An investigation into the systematics and population diversity of *Tylosema* (Leguminosae).

by

Lizo Masters

A dissertation submitted in partial fulfilment of the requirements for the degree

Magister Scientiae

In the Department of Plant and Soil Sciences
University of Pretoria
Pretoria

Supervisor: Prof Nigel P. Barker

Co-Supervisor: Prof Martin P.A. Coetzer

May 2020

Dissertation Summary

The following dissertation in the product of research conduction for an MSc in phylogenetics and population genetics of southern African *Tylosema* sp., with a specific focus on how this research benefits the effort to domesticate these plants. The dissertation comprises of a literature review, two research chapters, and a final synopsis examining implications of the results obtained and perspectives on future research. The two research chapters are written as stand-alone manuscripts concerning two independent research aims, namely: the phylogenetic assessment of southern African *Tylosema sp*., and the development and testing of microsatellite markers in *Tylosema sp.* Formatting and referencing follow the guidelines of the journal 'South African Journal of Botany'. A single list of references and supplementary data are provided at the end.

Chapter 1 is a review of current available literature on *Tylosema* and includes relevant information regarding the nutritional and agricultural value of *Tylosema* species, as well as past taxonomic treatments and genetic variability studies conducted to date. This chapter also includes relevant background information on the methodologies utilized in this study and the motivations behind choosing them. The chapter concludes with the aims of the study and a summary of the approaches taken.

Chapter 2 presents a molecular phylogenetic assessment of species limits between *T. esculentum, T. fassoglense,* and *T. angolense* using two chloroplast and two nuclear markers. One of the nuclear markers utilized, serine hydroxymethyltransferase (SHMT), contained an insertion/deletion (indel) event in *T. esculentum* specimens which was assessed using the free online software Indelligent. This increased the phylogenetic utility of the SHMT marker and allowed for well-resolved phylogeny in the combined dataset set. Species limits between the three taxa were confirmed and

evidence of two distinct lineages was found in *T. esculentum*. Shortcomings in the results are discussed and means by which they can be overcome in future researched are outlined in detail.

Chapter 3 presents the development of microsatellite markers using genomic data of *T. esculenum* through the program MSATCOMMANDER. Select microsatellite markers were successfully amplified in *T. esculentum* and *T. fassoglense,* as well as evidence of polymorphisms across both species. The utility of these markers and potential caveats as a result of polyploidy in the genus are also discussed.

Chapter 4 presents a summary of the results obtained and provides context for how they may inform future research. The agricultural significance of *Tylosema sp.* is reiterated in light of these results.

Abstract

Tylosema esculentum, *T. fassoglense*, and *T. angolense* are legume species found across southern Africa that produce highly nutritious seeds and tubers. These species have significant agricultural potential but are severely understudied in the wild. As a result of this neglect the taxonomy of these three species is uncertain. Additionally, little is known about the genetic diversity within and between wild populations of the three species largely due to the fact that very few molecular markers have been developed in *Tylosema*. Both of these issues will greatly reduce the efficacy of any domestication attempts provided they are not addressed.

This MSc addresses the taxonomic issues present in *T. esculentum*, *T. fassoglense*, and *T. angolense*. A molecular phylogenetic assessment using the chloroplast markers *trn*L-F intron and spacer and *psb*A-*trn*H intergenetic spacer, the high copy nuclear marker ITS, and the low copy nuclear gene SHMT was produced using Bayensian inference methods. To aid in future genetic diversity studies within the genus, microsatellite markers were also developed using genomic DNA from *T. esculentum* provided by Dr Chris Cullis, Case Western Reserve University, Ohio, USA. These markers were tested for amplification and polymorphisms in both *T. esculentum* and *T. fassoglense* populations to determine whether they were suitable for genetic assessments across the *Tylosema* genus.

Phylogenetic analyses confirmed the species level description of the recently described *T. angolense*. Both *T. angolense* and *T. fassoglense* belong to a separate lineage to *T. esculentum*, which itself is split into two lineages: a Kalahari Desert lineage and a South African Highveld grassland lineage. This study has demonstrated that the evolutionary history of *Tylosema* is likely more complex than current taxonomic treatments indicate. Future research should entail

molecular/genomic phylogenetic analyses across the entire genus in order to determine accurate species limits in *Tylosema*.

