	2	0
3,627	28,921	0	1	1	0	1
3,780	205,347	24	15	9	3	2
4,759	59,003	2	10	5	0	0
6,443	57,098	1	6	3	0	0
12,637	107,269	0	10	4	3	2
13,028	59,370	1	8	5	1	2
18,989	50,321	2	1	3	1	0
19,048	242,352	15	17	20	6	3
25,311	271,728	11	15	4	3	2
Total	4,622,387	148	249	168	64	42

Supplementary Table S4.3: Estimated cost analysis of marker screening using ABI vs NGS technology

ABI Technology

Synthesis of fluorescently labelled primers (6 primers)	R4 250
PCR amplification	
Screening 2 panels (3 markers each) on 60 individuals	R7 000
Total	R11 250
Timeline: 6 injections on ABI (40 min per injection)	4 hours

NGS Technology

Non-fluorescently labelled primers	R1 200
PCR amplification	
Library preps (1 prep, using 60 barcodes)	R1 700
MiSeq run (1 run, 6Gb), approximately 100Mb per sample	R40 000
Total	R42 900
Timeline: 1 MiSeq runs	24 hours

CHAPTER 5

APPLICATION OF CHLOROPLAST PHYLOGENOMICS TO RESOLVE SPECIES RELATIONSHIPS WITHIN THE PLANT GENUS *AMARANTHUS*

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 (*matK*, *rbcL*, 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.

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 *atpF-H*, *matK*, *psbK-I*, *rbcL*, *rpoC1*, *rpoB*, *trnH-psbA*, *trnL-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 *matK*, *rbcL*, *trnH-psbA*, the

nuclear ITS and more recently, the *ycf1* 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 (*matK*, *rbcL* 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 5.1). 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 5.1). 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 5.1: *Amaranthus* genus germplasm sets

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

^aIdentified based on preliminary morphological analysis (data not shown)

^bBased on whole chloroplast sequencing - Fig 5

^cBased on *matK* 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 S5.1) 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 S5.2). 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 S5.3) 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 *matK*, *rbcL* and nuclear ITS) as previously suggested (Dong *et al.* 2015). The *matK* gene was amplified using primers *matK*-F (5'-CACTATGTATCATTTGATAACCCTC-3') and *matK*-R (5'-TATTACAATCAACATTTTCAGAATAG-3') (Burgess *et al.* 2011). Primers *rbcL*-F (5'-ATGTCACCACAAACAGAGACTAAAGC-3') and *rbcL*-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 MgCl₂) (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 *matK*, *rbcL* and ITS data sets. Sequences obtained for the three barcoding genes of the 45 GRIN *Amaranthus* accessions (Table 5.1) have been deposited at GenBank as follows: chloroplast *matK* (KX079543-KX079587); chloroplast *rbcL* (KX079588-KX079632); and nuclear ITS (KX079498-KX079542). Sequence data for *matK*, *rbcL* 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 (*matK*) and 606 (*rbcL*) 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 v1.31 was 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 S5.4.

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 5.2). Sequences from these individual regions were subsequently concatenated to the original barcoding data set (*matK*, *rbcL* 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 5.2: Additional barcoding region statistics

Barcoding region	Size (bp)	Chloroplast region	Informative sites (%)	TOPD Results	
				Disagreement (%)	Split distance
<i>ndhD</i>	1,503	SSC	1.9	43	0.58
<i>rpoC2</i>	2,584	LSC	2.9	50	0.72
<i>atpE</i>	409	LSC	1.4	52	0.65
<i>rpl22</i>	592	LSC	5.6	52	0.65
<i>accD</i>	1,659	LSC	1.6	54	0.72
<i>atpF</i>	503	LSC	1.8	54	0.72
<i>psbA</i>	1,068	LSC	1.2	54	0.70
<i>rps14</i>	311	LSC	3.5	56	0.62
<i>rpoB</i>	3,249	LSC	1.2	57	0.67
<i>ndhG</i>	534	SSC	1.9	58	0.69
<i>rps3</i>	656	LSC	2.13	58	0.58
<i>psbI-trnS</i> IGR	226	LSC	27.2	59	0.74
<i>trnQ</i>	91	LSC	18.6	59	0.76
<i>ndhH</i>	1,204	IR	3.9	60	0.74
<i>petA</i>	1,005	LSC	14	60	0.81
<i>psaC</i>	246	SSC	1.6	60	0.69
<i>atpA</i>	1,524	LSC	1.0	63	0.67
<i>ndhI</i>	514	SSC	1.16	63	0.63
<i>ndhK</i>	861	LSC	2.0	65	0.70
<i>rpl16</i>	401	LSC	2.0	65	0.67
<i>rpoA</i>	1,012	LSC	2.4	65	0.72
<i>rpoB-trnC</i> IGR	1,197	LSC	5.4	65	0.72
<i>trnR-atpA</i>	109	LSC	13.7	65	0.69
<i>trnT-psbD</i> IGR	1,326	LSC	5.8	65	0.72
<i>ycf1</i>	1,407	IR	0.6	65	0.72
<i>atpB</i>	1,495	LSC	1.3	67	0.74
<i>cemA</i>	702	LSC	1.13	67	0.65
<i>petB</i>	648	LSC	1.0	67	0.65
<i>petD</i>	544	LSC	7.0	67	0.72
<i>psbM</i>	103	LSC	7.7	67	0.74
<i>rps2</i>	708	LSC	1.0	67	0.74
<i>trnS</i>	89	LSC	11.2	67	0.72
<i>ccsA</i>	972	SSC	2.9	69	0.74
<i>rps8</i>	404	LSC	1.5	69	0.72
<i>atpH-atpI</i> IGR	679	LSC	4.8	70	0.69
<i>trnG</i>	83	LSC	18.0	70	0.72
<i>trnN</i>	73	IR	21.9	70	0.72
<i>psbD</i>	1,092	LSC	15.0	71	0.81
<i>trnS-rps4</i>	340	LSC	12.35	71	0.72
<i>petG-trnW</i> IGR	136	LSC	13.2	74	0.72
<i>psbC</i>	1,476	LSC	11.9	74	0.88
<i>psbI</i>	156	LSC	17.0	74	0.74
<i>rps16-trnQ</i>	1,555	LSC	3.2	74	0.74
<i>rrn23-rrn4.5</i> IGR	98	IR	10.2	74	0.72
<i>trnC</i>	72	LSC	6.9	74	0.72
<i>ndhF</i>	2,316	IR	4.0	76	0.81
<i>rps18</i>	309	LSC	13.9	78	0.79
<i>rrn16-trnI</i> IGR	302	IR	22.5	78	0.86
<i>rpl16-rps3</i> IGR	1,588	LSC	10.1	85	0.83
<i>rrn16</i>	1,549	IR	18.0	87	1.00
<i>rrn23</i>	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 S5.1. 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 S5.1), resulting in a total *A. tricolor* complete chloroplast genome sequence of 150,027 bp (Figure 5.1).

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 5.3), 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 5.3).

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.

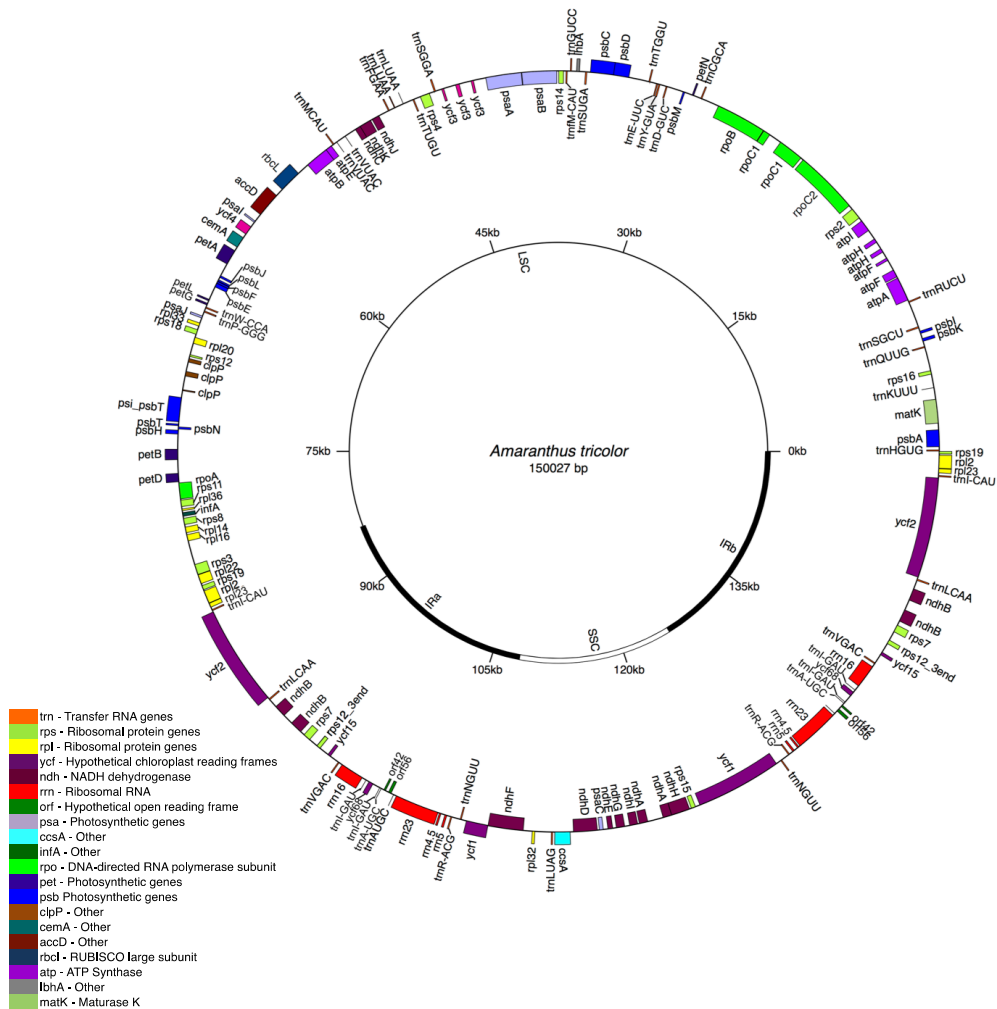


Figure 5.1: Circular gene map for *Amaranthus tricolor* chloroplast genome. Genes indicated on the outer region are transcribed clockwise while genes on the inside are transcribed counter-clockwise. Genes with similar functions are grouped together and colour coded