A total of 46 microsatellite markers were developed using the program MSATCOMMANDER, five were selected for amplification and polymorphism detection through gel electrophoresis. All five markers successfully amplified in both *T. esculentum* and *T. fassoglense*, four of which were polymorphic in both species, one showing clear length polymorphisms between the two species. These results confirm that microsatellite markers designed in *T. esculentum* can likely be used across *Tylosema* and potentially other closely related genera.

The results of this study can provide useful insight into the value of *Tylosema* species for agriculturalists and researchers investigated in their domestication. Researchers should take note of the phylogenetic relations between southern African *Tylosema sp.*, especially with *T. esculentum* where Kalahari and Highveld populations belong to different lineages and may have different genetic/chemical properties relavent to food scientists and agricultural breeders. The SSR markers developed can also aid in assessing genetic diversity in these wild populations and help ensure than unique genetic lineages within *Tylosema* species are protected.

Declaration

I, the undersigned, declare that the dissertation which I hereby submit for the degree Magister Scientiae at the University of Pretoria, is my own independent work and has not previously been submitted by me for a degree at this or any tertiary institution.

Signature:……..

Name: Lizo Masters

Date: 11/05/2020

Acknowledgements

To the following individuals and institutions, I would like to extend a word of deep and sincere gratitude. Without their support and assistence the completion of this degree would not have been possible. To the South African National Seed Organisation for providing me with funding and mentorship through their scholarship program. To Dr Chris Cullis of Case Western Reserve University, for allowing us to use genomic data his lab had produced and for allowing me to assist him and his class with fieldwork in Namibia, a highlight of my university career. To Mr Arnold Frisby (UP) for providing me with assistance and good company during the long hours of sample collecting. To Dr Juan Forster (UP), Mr Jason Sampson (UP), Dr Bronwyn Egan (University of Limpopo), and Mr John Burrows (Buffelskloof Private Nature Reserve) for providing additional plant material for DNA work. To Dr Kenneth Oberlander and Ms Magda Nel from the H.G.W.J. Schweickerdt Herbarum at UP for all their advice throughout the degree and allowing me to steal their time almost weekly. To my co-supervisor, Prof Martin P.A. Coetzer, for his expertise in microsatellite development and his invaluable contributions toward the research conducted. Finally, to my primary supervisor, Prof Nigel P. Barker, who has been a phenomenal mentor and teacher to me for the past three. Without him my love for phylogenetics, taxonomy, and for marama beans, would not have taken root.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction and Literature review

Tylosema (Schweif.) Torre and Hillcoat, is a genus of perennial legumes endemic to sub-Saharan Africa. It comprises five recognized species at present (*T. esculentum* (Burch.) A. Schreib., *T. fassoglense* (Schwief) Torre and Hillc., *T. angolense* P. Silveiro & P. Castro, *T. humifusa* (Pic.Serm. & Roti. Mish) Brenan, and *T. argentea* (Chiov.) Brenan). Of these, *T. fassoglense* has the largest distribution of the five species, spanning from South Sudan and Ethiopia down and across southern Africa as far south as Swaziland (Coetzer and Ross, 1976; Coetzer et al., 2011). The remaining four species have comparatively restricted distributions with *T. argentea* and *T. humifusa* overlapping in the Somali-Masai regional center of endemism (Kenya, Somalia, and Ethiopia); *T. esculentum* is limited to the Kalahari Desert and parts of North-Western South Africa; and *T. angolense* is endemic to Angola (Gillet et al., 1971; Coetzer and Ross, 1976; Castro et al., 2005; Lewis and Forest., 2005).

Tylosema species can be found across, and often share, a wide range of habitats and biomes ranging from wooded savannas and bushveld, open plains and grasslands, rocky outcrops, as well as the aforementioned Kalahari Desert and other arid areas (Coetzer and Ross, 1976). Morphologically, the genus is characterized by its large underground tuber used for water and nutrient storage and above-ground sprawling lianas. Most *Tylosema* species are characterized by the presence of tendrils, with the exception of *T. angolense* and *T. humifusa* (Castro et al., 2005). The flowers of *Tylosema* have nine or ten stamens, of which only two are fertile (Verdoorn, 1959; Coetzer and

Ross, 1976), lobed calyxes, and are functionally heterostylous. This is the first known instance of heterostyly in the Fabaceae family (Hartley et al., 2002). A unique character found in *T. esculentum* is a mucilaginous substance coating the anthers which putatively aids in pollen attachment to pollinators (De Frey et al., 1992). Studies on *T. esculentum* greatly outnumber those of other *Tylosema* species, though it is reasonable to assume that many characters (both morphological and otherwise) seen in *T. esculentum* are likely present in *Tylosema* as a whole due to the relative morphological and genetic similarity between species (Coetzer et al., 2011).

Tylosema: food source and domestication

Tylosema esculentum seeds have been eaten by numerous cultures and tribes across southern Africa for possibly millenia (Keith and Renew, 1975; Van Wyk and Gericke, 2000). Colloquially, the species is known by numerous names: gemsbok boontjie or braaiboontjie in Afrikaans, ombanui in Herero, and most commonly marama, from SeTswana (Jackson et al., 2010). The large seeds, or beans, of marama plants have numerous culinary uses in southern Africa. The seeds are roasted or boiled which gives them an improved flavour and texture, comparable to cashews or almonds (Jackson et al., 2010). Cooking marama beans is essential for human consumption as heat is required to denature strong trypsin inhibitors in the cotyledons which are harmful to humans (Powell, 1987; Bower et al., 1988; Van der Maesen, 2006). The roasted seeds can then be ground into a fine powder to make porridge and bread, boiled and strained to use the reserved liquid as a drink, or simple eaten whole as a snack (Van der Maesen, 2006; Jackson et al., 2010; Coetzer et al., 2011). In the Kalahari, marama beans have been a major food staple for various cultures, contributing up to 75% of their plant-based food intake in certain tribes (Chimwamurombe, 2010).

Marama beans are an incredibly nutritious food source despite being entirely wild plants. Their lipid content and yield ranging between 29-42%, which is comparable to many commercial vegetable oil sources such as sunflower seeds, rapeseed, soybean, and peanuts (Bower et al., 1988; Ketshajwang et al., 1999; Holse et al., 2010). Marama bean oil contains a peroxide value (PV) of 20.3 meq/kg (milliequivalents per kilogram), meaning it would require industrial refinement for improved shelf life, though the recommended PV for extra virgin olive oil is 20 meq/kg (Ketshajwang et al., 1999). Marama bean protein content is also unusually high for a wild legume with researchers estimating levels between 30%-39%, making them comparable to a number of varieties of soybeans (Mmonatau, 2005; Jackson, 2010; Holse et al., 2010). In fact, marama beans have a higher protein content than a range of agricultural legumes namely peas (23%), broad beans (23%), and lupin (31%) (Bower et al., 1988). It is worth noting that Holse et al. (2010) found that the protein content of *T. esulentum* beans from South African (34-36%) populations is statistically significantly higher than those from Botswana and Namibia (29-32% and 30-35% respectively).

The young tubers of *T. esculentum* are also edible and can be prepared in similar ways to most root vegetables. Nutritionally, the tubers have a protein content of 9%, more than double that of most potato or sweet varieties (Dakora et al., 1999). Young tubers (1 or 2 years) are preferred for consumption as the tubers becomes more fibrous and develops an astringent taste as they mature, though in the wild large tubers are an important source of water for both humans and animals (Coetzer and Ross, 1976; Keegan and van Staden, 1981). Dakora et al., (1999) found that the tubers do not form nodules with nitrogen fixing bacteria, even though the nitrogen content of *T. esculentum* leaves, stem, and tuber (roots) were greater than those of three nodulating, N2 fixing *Acacia* species. *T. esculentum* grows in nutrient poor soils in the wild, but nitrogen isotope analysis

shows that these soils are the source of its nitrogen (Dakora et al., 1999). *T. esculentum* efficiently harnesses the trace amounts of nitrogen in the ambient soil, storing nitrogen in high concentrations in its tubers and leaves (Thomas, 2004). Indigenous tribes throughout the Kalahari grind *T. esculentum* leaves into a paste used to treat open wounds (Chimwamurombe, 2010). Leaves are also supposedly eaten by livestock, though there are conflicting reports of their palatability (Van der Maese, 2006). Extracts from *T. esculentum* tubers and leaves have been used by traditional African healers to treat diarrhoea (Chingwaru et al., 2007). Chingwaru et al., (2011) found that extract from marama seeds and seed coats extracts could be used as antivirals against rotavirus infections; viruses responsible for incidences of lethal diarrhoea in infants and livestock across Africa.