Table 5.3: Classification of gene regions identified on chloroplast genome

RNA genes

Ribosomal RNA genes

*rrn16** *rrn23** *rrn4.5** *rrn5**

Transfer RNA genes

<i>trnA</i> -UGC*#	<i>trnC</i> -GCA	<i>trnD</i> -GUC	<i>trnE</i> -UUC	<i>trnF</i> -GAA	<i>trnM</i> -CAU	<i>trnG</i> -UCC
<i>trnH</i> -GUG	<i>trnI</i> -CAU*	<i>trnI</i> -GAU*#	<i>trnK</i> -UUU	<i>trnL</i> -CAA*	<i>trnL</i> -UAA#	<i>trnL</i> -UAG
<i>trnM</i> -CAU	<i>trnN</i> -GUU*	<i>trnP</i> -GGG	<i>trnP</i> -UGG	<i>trnQ</i> -UUG	<i>trnR</i> -ACG*	<i>trnR</i> -UCU
<i>trnS</i> -GCU	<i>trnS</i> -GGA	<i>trnS</i> -UGA	<i>trnT</i> -GGU	<i>trnT</i> -UGU	<i>trnV</i> -GAC*	<i>trnV</i> -UAC#
<i>trnW</i> -CCA	<i>trnY</i> -GUA	<i>trnM</i> -CAU				

Polypeptide genes

Ribosomal protein genes

<i>rpl14</i>	<i>rpl16</i>	<i>rpl2*</i>	<i>rpl20</i>	<i>rpl22</i>	<i>rpl23*</i>	<i>rpl32</i>
<i>rpl33</i>	<i>rpl36</i>					
<i>rps11</i>	<i>rps12α</i>	<i>rps12_3end*</i>	<i>rps14</i>	<i>rps15</i>	<i>rps16</i>	<i>rps18</i>
<i>rps19*</i>	<i>rps2</i>	<i>rps3</i>	<i>rps4</i>	<i>rps7*</i>	<i>rps8</i>	
<i>orf42*</i>	<i>orf56*</i>					

Transcription/Translation genes

rpoA *rpoB* *rpoC1#* *rpoC2* *infA*

Photosynthetic genes

<i>rbcL</i>						
<i>psaA</i>	<i>psaB</i>	<i>psaC</i>	<i>psaI</i>	<i>psaJ</i>		
<i>psbA</i>	<i>psbC</i>	<i>psbD</i>	<i>psbE</i>	<i>psbF</i>	<i>psbH</i>	<i>psbI</i>
<i>psbJ</i>	<i>psbK</i>	<i>psbL</i>	<i>psbM</i>	<i>psbN</i>	<i>psbT</i>	<i>psi_psbT</i>
<i>petA</i>	<i>petB</i>	<i>petD</i>	<i>petG</i>	<i>petL</i>	<i>petN</i>	
<i>atpA</i>	<i>atpB</i>	<i>atpE</i>	<i>atpF#</i>	<i>atpH#</i>	<i>atpI</i>	
<i>ycf3§</i>	<i>ycf4</i>					

NAHD dehydrogenase genes

<i>ndhA#</i>	<i>ndhB*#</i>	<i>ndhC</i>	<i>ndhD</i>	<i>ndhE</i>	<i>ndhF</i>	<i>ndhG</i>
<i>ndhH</i>	<i>ndhI</i>	<i>ndhJ</i>	<i>ndhK</i>			

Other protein genes

<i>accD</i>	<i>clpP§</i>	<i>ccsA</i>	<i>matK</i>	<i>cemA</i>	<i>ihbA</i>
-------------	--------------	-------------	-------------	-------------	-------------

Open Reading Frames

<i>ycf1Ψ</i>	<i>ycf15*Ψ</i>	<i>ycf2*</i>	<i>ycf68*</i>
--------------	----------------	--------------	---------------

*	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 5.2) 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.

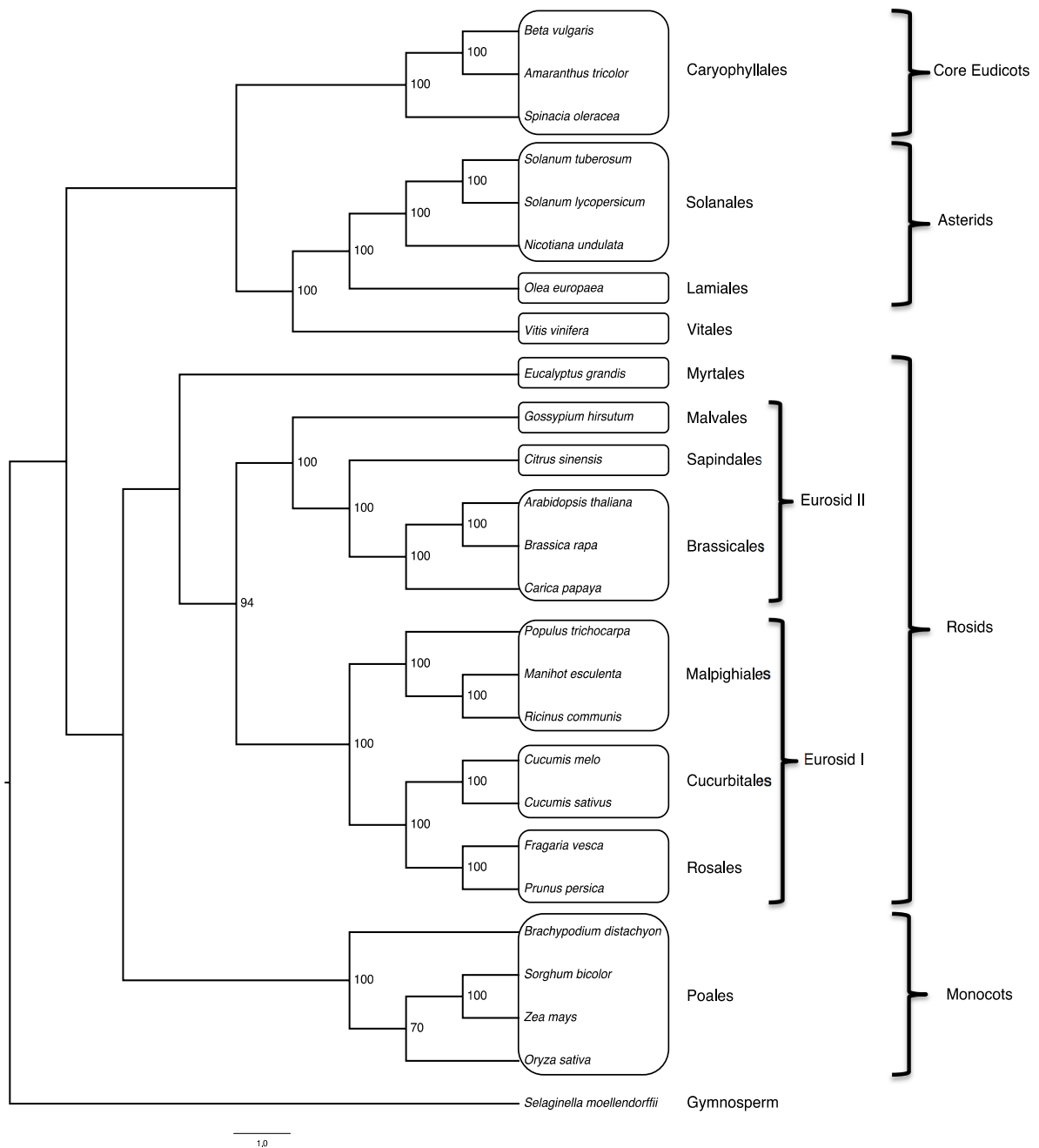


Figure 5.2: Classification of species within the angiosperm plant group based on maximum likelihood phylogeny (Model GTR, 1000 bootstrap replicates) of whole chloroplast genomes. Bootstrap confidence values (> 60%) are indicated at branch nodes. The addition of the newly assembled *A. tricolor* chloroplast 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 *matK*, chloroplast *rbcL* 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 5.4). 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 *matK*) and HKY (*rbcL*) evolutionary models (Table 5.4).

Table 5.4: Barcoding nucleotide site analysis

	Total length (bp, no gaps)	Constant sites (bp)	Variable sites (bp)	Parsimoniously informative sites (bp, %)	JModelTest
<i>MatK</i>	837	723	96	18	2.5%
<i>rbcL</i>	625	592	26	7	1.1%
ITS	702	434	231	37	5.3%
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 5.3). 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 5.3).

Although only about one third of the total nodes were supported with a BS > 60%, certain clades and sub-clades could be identified (Figure 5.3). *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.

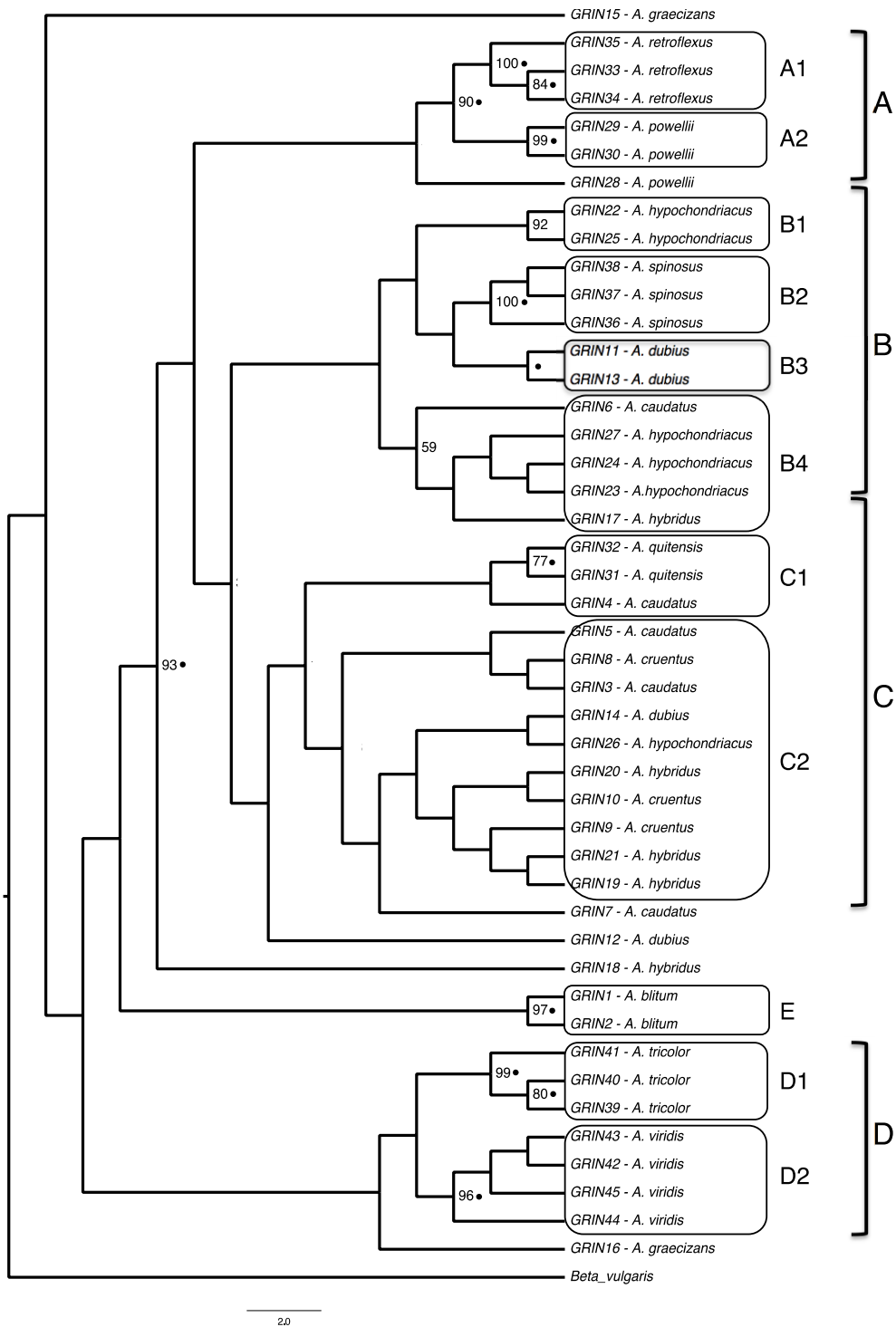


Figure 5.3: *Amaranthus* phylogeny based on DNA barcoding of chloroplast *matK*, *rbcL* and nuclear ITS gene regions. The phylogeny was constructed using a partitioned maximum likelihood analysis (Model GTR+G (ITS, *matK*) and HKY (*rbcL*), 1000 bootstrap replicates), and bootstrap confidence values (> 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 5.1)

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 5.3).