T. esculentum has been described as an 'orphan' or 'lost' crop since it has huge agricultural potential but has received relatively little scientific attention (National Research Institution, 2006; Cullis and Kunert, 2017). Thankfully, this trend has gradually changed in recent years as scientists and agriculturalists around the world have noted the plant's potential in food and medicine (Powell, 1987; Bower et al., 1988; Travlos et al., 2006; Chimwamurumbe, 2011; Cullis and Kunert, 2017; Cullis et al., 2019). What makes the species such an attractive candidate for domestication in so many countries is the combination of highly nutritious seeds and tubers, and the ability to grow in nutrient poor soils in the water restricted Kalahari Desert (Museler and Schonfeldt, 2006; Cullis et al., 2019). Current projections show that climate change will have a severe negative impact on agricultural production in Africa, and more so for small-scale subsistence farmers (Muller et al., 2011). African farmers are especially vulnerable to the negative effects of climate change. This is due to a lack of quick adaptability in the face of changing environmental conditions (Hassan and

Nhemachena, 2008). Farming remains the main source of food and income for poor communities across sub-Saharan Africa, a region which is expected to see frequent changes in weather patterns, with an increase in the occurrence of droughts and floods (Nhemachena and Hassan, 2007; Gbetiouo et al., 2010). The need for hardy crops that can survive these harsh climatic changes should expedite the domestication of *T. esculentum*.

T. esculentum domestication has gained attention beyond southern Africa, with projects initiated as far as Australia, Israel, and the USA. Powell (1987) planted marama seeds for field trials in arid, controlled regions in Texas. His trials showed that plants began producing large seed crops between 4 and 5 years of age. Seed germination trials have shown that a number of mechanical scarification and water immersion methods result in successful germination and that there is likely no physiological dormancy present in *T. esculentum* (Travlos et al., 2006). Marama seeds also germinate best in loose, sandy soils and emergence is low in soils with clay like properties (Travlos et al., 2007). Observations of marama plants under drought stress, both in the wild and in controlled experiments, indicate that a variety of mechanisms for drought avoidance allow *T. esculentum* to survive periods of water shortage. Photosynthesis in *T. esculentum* is comparable to other C3 photosynthetic plants, in that it is not adapted for water efficiency (Mitchell et al., 2005). Instead, stomatal conductance preempts rising temperatures at midday, the bilobed leaves fold inwards onto each other to reduce the surface area exposed to sunlight, and water and sugars are mobilized from the underground tuber to maintain water potential and turgor in above ground organs (Mitchell et al., 2005; Karamanos and Travlos, 2012).

Food and nutrition scientists have moved forward in measuring the physicochemical properties of marama beans to determine other applications the seeds could have in food production (Mnnonatau, 2005; Museler, 2005; Maruatona et al., 2010; Jackson et al., 2010; Kayitesi et al., 2012; Nyembwe et al., 2015; Nyembwe et al., 2018). The high nutrient content in marama seeds makes them ideal for supplementing low nutrition staple foods e.g. bread, sorghum, maize meal, etc. (Mmonatau, 2005). Preliminary tests to determine the marketability of marama beans have been conducted, indicating that marama food products are well known and even regularly sold within rural communities in Namibia, Botswana, and South Africa (Faria et al., 2011). These findings were supported by Mahgoud et al. (2013), who elaborated that marama seeds were a recognised food source in rural Botswana communities, but noted that the general public was largely unaware of the nutritional benefits the seeds offer or that efforts to domesticate the plant were underway.

The value *T. esculentum* seeds and tubers could have in arid countries is well established given the extent of scientific literature and indigenous knowledge available. However, few studies have been conducted on *T. esculentum* populations in the wild; how vulnerable they are to extinction and the impact urban expansion has on their survival. Little is known about the genetic variability of *T. esculentum* across its entire distribution, knowledge that would greatly benefit domestication and breeding strategies (Chimwamurombe, 2011). Additionally, there seems to be no interest in domesticating other *Tylosema* species. *T. fassoglense* seeds and young tubers are also edible and have been utilized for food and medicine by many people across Africa (Coetzer et al., 2011). In southern Africa the common name ‘marama’ is also applied to *T. fassoglense* seeds (Coetzer et al., 2011). The nutritional value of *T. fassoglense* seeds is similar to *T. esculentum*, with equally high

protein and lipid content that rivals many commercial legumes (DuBois et al., 1995). In certain parts of South Africa, the two species even share the common name, ‘marama’ (Coetzer et al., 2011). *T. angolense* is closely related to these two species and is also edible. It could thus also be a valuable food source in Angola or beyond (Castro et al., 2005).