A Bayesian analysis of the partitioned barcoding analysis is provided in Supplementary Figure S5.2. 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.

***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 5.5. Both partitioned and unpartitioned analysis resulted in highly similar tree topologies (data not shown).

Table 5.5: 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 5.4). 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 5.4). 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 5.3). 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 5.3) 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 5.3) 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 5.4). 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 5.4). 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*).

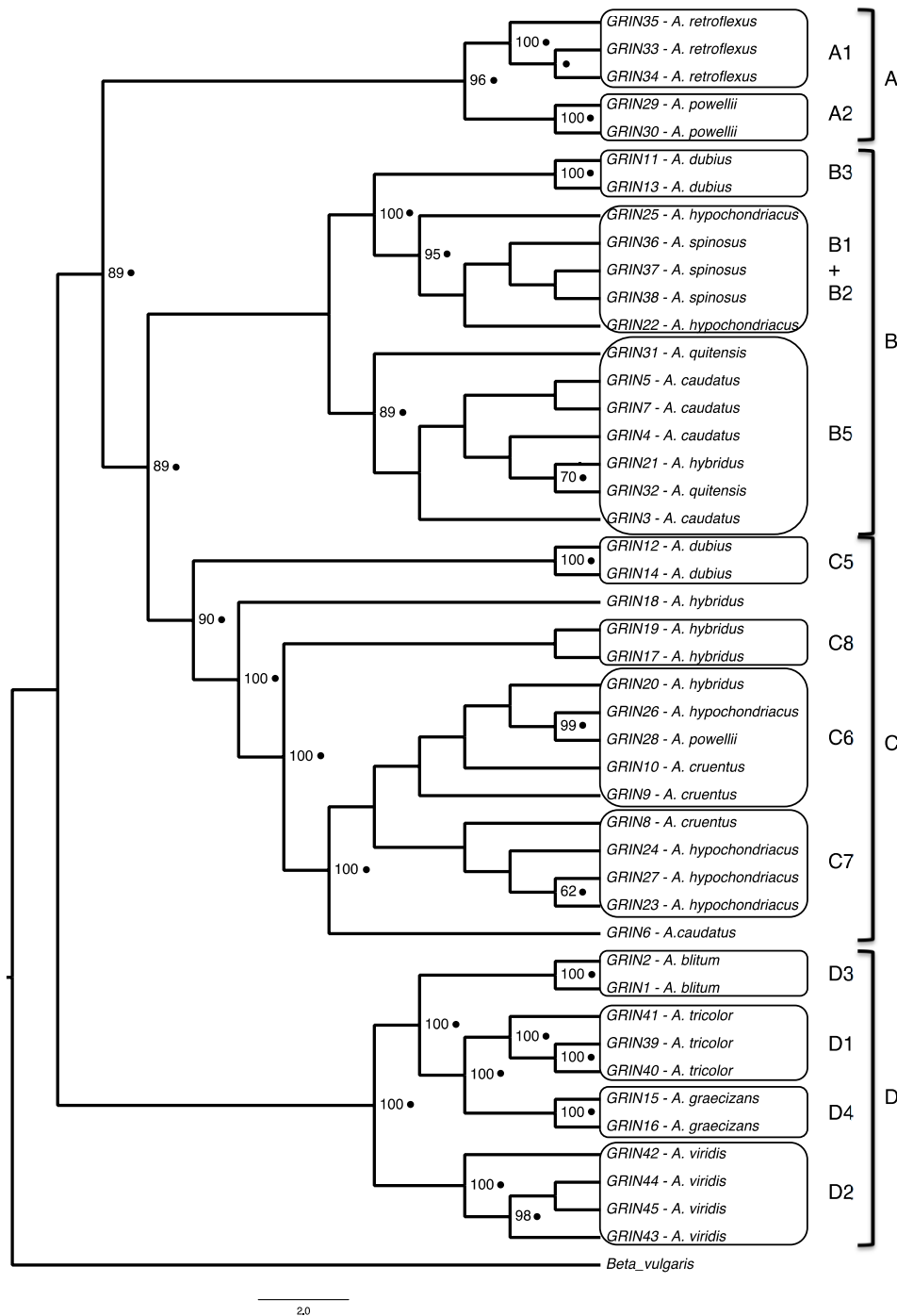


Figure 5.4: Phylogeny of taxonomically described *Amaranthus* accessions. The analysis was based on a partitioned maximum likelihood analysis of 45 GRIN amaranth accessions. (Model GTR+G, 1000 bootstrap replicates) of whole chloroplast genome sequences. Bootstrap confidence values (> 60%) are indicated 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 5.3), 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 5.4). 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 5.3) 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 S5.3). 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 5.6.

Table 5.6: 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 5.4) was observed for the combined analysis of GRIN and the unknown SAG accessions (Figure 5.5). 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 5.5), and their updated identities have been included in Table 5.1. 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 *matK* 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.

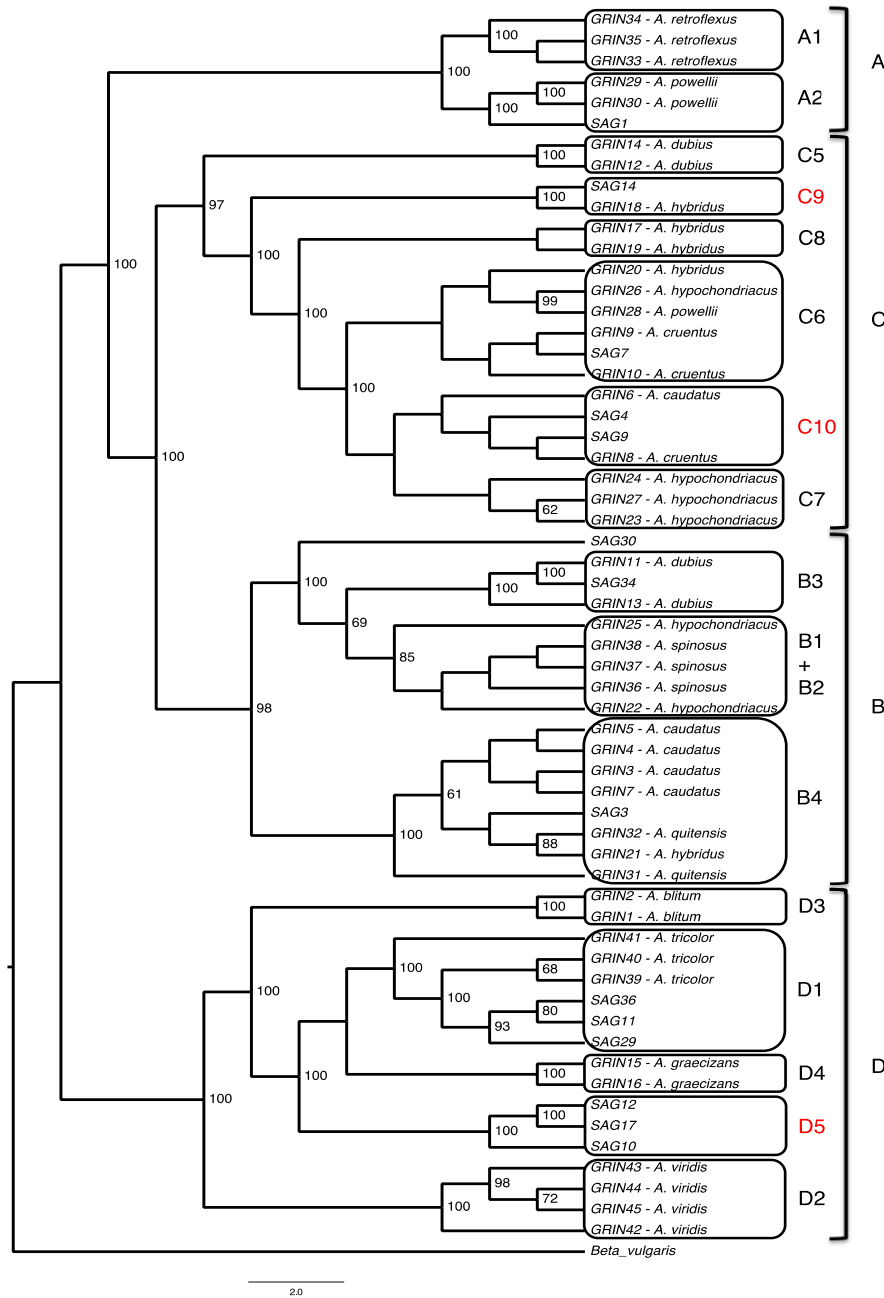


Figure 5.5: Phylogenetic distribution of unknown *Amaranthus* accessions in relation to previously described *Amaranthus* species. The analysis was based on a partitioned maximum likelihood analysis of 45 GRIN and 14 SAG amaranth accessions (Model GTR+G, 1000 bootstrap replicates) of whole chloroplast 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, *matK* and *rbcL* 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, *matK* and *rbcL*) 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 S5.4). 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 *ndhD* (43% disagreement), *rpoC2* (50% disagreement) and *rpl22* (52% disagreement) (Supplementary Figure S5.4). The addition of the *ndhD*, *rpl22* and *atpE* 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 *ndhD* 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, *rpl22* and *atpE* were not considered for further analysis.