The research bias towards *T. esculentum* is understandable given the potential role it could play in food security in drought stressed environments. However, this bias highlight two central issues: Firstly, little is known about the ecology, pollination, and genetic diversity within the species in the wild. Even species limits within the genus *Tylosema* are questionable (see later). Addressing these gaps in our knowledge helps speed the domestication of wild plants (Hickey et al., 2019). The second issue was briefly mentioned above; the remaining *Tylosema* species are also edible and hardy plants that could prove useful in agriculture. The neglect *Tylosema* has seen from contemporary science means that foundational taxonomic and diversity information is missing, and this problem must be addressed to ensure domestication efforts are as efficient as possible.

Taxonomy

The taxonomic history of *Tylosema* is plagued with ambiguities and inconsistencies. Broadly, *Tylosema* is a member of the monophyletic legume subfamily Cercidoideae (Legume Phylogeny Working Group [LPWG], 2017). Originally, specimens of *Tylosema* were lumped into the genus *Bauhinia* L., a genus with arguably an even more complicated taxonomic history. The earliest description of *T. esculentum* by a Western botanist appears in William Burchell’s *Travels in the Interior of Southern Africa* (Burchell, 1822). Therein he describes the plant species *Bauhinia*

esculenta (B. Catal. George. 2414.), as having ‘long slender branches spreading on the ground’ and ‘rounded leaves nearly divided in two’ and deems it ‘the only species of *Bauhinia* hitherto discovered in Southern Africa’. Burchell also makes reference to the edible ‘roots’ (tuber) and seeds in his description; characters expounded upon earlier in this review.

Though members of *Tylosema* have historically been included within *Bauhinia*, they have been recognized as occupying an unofficial though distinct subgroup therein (Coetzer et al., 2011). Subsequently, numerous authors have proposed that *Tylosema* be elevated to the rank of genus, though *Tylosema* is not unique in this regard. Several subgroups within *Bauhina* s.l. have been proposed to be ranked as genera e.g. *Phanera*, *Barklya*, *Gigasiphon* etc. (Schery, 1951; Torre and Hillcoat, 1955; De Wit, 1956; Wunderlin et al., 1987). This taxonomic uncertainty is unsurprising given the global distribution of taxa as well as the wide variety of growth forms and morphological diversity members of *Bauhinia* s.l. exhibit (trees, shrubs, lianas, etc.) (Hao et al., 2003; Sinou et al., 2009). At present *Bauhinia* s.l. remains an unresolved complex with numerous classifications proposed to clarify generic limits (Hao et al., 2003; Zhang, 1995; Lewis and Forest 2005; Wojciechowski et al., 2004).

However, *Tylosema* is a well-supported monophyletic group as shown in molecular phylogenetic studies based on chloroplast DNA data (Lewis and Forest 2005; Sinou et al., 2009; Wang et al., 2018; LPWG 2017; Bruneau et al., 2008). Additionally, studies have shown *Tylosema* chloroplast genomes (plastomes) contain unique features not present in other members of *Bauhinia* s.l. or Cercidoideae in general. Kim and Cullis (2017) found that *T. esculetum* plastomes contain a unique inversion of 7479 base pairs (bp), containing six genes, in the large single copy (LSC) region, and found peculiar intraspecific variations within *T. esculentum* plastomes; effectively two different chloroplast genomes were found within a single specimen.

Wang et al. (2018) evaluated the evolution of chloroplast genomes within Cercideae. There they elaborated on the findings of Kim and Cullis (2017), confirming the presence of the LSC region inversion in *T. fassoglense* but not in *Bauhinia* s.l., meaning the inversion is a likely synapomorphy in *Tylosema*. Additionally, Wang et al. (2018) found a second 38k bp inversion in *Tylosema* taxa tested (*T. esculentum* and *T. fassoglense*), which is mediated by a 29 bp inverted repeat at either end of the 38k bp inversion. Interestingly, it was found that within a single *Tylosema* specimen two chloroplast genomes could be present with respect to this 38k bp inversion (i.e. some plastomes with and without the inversion are both present in the same individual). The predominance of one plastome over the other in a single *Tylosema* specimen differed from individual to individual. Wang et al. (2018) had not commented on whether or not this finding was in accordance with Kim and Cullis's (2017) finding of two plastomes present in an individual *T. esculentum* though noted it was unique in Cercidoideae. Inversions within plastomes are fairly common in papilionoid legumes but in no other legume subfamily (Doyle et al., 1996).