The new barcoding phylogeny based on *ndhD* as well as ITS, *matK* and *rbcL* (Supplementary Figure S5.5) 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 *ndhD* 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 *ndhD* 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 *matK*, *rbcL* 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 (*trnA*-UGC, *trnI*-GAU, *trnL*-UAA, *trnV*-UAC, *rpoC1*, *atpF*, *atpH*, *ndhA*, *ndhB*), while two genes contained two introns each (*ycf3*, *clpP*). The remaining genes contained no intronic regions. The intron that was found in *trnK*-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 *trnL-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 *ycf1* and *ycf15* 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 (*ycf1*, *ycf2*, *ycf3*, *ycf4*, *ycf15* and *ycf68*) 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 *ycf1* 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 (*orf42* and *orf56*) 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 reconstruct a broad phylogeny of several different plant orders. The newly assembled chloroplast of *A. tricolor* was placed within the *Caryophyllales* together with *B. vulgaris* 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, Waselkov *et al.*, 2018). Our results were highly congruent to these groupings, with *A. hypochondriacus*, *A. cruentus*, *A. caudatus*, *A. hybridus* and *A. quitensis* falling within the same broad clade, representing the *Amaranthus* sub-genus. The accessions of *A. hypochondriacus* and *A. cruentus* form close sister-clades, with several accessions of *A. hybridus* also scattered in between. This result supports the hypothesis that *A. hybridus* is the possible progenitor of *A. caudatus* and *A. hypochondriacus* by domesticating independently at least twice in two different geographical regions (Sauer, 1967; Waselkov *et al.*, 2018).

As in Waselkov *et al.*, the *A. spinosus* and *A. dubius* accessions also fall within this clade, although they have not previously been considered part of the Hybridus complex (Kietlinski *et al.*, 2014). Interestingly, in the analysis using SSR markers (Chapter 4), the *A. dubius* accessions were found to be more closely related to the members of the Hybridus complex, while in the chloroplast analysis they formed a close sister-branch to accessions of *A. spinosus*. The same result was observed by Waselkov *et al.*, (2018); *A. dubius* grouping with Hybridus complex accessions when nuclear genes were analysed, and with *A. spinosus* when chloroplast genes were analysed. *Amaranthus dubius* is a known allotetraploid species, and previous

literature has suggested that one chromosome set was obtained from *A. spinosus*, while the other was obtained from *A. hybridus/A. quitensis* (Sauer, 1967). The placement of *A. dubius* with *A. spinosus* in the chloroplast analysis possibly indicates that *A. spinosus* was the maternal parent, while the paternal parent was from the Hybridus complex. Waselkov *et al.*, (2018) mention that the Hybridus complex species were also poorly resolved during their study (based on nuclear and chloroplast gene regions) and that the accessions in this group mostly form a polytomy. As in their study, two accessions of *A. hypochondriacus* (not the same accessions as in the current study) group with accessions of *A. spinosus* in the current analysis. In the current analysis, *A. caudatus* groups with *A. quitensis* (bootstrap 89%), indicating that the hypothesis set out in Sauer (1967) that *A. quitensis* is the progenitor of *A. caudatus* could be true. The same topology was found in Waselkov *et al.*, (2018) who hypothesize that possible gene flow between these two accessions could explain their genetic similarity. 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*), which was also found during the analysis of targeted chloroplast sequences (Waselkov *et al.*, 2018). 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 *ndhD* 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 *ndhD* gene had the same level of informativeness (1.9% parsimoniously informative sites) as the previously used chloroplast *matK* (1.9%) and *rbcL* (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 *matK*, *rbcL* 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 5.4), 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 5.4). 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. Species with multiple accessions were recovered as monophyletic, except *A. hybridus*, which was distributed across several different clades in the *Amaranthus* sub-genus. 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. However, whole chloroplast sequence analysis facilitated the more confident identification of unknown *Amaranthus* accessions in the South African genebank (Figure 5.5, Table 5.1), compared to traditional barcoding techniques (Figure 4.9). Based on the additional identities and relationships that could be derived from this study, the hypothesis that whole chloroplast phylogenomics is a superior technique to use for phylogenetic placement of morphologically difficult species groups compared to traditional barcoding techniques, is accepted.

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. Within species variation is an important point to investigate in future *Amaranthus* species phylogeny studies, especially if a more complete representation of different species from different geographical locations are added. Information on the historical distribution, centers of origin and domestication patterns would be revealed by investigating the within-species variation of the chloroplast sequence. Individual chloroplast sequences (or whole chloroplast genomes) are generally conserved, but enough variation does exist for plant population and evolutionary studies. Examples include rice landrace strain identifications (Tong *et al.*, 2015) and wild rice species (Asaf *et al.*, 2017), the use of chloroplast SSRs to delimit species of *Triticum* and *Aegilops* (Yamane and Kawahara, 2017), identification of micro-species in the Asteraceae family (Salih *et al.*, 2017). It has recently been shown that the complete chloroplast genome contains as much variation as the COI gene for animals and can be used as barcodes for species identifications (Li *et al.*, 2014).

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. The same whole plastid sequencing approach has been described when performing evolutionary studies in humans. In the past, only the control regions of the mitochondrial genome were used to investigate human evolution, but recent studies are focussing more on the use of the whole mitochondrial genome to provide in-depth information on the maternal lineage of our species (Ingman *et al.*, 2000).

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.

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 JJ, 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*+ *matK* 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 *rbcL*, *atpB*, and *matK* 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) *ycf1*, 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 *rbcL* gene complements the non-coding *trnH-psbA* 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* in 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 S5.1: 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	TCTTTCATTTCCGGTCGGAAC		
CSS7 F	TTTCTCACTTTCACCTCGATTTT	59	717
CSS7 R	TAATGAATCGGAGCACATGG		
CSS8 F	TTTTTCGTTGGCGATATCCT	59	706
CSS8 R	TTCATAAGAAATCGCGTGGT		
CSS9 F	ACCCGCAATGAATAGGGAAG	59	675
CSS9 R	GCCTCAACAATTGGGTTCAAT		
CSS10 F	GATTTTCAATGCGCGATTTT	59	683
CSS10 R	CGTTCGGCTCGATCTATTTT		
CSS11 F	TCCGAAACAAATACGATTACCC	60	676
CSS11 R	ACTAAGGGGGTTGGTGAAA		
CSS12 F	GACCACGATTGGAATTGTT	59	705
CSS12 R	ACATCCCGGATCTGATGAAA		
CSS13 F	CCTCATACGGCTCCTCAAGA	60	736
CSS13 R	CGGGTTGAAACGAATTATGG		
CSS14 F	CCGCACTTTTTGGGGTACTA	59	721
CSS14 R	TGGCTAATCGAACAAGGACA		
CSS15 F	CCACTCAGCAGTTTGAATA	60	1100
CSS15 R	CGGGTTTGAAAAATCCCTTC		
CSS16 F	TCGGCCCTTTATTCGAGTA	54	1355
CSS16 R	GCCGTGGGAATCAAAAAA		
CSS17 F	GCGTGTAACGAGGTGCTCT	60	689
CSS17 R	AAAAGACCATGCCGCTACTG		
CSS18 F	GAGGAGGAAGCTTATAAAGA	53	1298
CSS18 R	GGCACAAAAATAGAGGGA		
IRa/SSC F	CTTCTCCTTCCCTTTTTTTC	52	1329
IRa/SSC R	GATGCGACCTTTTCTTAC		
SSC/IRb F	GGCTTCCTTTGTTCGGTT	55	1269
SSC/IRb R	CCTTCTCCTTCCCTTTTTTT		
IRb/LSC F	GGACAAGTGGGAAATGTTGG	60	707
IRb/LSC R	GCAATGCCGTTTCTTGTGTT		
LSC/IRa	GCATAACACGGAACAAAG	54	1027
LSC/IRa	AGAAAAAACCCACAACC		

Supplementary Table S5.2: 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 S5.3: Whole angiosperm chloroplast sequences obtained from Genbank (NCBI)

Angiosperm species	Common name	NCBI Accession Nr	Size (bp)	GC content (%)	Family
<i>Arabidopsis thaliana</i>	Thale cress	gi 5881673	154,478	36.3	Brassicaceae
<i>Amaranthus tricolor</i>	Leafy amaranth	PRJNA318736	150,027	36.6	Amaranthaceae
<i>Beta vulgaris</i>	Sugar Beet	gi 148607972	149,696	37.0	Amaranthaceae
<i>Brachypodium distachyon</i>	Purple False Brome	gi 194033128	135,199	38.6	Poaceae
<i>Brassica rapa</i>	Turnip	gi 323481965	153,482	36.4	Brassicaceae
<i>Carica papaya</i>	Pawpaw	gi 166344111	160,100	36.9	Caricaceae
<i>Citrus sinensis</i>	Orange	gi 113952601	160,129	38.5	Rutaceae
<i>Cucumis melo</i>	Melon	gi 346578170	156,047	36.9	Cucurbitaceae
<i>Cucumis sativus</i>	Cucumber	gi 67511377	155,293	37.1	Cucurbitaceae
<i>Eucalyptus grandis</i>	Rose gum	gi 309322431	160,137	37.9	Myrtaceae
<i>Fragaria vesca</i>	Strawberry	gi 325126844	155,691	37.2	Rosaceae
<i>Gossypium hirsutum</i>	Cotton	gi 329317136	160,256	37.2	Malvaceae
<i>Manihot esculenta</i>	Cassava	gi 169794052	161,453	35.9	Euphorbiaceae
<i>Nicotiana undulata</i>	Tobacco	gi 347453879	155,863	37.9	Solanaceae
<i>Olea europaea</i>	Olive	gi 334084552	155,657	37.8	Oleaceae
<i>Oryza sativa</i>	Rice	gi 109156581	134,496	39.0	Poaceae
<i>Populus trichocarpa</i>	Poplar	gi 133712039	157,033	36.7	Salicaceae
<i>Prunus persica</i>	Peach	gi 313183801	157,790	36.8	Rosaceae
<i>Ricinus communis</i>	Caster oil	gi 372450118	163,161	35.7	Euphorbiaceae
<i>Solanum lycopersicum</i>	Tomato	gi 113531108	155,461	37.9	Solanaceae
<i>Solanum tuberosum</i>	Potato	gi 82754608	155,312	37.9	Solanaceae
<i>Sorghum bicolor</i>	Sorghum	gi 118614470	140,754	38.5	Poaceae
<i>Spinacia oleracea</i>	Spinach	gi 7636084	150,725	36.8	Amaranthaceae
<i>Vitis vinifera</i>	Grape Vine	gi 91983971	160,928	37.4	Vitaceae
<i>Zea mays</i>	Maize	gi 11994090	140,384	38.5	Poaceae
Gymnosperm species	Common name	NCBI Accession Nr	Size (bp)	GC content (%)	Family
<i>Selaginella moellendorffii</i>	Lycophyte	gi 296399203	143,775	51.0	Selaginellaceae

Supplementary Table S5.4: Genic information for partitioned phylogenomic analysis

Chloroplast Region	Strand	Fragment size	Pi* Sites %	Model
<i>accD</i>	+	1,658	1.9	TVM+I
<i>atpA</i>	-	1,523	1.4	TPM1uf+G
<i>atpB</i>	-	1,492	0.5	GTR
<i>atpE</i>	-	405	1.5	TPM2uf
<i>atpF</i>	-	160	0.0	F81
<i>atpI</i>	-	754	0.9	TVM
<i>ccsA</i>	+	973	3.0	TPM1uf+G
<i>cemA</i>	+	701	1.1	TrN+I
<i>clpP</i> Exon1	-	227	0.9	TPM1uf
<i>clpP</i> Exon2	-	292	1.7	TPM1uf
<i>infA</i>	-	232	0.9	TPM3uf
<i>matK</i>	-	1,515	3.6	TVM
<i>ndhA</i> Exon1	-	539	2.0	TIM1
<i>ndhA</i> Exon2	-	554	0.9	TPM1uf
<i>ndhD</i>	-	1,502	2.3	TVM+G
<i>ndhE</i>	-	309	1.0	TPM2uf
<i>ndhF</i>	-	3,452	3.3	TVM+I+G
<i>ndhG</i>	-	533	1.9	TPM2uf
<i>ndhH</i>	-	1,181	0.8	TVM+I+G
<i>ndhI</i>	-	512	1.2	TPM1uf
<i>ndhJ</i>	-	476	1.1	TIM3
<i>ndhK</i>	-	862	2.2	TVM+G
<i>petA</i>	+	961	0.9	TVM+I
<i>petB</i>	+	648	1.1	TPM1uf+G
<i>petD</i>	+	521	1.0	TPM1uf
<i>petG</i>	+	113	2.7	TrN
<i>psaA</i>	-	2,251	0.6	TVM+G
<i>psaB</i>	-	2,202	0.8	TVM+I
<i>psaC</i>	-	243	1.6	TrN
<i>psbA</i>	-	1,061	1.2	TIM1+I
<i>psbC</i>	+	1,456	0.3	HKY
<i>psbD</i>	+	1,061	0.8	TVM
<i>psbH</i>	+	239	2.1	TPM2uf+I
<i>psbI</i>	+	163	4.3	TPM1uf+G
<i>psbK</i>	+	184	1.6	TPM1uf
<i>psbT</i>	+	107	1.9	TPM1uf+I
<i>psi psbT</i>	+	1,526	1.2	GTR+G
<i>rbcL</i>	+	1,426	0.7	TPM1uf+G
<i>rpl14</i>	-	365	0.5	TIM2
<i>rpl16</i>	-	399	2.3	TVM+I
<i>rpl22</i>	-	591	4.6	TVM+G
<i>rpl32</i>	+	177	6.2	F81+G
<i>rpl36</i>	-	111	1.8	TIM1
<i>rpoA</i>	-	1,012	2.4	TPM1uf
<i>rpoB</i>	-	3,214	1.4	GTR+I
<i>rpoC1</i> Exon1	-	1,610	2.0	TVM+G
<i>rpoC1</i> Exon2	-	447	0.7	TPM1uf
<i>rpoC2</i> Exon1	-	2,439	1.7	TVM+G
<i>rpoC2</i> Exon2	-	1,645	1.0	TIM1+I
<i>rps11</i>	-	457	1.3	TPM1uf
<i>rps14</i>	-	303	1.3	TPM3uf
<i>rps15</i>	-	271	2.2	HKY+I
<i>rps16</i>	-	228	0.9	TVM
<i>rps18</i>	+	411	4.9	TrN
<i>rps19</i>	-	276	1.1	TPM1uf
<i>rps2</i>	-	710	1.0	TPM1uf
<i>rps3</i>	-	655	2.3	TPM2uf+I
<i>rps8</i>	-	402	1.5	TPM1uf+G
<i>vcf1</i>	+	1,406	0.6	TPM1uf
<i>vcf2</i> Exon1	+	2,077	0.2	TPM1uf
<i>vcf2</i> Exon2	+	4,635	0.6	TPM1uf+I
<i>vcf3</i> Exon1	-	152	0.0	TrNef
<i>vcf3</i> Exon2	-	230	0.9	HKY
<i>vcf4</i>	+	555	1.4	TIM3

*Pi: Parsimoniously informative

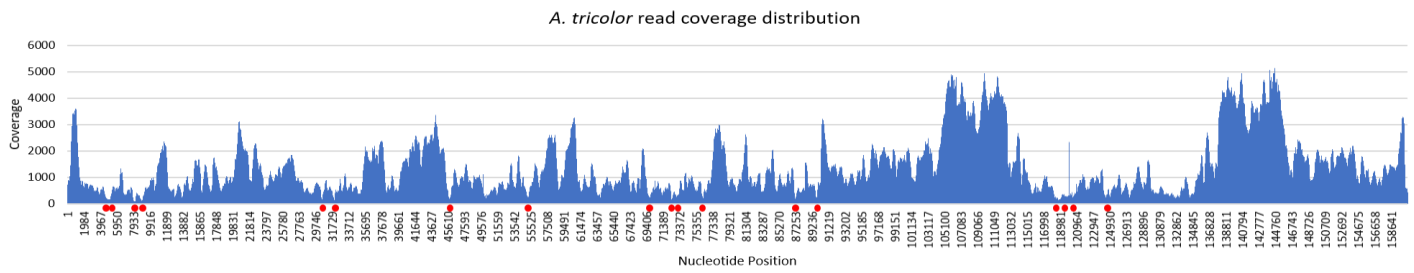


Figure S5.1: Mapping of raw *Amaranthus tricolor* Illumina sequencing reads to the *Beta vulgaris* chloroplast sequence resulted in a mostly uniform coverage distribution. Areas with coverage below 50x were identified (marked by red circles) and were further confirmed/resolved through primer design and Sanger sequencing

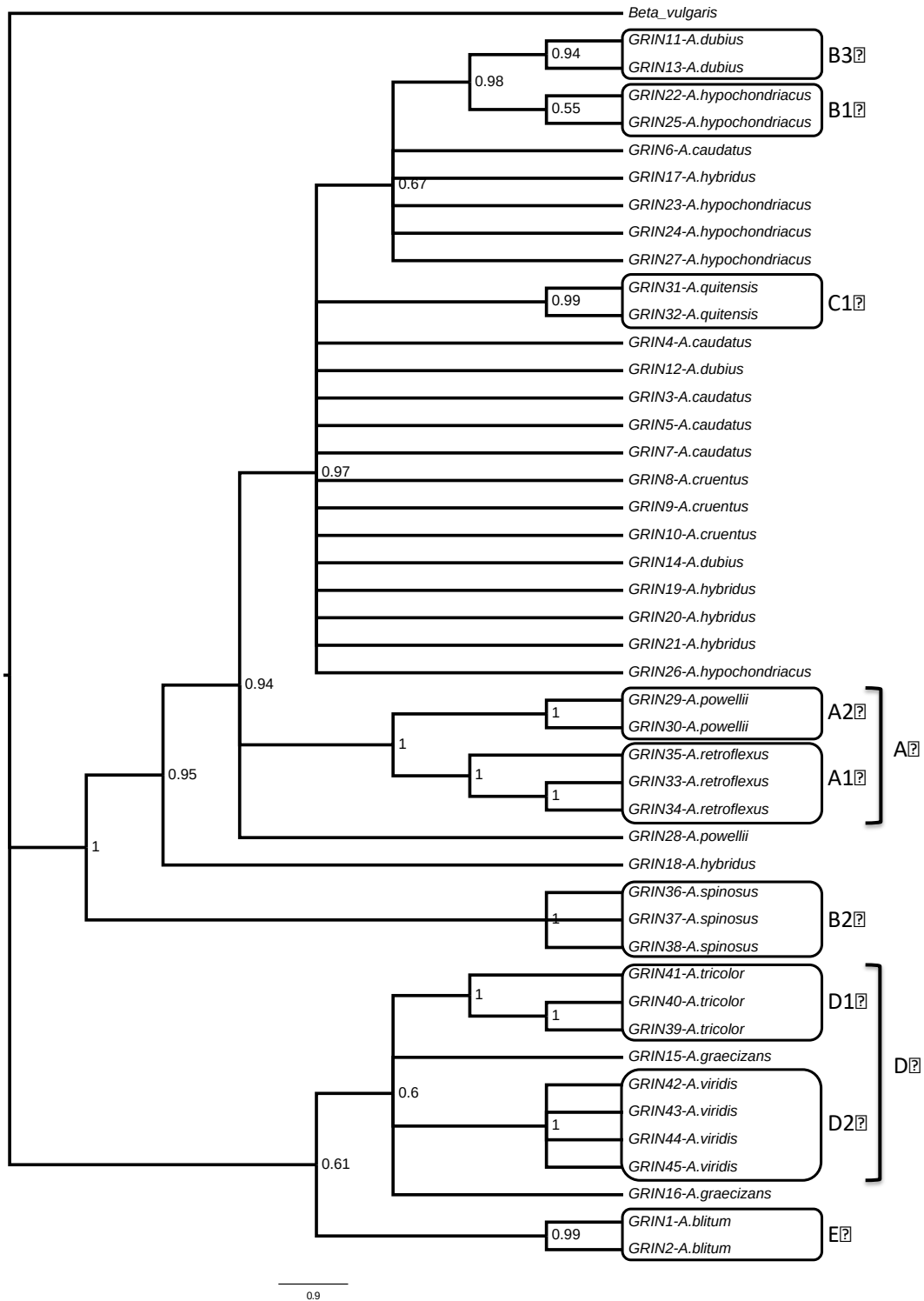


Figure S5.2: Amaranth phylogeny based on DNA barcoding of chloroplast *matK*, *rbcL* and nuclear ITS gene regions using Bayesian analysis. Percentage posterior probability values are indicated on nodes. Only nine sub-clades could be identified compared to the RAxML analysis, but their probability values are higher. The remaining nodes have collapsed, and identifications become questionable

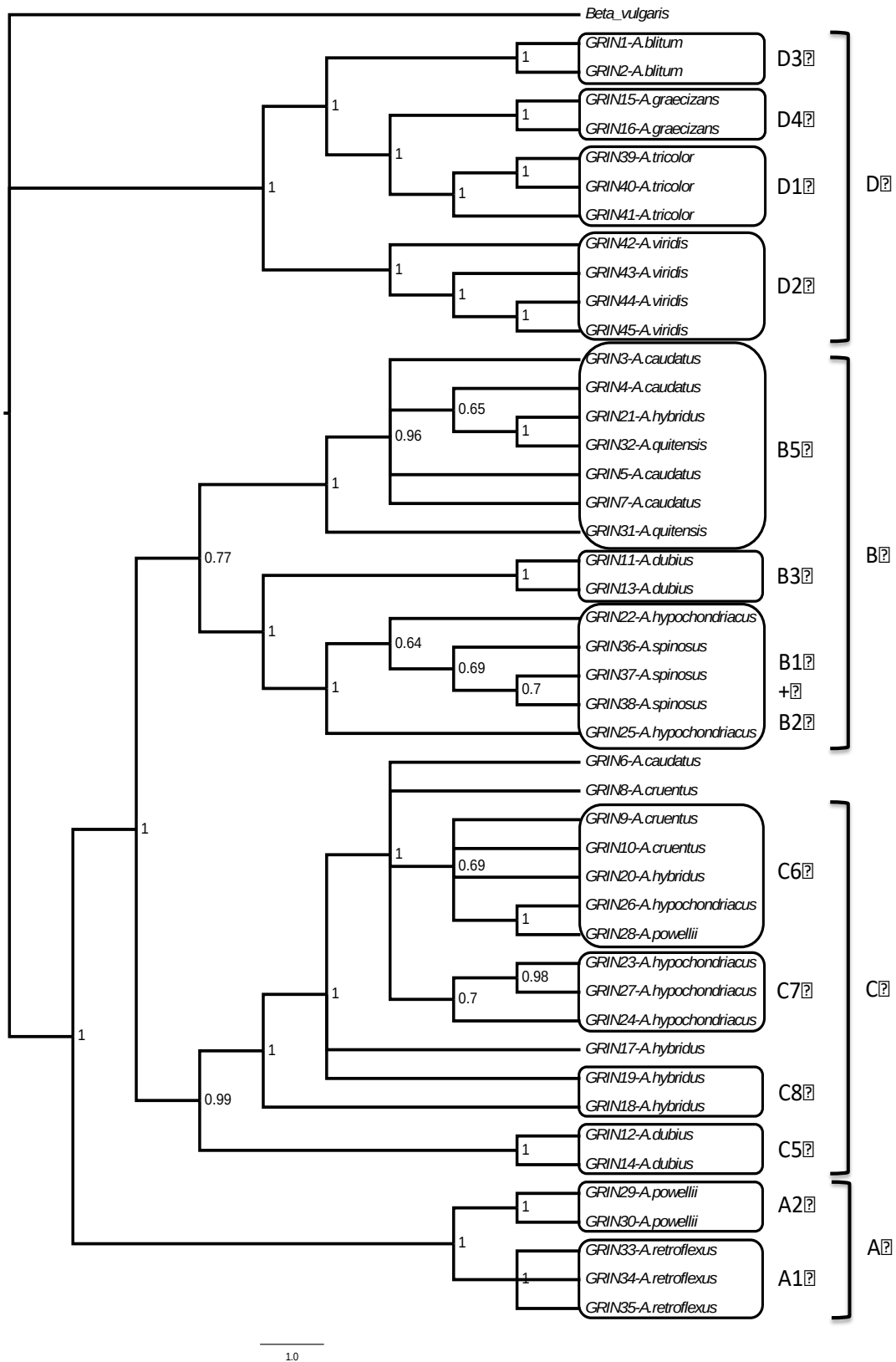


Figure S5.3: Amaranth phylogeny based on a Bayesian analysis of whole chloroplast genome GRIN sequences.

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.

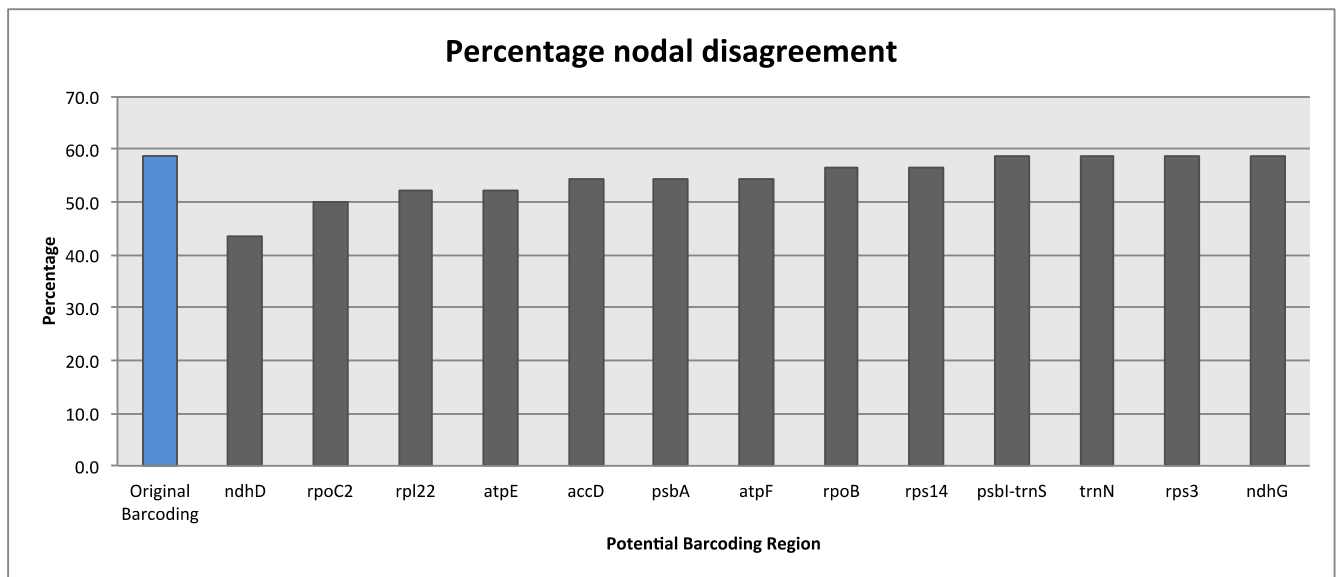


Figure S5.4: Bar graph indicating the percentage disagreement between a new barcoding phylogeny vs. the whole chloroplast phylogeny using the TOPD/FMTS software. The blue bar reflects the percentage disagreement between the original barcoding set (*matK*, *rbcL* and ITS) and the whole chloroplast phylogeny. 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

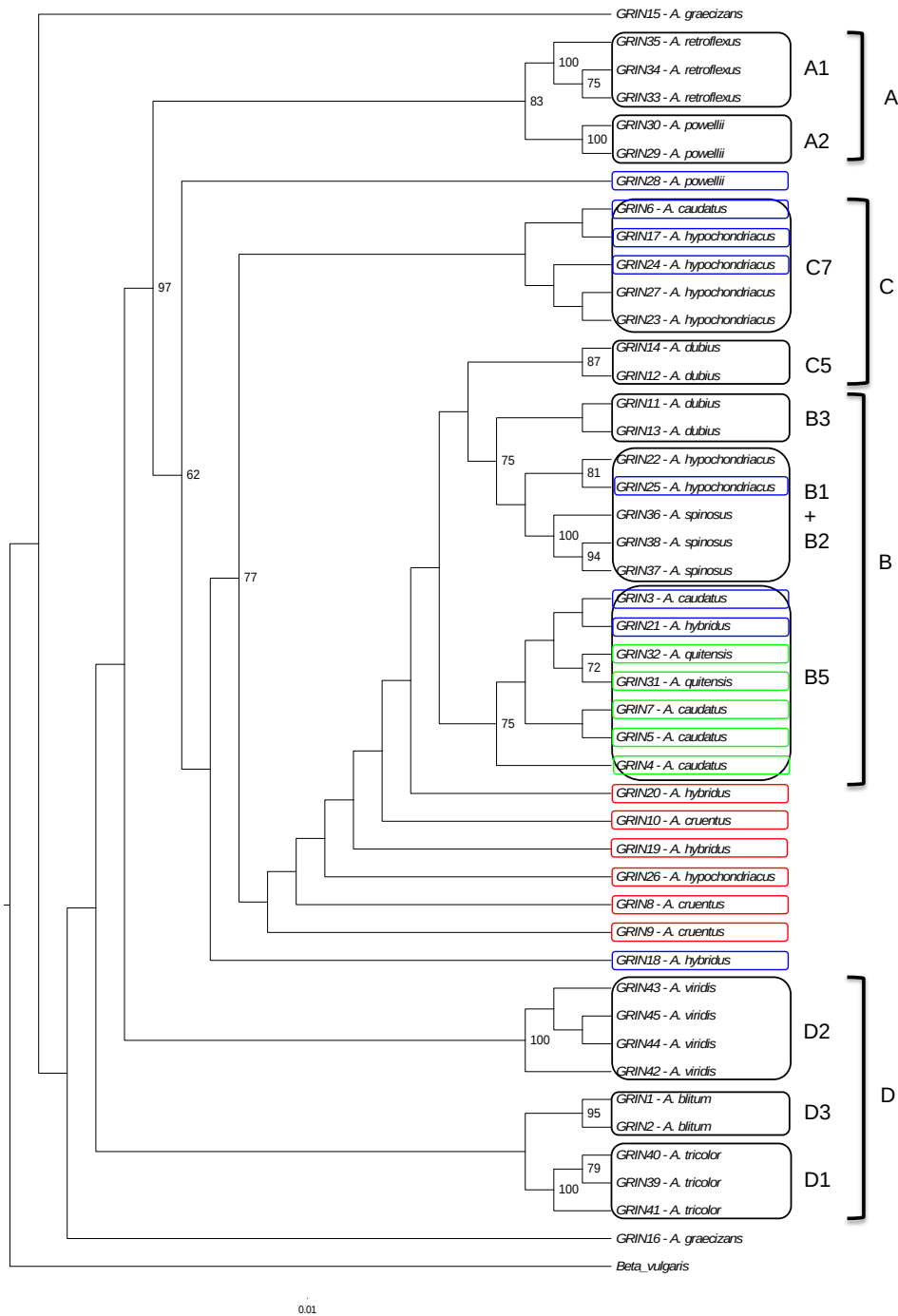


Figure S5.5: *Amaranthus* phylogeny based on DNA barcoding of chloroplast *matK*, *rbcL*, *ndhD* and nuclear ITS gene regions on the entire germplasm set. The phylogeny was constructed using a partitioned maximum likelihood analysis (Model GTR+G (ITS, *matK*, *ndhD*) and HKY (*rbcL*), 1000 bootstrap 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)

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

The observable effects of global warming (temperature changes, extreme weather patterns, rising sea levels) have become more prominent over the last few decades, especially in the agricultural sector (Ziervogel *et al.* 2014). Changes in weather patterns will have an adverse effect on global agriculture, since most plant species will not be able to harness or develop adaptive traits in time to sustain growth during severe weather fluctuations. Coupled with the escalating population size (and subsequent loss of land due to urban development), the world is facing a crisis to sustain food security (Tumushabe, 2018, Ncube *et al.*, 2018). Most countries are dependent on only a few commercialized plant species for their dietary needs, which becomes a problem if weather anomalies influence their growth and development (Mlakar *et al.* 2010). Plants that are native or indigenous to a region naturally contain the necessary genetic diversity to adapt to their environment, and often a more diverse nutrient profile to aid in a more comprehensive diet for humans (Dawson *et al.* 2009; Daru *et al.* 2015).

The genus *Amaranthus* is among an array of traditional vegetables that have been receiving increased attention for future development in addressing the effects of hidden hunger and malnutrition, especially in Africa (Achigan-Dako *et al.* 2014). Within this genus are specific species that can be utilized for their grain, which contains considerable amounts of important amino acids such as lysine and methionine, as well as essential oils. In addition, the leafy amaranth species are excellent sources of protein and vitamins (especially β -carotene) and supplies essential minerals for a balanced diet (Akond *et al.* 2013; Rastogi and Shukla 2013). A few commercialized grain *Amaranthus* species are currently grown in the USA, Central Mexico, India and parts of Africa (Achigan-Dako *et al.* 2014). In particular, *A. tricolor* is considered an important leafy vegetable in Africa, due to its ability to withstand drought conditions over an extended period (Ribeiro *et al.* 2017) and the palatability of the leaves. *Amaranthus tricolor* plants are well adapted to African environmental conditions and are considered low input crops in terms of water application and maintenance. In addition, the leaves contain appreciable amounts of essential minerals (calcium, iron and zinc) and micronutrients (β -carotene) to enrich the diets of resource poor nations (Grubben 1980; van Rensburg *et al.* 2007). A long-term aim for *Amaranthus* breeding includes crossing a grain type species with a leafy type species, such as *A. tricolor*, to incorporate the respective beneficial traits in one genotype without compromising yield.

In the present study, the main hypothesis was that next generation sequencing of the traditional leafy vegetable *A. tricolor* could be used to develop genomic tools that would identify and classify unknown *Amaranthus* genus accessions based on their genetic diversity and evolutionary relationships to previously identified species. This hypothesis was accepted based on the development of a draft, annotated genome of *A. tricolor*, the development of informative inter/intra-species molecular SSR markers and the whole chloroplast assemblies of several *Amaranthus* species for downstream phylogenomic analysis.

The draft genome of *A. tricolor* developed during this study is a first step to the development of a full-length, completely annotated genomic assembly. In total, 135.5 Gb reads were produced from paired end and long mate pair NGS (Illumina) sequencing data, resulting in a total scaffold size of 323 Mb, with 15.72% gapped regions and an N50 size of 51 Kb. Using the transcriptomic data as evidence of gene predictions, the genome was annotated with 16 495 coding regions with a median length of 4 461 bp (Chapter 3, Table 3.1 and Table 3.2). The genes and gene transcripts revealed will be useful to investigate important gene functions (for potential transgenic studies) and gene syntenic studies within the *Amaranthus* genus and other crops species. Genomic rearrangements, which could have an influence on postzygotic barriers leading to hybrid sterility, are important features to study if hybridizations between different species groups are to be attempted. Furthermore, the development of an extensive array of genome wide informative genetic marker systems (such as SNP's or SSR's) becomes highly feasible. Coupled with phenotypic data, fine mapping of beneficial traits and association mapping experiments should simplify the identification of focused genes for genetic manipulation or targeted breeding. The transcriptomic data produced during the current study can also be used for studying gene expression of important trait-related genes when compared to a previously published, cultivated grain amaranth type, *A. hypochondriacus*. Knowledge on how target genes are expressed (and the identification of eSNP's or eQTL's) could shed light on how to further improve current cultivated types of amaranth, potentially bypassing more difficult traditional breeding for enhanced traits.

Malnutrition caused by insufficient amounts of essential micronutrients such as lysine are prevalent in Africa, mainly due to the lack of sufficient levels of lysine in maize (which forms a staple food). The high lysine content within the grains of *A. hypochondriacus* has previously been investigated using a draft genome and transcriptome sequence (Sunil *et al.* 2014). The authors suggest that the absence of orthologues for two lysine sensitive aspartate kinase (AK) genes or the transcriptomic expression levels of

dihydrodipicolinate synthase (DHDPS) may be responsible for the high lysine content observed within *A. hypochondriacus*. With the availability of the genome and transcriptome of *A. tricolor*, these genes (or their expression) becomes testable hypotheses for future research. Targeted breeding could potentially enable researchers to upregulate the lysine-producing genes within *A. tricolor*, making the grains more nutritious in terms of essential amino acids. Coupled with the already beneficial traits of leaf palatability and drought resistance of *A. tricolor*, the addition of pale seed colour and taller plants from *A. hypochondriacus* will yield superior genotypes encompassing most of the consumer driven traits for amaranth plants.

Future work to enhance the current genome and transcriptome assembly from multiple contig/scaffold stage to full-length chromosomal stage would include long read sequencing with PacBio or NanoPore technologies to produce a chromosome-scale assembly with high coverage and low gapped regions. Recently, a full-length genome (composed of 16 scaffolds corresponding to the 16 chromosomes) of *A. hypochondriacus* was released (Lightfoot *et al.* 2017). Syntenic studies will enable researchers to efficiently identify and compare homologous regions during genome wide association studies. Once a higher quality genome sequence is available for *A. tricolor*, shallow NGS sequencing of alternative *Amaranthus* species would facilitate concurrent multiple gene target identifications for downstream breeding of genotypes.

By filtering the genome sequence of *A. tricolor*, it was possible to extract, assemble and annotate a complete chloroplast genome sequence. Since chloroplast sequences are predominantly maternally inherited, they are less susceptible to genomic recombination events and are generally slow mutating organelles (Greiner *et al.* 2015). However, chloroplast genic and intergenic regions will contain large numbers of informative characters for reconstructing phylogenetic relationships between species, which would result in higher clade support compared to targeted gene barcoding. In the current study, the full-length chloroplast sequences were assembled for 59 amaranth accessions, representing 14 different species groups (Chapter 5, Table 5.1).

The main conclusions from the phylogenetic analysis included the highly confident confirmation of species separation in their respective sub-genus groups, as well as re-identification of certain accessions that had previously been assigned an incorrect species name based on morphological characteristics or limited genetic marker analysis (Chapter 5, Fig 5.4 and Fig 5.5). *Amaranthus tricolor* formed a monophyletic group with *A. graecizans*, while *A. blitum* and *A. viridis* formed close sister groups. These species are hypothesized to be

introductions of European and Asian origin and are classified as *Amaranthus* subgenus *Albersia*. The two main weedy amaranths (*A. retroflexus* and *A. powellii*) resided within the same clade based on the phylogenetic analyses, confirming their initial identifications. An exception is the *A. powellii* accession collected in Mexico (GRIN28), which groups with *A. hypochondriacus* accessions (also collected in Mexico) indicating that GRIN28 may have been miss-identified initially. 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.* 2016) that *A. caudatus* accessions are the result of partial domestication from *A. hybridus* including gene flow from *A. quitensis*. *Amaranthus hypochondriacus* and *A. cruentus* could be domesticated versions of different geographical isolates of *A. hybridus*, as previously reported (Stetter *et al.* 2016; 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*) in one large clade conforms 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. The origin of the leafy amaranths from Eurasia is supported by the robust separation from the Central and South American weedy/grain amaranth types (Mosyakin and Robertson 1996). In addition, it was possible to assign a putative identity to 92% of the previously unknown amaranth accessions obtained from a local South African germplasm set. Further investigations of the whole chloroplast sequence to identify additional targeted 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. These findings support the hypothesis that increased phylogenetic resolution would be obtained by using whole chloroplast sequences rather than individual barcoding regions, due to the increased number of informative sites distributed across the chloroplast genome.

The phylogenomic results emphasize the power of adding genetic data to morphological observations in plant classifications. Future studies of the evolutionary relatedness within the *Amaranthus* genus should also focus on the addition of more geographically diverse species, especially within the Hybridus complex, to increase genetic variation associated with geographical separation. In this manner, a more complete picture could be obtained of grain amaranth domestication and the role of their weedy ancestors. 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.

Potential genomic SSR marker loci were identified from the assembled genome sequence and investigated across the germplasm accession set to determine their polymorphic content across different *Amaranthus* species. The screening of SSR markers using NGS platforms (and custom bioinformatic scripts) was validated through direct comparisons to results obtained through ABI technologies. The hypothesis that the NGS platform was superior to ABI in terms of the number of samples that can be multiplexed in one run, the cost involved and the increased informativeness of the allele call data output, was supported. The NGS approach would be particularly useful when screening large population sizes (such as global germplasm banks), with high numbers of multiplexed marker loci. Not only would it be possible to corroborate/contradict previous identifications based on morphology or limited genetic information, but it would also remove inadvertent duplications within amaranth collections for a more accurate species representation for downstream breeding applications.

During the current study, 6 marker loci were identified that proved to be polymorphic between different species groups within the *Amaranthus* genus. The resulting marker dendrogram (Chapter 4, Figure 4.10) corroborated the genetic relationships observed during the whole chloroplast phylogenomic analysis, namely that the accessions could be separated into their respective sub-genus clades and confirmed the basal relationships between leafy, weedy and grain amaranth. Of the unknown amaranth accessions,

approximately 70% could be assigned a species identity, which was less than for the whole chloroplast phylogeny. However, the accessions that were assigned “new” identifications were highly comparable between the two methods. The hypothesis that molecular markers would reveal the genetic relatedness between *Amaranth* accessions was supported in part, but the inclusion of additional markers would increase the present resolution substantially. In addition to inter-species classifications, the SSR markers will also be useful for intra-species classifications. Further mining of the *A. tricolor* genome will reveal additional SSR loci that can be screened for their polymorphic content across *A. tricolor* accessions. Investigating the genetic diversity of a globally diverse germplasm set of *A. tricolor* may shed light on their evolutionary history, as well as potential domestication events that have occurred independently in geographically isolated species.

Additional SSR marker loci can further be employed in expanded population studies within the *Amaranthus* genus, to determine the genetic distance or to estimate gene flow during hybridization events. Since both alleles of a heterozygous locus can be identified with SSR markers, the total number of alleles per locus identified can be representative of the ploidy level at that locus (Yagi *et al.* 2009; Feng *et al.* 2018). Therefore, the current and future markers developed from the *A. tricolor* genome could be used to determine the ploidy of accessions within different *Amaranthus* species, shedding light on their hybrid status. These profiles will be particularly useful in choosing potential parental lines for downstream breeding programs. The construction of linkage maps, mapping loci involved in quantitative traits and marker assisted selection will enable researchers to fine map traits of interest in *Amaranthus*.

Overall, the results from this study reveal the potential of powerful downstream analysis tools that can be developed from a draft genome assembly. Important aspects of the genetic relatedness between different international and local *Amaranthus* species/accessions were revealed using whole chloroplast phylogenomics and dendrogram analysis. Breeding programs currently being investigated in Africa (including South Africa) includes producing hybrid accessions between leafy and grain amaranths to incorporate agriculturally important traits of both into one genotype. Particularly, the pale seed morphology and nutritional profile of grain type amaranth and the palatability and drought resistance profiles of leafy type amaranth are important targets to incorporate into hybrid offspring. The information produced during the present study will enable researchers to make informed decisions on which amaranth accessions within the local germplasm set would

be most useful for breeding. The whole chloroplast phylogeny provides information on the genetic relationships between the accessions and therefore sheds light on which accessions would be most likely to hybridize, while the polymorphic SSR markers can be used for marker assisted selection of offspring based on the agriculturally targeted genes identified from the whole genome sequence. Eventually, the further development of these orphan crop species could significantly influence sustained food security in resource poor nations.

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
- 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
- Daru BH, Berger DK, Wyk AE (2015) Opportunities for unlocking the potential of genomics for African trees. *New Phytologist* 210:772-778
- Dawson IK, Hedley PE, Guarino L, Jaenicke H (2009) Does biotechnology have a role in the promotion of underutilised crops? *Food Policy* 34:319-328
- Feng B, Soojin VY, Zhang M, Zhou X (2018) Development of novel EST-SSR markers for ploidy identification based on *de novo* transcriptome assembly for *Misgurnus anguillicaudatus*. *PLoS One* 13:e0195829
- Greiner S, Sobanski J, Bock R (2015) Why are most organelle genomes transmitted maternally? *Bioessays* 37:80-94
- Grubben G (1980) Germplasm of vegetable and grain amaranths. *Plant Genetic Resources Newsletter*:33-35
- Lightfoot D, Jarvis D, Ramaraj T, Lee R, Jellen E, Maughan P (2017) Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (*Amaranthus hypochondriacus*) chromosomes provide insights into genome evolution. *BMC Biology* 15:74
- 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
- Mosyakin SL, Robertson KR (2016) New infrageneric taxa and combinations in *Amaranthus* (Amaranthaceae). In: *Annales Botanici Fennici* JSTOR, pp: 275-281
- Ncube B, Mupangwa W, French A (2018) Precision agriculture and food security in Africa. In: *Systems Analysis Approach for Complex Global Challenges*. Springer, Cham pp: 159-178
- Rastogi A, Shukla S (2013) Amaranth: a new millennium crop of nutraceutical values. *Critical Reviews in Food Science and Nutrition* 53:109-125
- Ribeiro JE, Pieterse PJ, Famba SI (2017) Vegetative growth of *Amaranthus hybridus* and *Amaranthus tricolor* under different watering regimes in different seasons in southern Mozambique. *South African Journal of Plant and Soil* 34:201-210
- Stetter MG, Müller T, Schmid K (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
- Sunil M *et al.* (2014) The draft genome and transcriptome of *Amaranthus hypochondriacus*: A C4 dicot producing high-lysine edible pseudo-cereal. *DNA Research* 6:585-602
- Tumushabe JT. (2018) Climate change, food security and sustainable development in Africa. In: *The Palgrave Handbook of African Politics, Governance and Development*. Palgrave Macmillan, New York pp: 853-868
- Van Rensburg WJ *et al.* (2007) African leafy vegetables in South Africa. *Water SA* 33:317-326
- Yagi M, Kimura T, Yamamoto T, Onozaki T (2009) Estimation of ploidy levels and breeding backgrounds in pot carnation cultivars using flow cytometry and SSR markers. *Journal of the Japanese Society for Horticultural Science* 78:335-343
- Ziervogel G *et al.* (2014) Climate change impacts and adaptation in South Africa. *Wiley Interdisciplinary Reviews: Climate Change* 5:605-620

APPENDIX A

MAND Data Script

Microsatellite Analysis from NGS Data (MAND)

@summary: Microsatellite finder using NGS data

@requirements: BioPython

@author:charles.hefer@gmail.com

@version:0.1

```
import sys, os
import optparse
from socket import gethostname
from multiprocessing import Pool

def parse_primer_file(filename):
    "Accepts a 3 column primer file, returns a dict for every SSR found"
    import csv
    handle = open(filename, "rU")
    #read the csv file into a dictionary, and return
    return list(csv.DictReader(handle, delimiter=","))

def find_ssr(ssr):
    "Steps over the fasta file, looking for SSRs"
    from Bio import SeqIO
    from Bio.Seq import Seq

    #Assign the required vars
    adaptor_fwd = ssr["Primer Forward"].rstrip()
```

```
adaptor_rev = ssr["Primer Reverse"].rstrip()
base = ssr["SSR"].replace(" ", "_").rstrip()
result_dir = "%s_Results" % base
results_dict = {}
```

```
#Prepare the output dir
```

```
if not os.path.exists(result_dir): os.makedirs(result_dir)
fasta_file = open(FASTA_INPUT, "r")
for entry in SeqIO.parse(fasta_file, "fasta"):
```

```
#Find the exact matches of the adaptors
```

```
if entry.seq.startswith(adaptor_fwd) and entry.seq.endswith(adaptor_rev):
    if len(entry.seq) not in results_dict:
        results_dict[len(entry.seq)] = [entry]
    else:
        results_dict[len(entry.seq)].append(entry)
```

```
#Create the FASTA files with the results
```

```
for entry in results_dict:
    out_handle = open("%s/%s_%s.fa" % (result_dir, base, entry), "w")

    SeqIO.write(results_dict[entry], out_handle, "fasta")
    out_handle.close()
```

```
def count_sequence_entries(filename):
    "Returns the number of entries in a fasta file"

    from Bio import SeqIO
    counter = 0
```

```
for entry in SeqIO.parse(open(filename, "r"), "fasta"):
    counter = counter + 1

return counter
```

```
def filter_ssrs(ssr):
    "Filter the allele frequency"
    ssr_count = []
    for filename in os.listdir("%s_Results/" % ssr):
        if filename.endswith(".fa"):
            ssr_count.append((filename, count_sequence_entries("%s_Results/%s" % (ssr,
            filename))))

    sum_counts = 0
    for filename, count in ssr_count:
        sum_counts = sum_counts + count

    outhandle = open("%s_Results/%s_data.csv" % (ssr, ssr), "w")
    outhandle.writelines("Len,Count\n")
    filterhandle = open("%s_Results/%s_data_filtered.csv" % (ssr, ssr), "w")
    filterhandle.writelines("Len,Count\n")

    for filename, count in ssr_count:
        outhandle.write(("s,%s\n" %
            (filename.replace(ssr + "_", "").replace(".fa", ""), count)))

    #Filter allele frequency
    if float(count)/sum_counts >= FILTER:
        filterhandle.write(("s,%s\n" %
            (filename.replace(ssr + "_", "").replace(".fa", ""), count)))

    outhandle.close()
    filterhandle.close()

def get_first_entry(filename):
```

"Returns the first entry in a fasta file"

```
from Bio import SeqIO
```

```
for entry in SeqIO.parse(open(filename, "r"), "fasta"):
```

```
    return [entry]
```

```
def representative_sequence(ssr):
```

"For every ssr, extract the first entry of the filtered dataset"

```
from Bio import SeqIO
```

```
for line in open("%s_Results/%s_data_filtered.csv" % (ssr, ssr)):
```

```
    if "Len" in line: continue
```

```
    length, count = line.split(",")
```

#convert the length to a filename

```
    filename = "%s_%s.fa" % (ssr, length)
```

```
    handle = open("%s_Results/%s_representative.fa" % (ssr, filename.replace(".fa", "")), "w")
```

```
    SeqIO.write(get_first_entry("%s_Results/%s" % (ssr, filename)), handle, "fasta")
```

```
    handle.close()
```

```
def run_repeatfinder(ssr):
```

"Run Tandem Repeat Finder"

```
for line in open("%s_Results/%s_data_filtered.csv" % (ssr, ssr)):
```

```
    if "Len" in line: continue
```

```
    length, count = line.split(",")
```

#convert the length to a filename

```
    filename = "%s_%s_representative.fa" % (ssr, length)
```

```
    cmd = "../trf407b.macos64 %s 2 5 5 80 10 30 30" % (filename)
```

```
    os.system("cd %s_Results; %s; cd ../" % (ssr, cmd))
```

```
def parse_repeatfinder(ssr):

    "Parse the results for Repeatfinder"

    class RepeatElement:

        def __init__(self):

            pass

    import re

    size_re = re.compile(".+Period size:\s+(\d+)")
    copynumber_re = re.compile(".+Copynumber:\s+(\d+\.\d+)")
    consensus_re = re.compile(".+Consensus\ssize:\s+(\d+)")
    pattern_re = re.compile("[A-Z]+")

    entries = []

    for line in open("%s_Results/%s_data_filtered.csv" % (ssr, ssr)):

        if "Len" in line: continue

        length, count = line.split(",")

    #convert to a filename

    filename = "%s_Results/%s_%s_representative.fa.2.5.5.80.10.30.30.1.txt.html" % (ssr, ssr,
length)

    handle = open(filename, "r")

    for line in handle:

        if "Copynumber" in line:

            entry = RepeatElement()

            entry.size = size_re.match(line).groups()[0]

            entry.copynumber = copynumber_re.match(line).groups()[0]

            entry.consensus_size = consensus_re.match(line).groups()[0]

            entry.ssr = ssr

            entry.length = length

        if "Consensus pattern" in line:
```

```

        line = handle.next()

        entry.pattern = pattern_re.match(line).groups()[0]

        entries.append(entry)

        entry = None

    handle.close()

handle = open("%s_Results/%s_summary.txt" % (ssr, ssr), "w")

handle.write(",".join(["SSR name", "SSR length", "Repeat size", "Repeat Copynumber", "Repeat
Cons size", "Repeat Pattern\n"]))

for entry in entries:

    handle.write(",".join([entry.ssr, entry.length, entry.size, entry.copynumber,
entry.consensus_size, entry.pattern + "\n"]))

handle.close()

def __main__():

    "Parse the command line arguments"

    parser = optparse.OptionParser()

    parser.add_option("-i", "--input", default=None, dest="input", help="The input file")

    parser.add_option("-p", "--primer", default=None, dest="primer",
        help="The 3 column SSR primer file (SSR, forward_primer, reverse_primer)")

    parser.add_option("-t", "--threads", default=2, dest="threads",
        help="The number of threads to use during parsing, defaults to 2")

    parser.add_option("-f", "--filter", default=0.4, dest="filter",
        help="The allele frequency filter, smaller than 1. Default = 0.4")

    (options, args) = parser.parse_args()

    global FASTA_INPUT

    FASTA_INPUT = options.input

    global FILTER

    FILTER = float(options.filter)

    #Parse the primer file into a dictionary

```

```
primers = parse_primer_file(options.primer)
```

```
#Obtain the list of SSRs from the input file
```

```
ssrs = [entry["SSR"].replace(" ","_") for entry in primers]
```

```
pool = Pool(processes=int(options.threads))
```

```
pool.map(find_ssr, primers)
```

```
#Filter the result files
```

```
pool.map(filter_ssrs, ssrs)
```

```
#Generate a representative sequence
```

```
pool.map(representative_sequence, ssrs)
```

```
#Run Tandem Repeat Finder
```

```
for ssr in ssrs:
```

```
    run_repeatfinder(ssr)
```

```
#Parse Repeatfinder
```

```
for ssr in ssrs:
```

```
    parse_repeatfinder(ssr)
```

```
if __name__ == "__main__":
```

```
    __main__()
```