Tylosema is also morphologically and geographically distinct within *Bauhinia* s.l., as mentioned earlier, but variation within *Tylosema* at both inter and intraspecific levels is extensive and has led to complications in determining species limits. Brummitt and Ross (1976) commented on the specimen described as *Bauhinia bainesii* Schinz in Schinz's own collection from South West Africa (Namibia). However, Schinz also makes reference to a collection from Rhodesia (Zimbabwe) by Baines, inferring that the two specimens are one and the same species. Brummitt and Ross examined both specimens based on vegetative characteristics (leaves and tendrils) and concluded that the Schinz's collection is 'undoubtedly' *T. esculentum* and Baines had collected *T. fassoglense*.

Coetzer and Ross (1976) presented a taxonomic treatment of *T. esculentum* and *T. fassoglense* in *Flora of Southern Africa*, owing to both being present in South Africa. They present clear morphological distinctions between the two species. *T. fassoglense* has demonstrably longer vines, longer functional tendrils, rusty pubescence on young growths and on the veins of leaves, longer petioles, leaves shallowly biloded, and larger, flatter fruit pods that are distinctly woody. *T. esculentum* is a much smaller plant with shorter vines, small and superficial tendrils, shorter petioles, and deeply bilobed leaf blades that are glabrous along the margins of veins. The authors also show the two species being geographically separated in southern Africa. *T. fassoglense* is found in Limpopo, Mpumalanga, and northern Kwa-Zulu Natal in South Africa, and in Swaziland. *T. esculentum* in South Africa is restricted to Gauteng, the North-West province, and parts of the Northern Cape; it is also present in Namibia and Botswana.

Castro et al. (2005) produced the most recent taxonomic treatment of *Tylosema* in an effort to further clarify species limits, and with the specific goal of describing strange *Tylosema* specimens found in southern Angola which did not exhibit diagnosable characteristics present in any of the then recognized *Tylosema* species. Castro et al. (2005) made use of numeric or phenetic analyses on a range of morphological characters, as well as an analysis of pollen morphology, for 258 *Tylosema* specimens spanning the geographical range of the genus. A notable caveat in this analysis was the omission of *T. argenteum* due to only a single herbarium specimen being available from a range of herbaria accessed. This drawback is due to the lack of sampling attention *Tylosema* specimens have received in general (Coetzer et al., 2011) and consequently compounds the already difficult task of establishing a rigorous taxonomy in *Tylosema*.

From their palynological assessment, Castro et al. (2005) determined minimal qualitative variation across *Tylosema* species for pollen morphology. *Tylosema* pollen grains are tricolporate, radially

symmetrical, and prolate in shape. Banks et al. (2014) confirmed these findings adding that *Tylosema* pollen type was unspecialized. Castro et al. (2005) were able to produce a dichotomous key for the four *Tylosema* species assessed using log functions of the P/E (length of polar axis/length of equatorial diameter) ratios. A Tukey's test showed that said key could identify the species of *Tylosema* from a pollen grain sample with a 95% certainty.

General pollen morphology detected in *Tylosema* species was consistent with other reports found in the literature though with some notable differences. For example, *T. fassoglense* pollen morphology matched findings by Coetzer et al. (1981) though the P and E values were larger in this study, and smaller than values obtained by Smith (1964) and Schmitz (1973). P and E values in *T. esculentum* were also larger in this study compared to Coetzer et al. (1981). Castro et al. (2005) also reported numerous aborted pollen grains and cites that this, as well as the variation in pollen size, is likely due to variation in ploidy level seen in the two *Tylosema* species (Goldblatt and Davids, 1977; Monaghan and Halloran, 1996), a topic which will be addressed in further detailed later.

Inquiry into the phenetic analyses conducted by Castro et al. (2005) reveals interesting complications in *Tylosema* species limits (Figure 1). The principal components analysis (PCA) (see Figure 1) showed *T. fassoglense* as a distinct group united by morphological features that corroborated with those described by Coetzer and Ross (1976). *T. humifusa* formed a single distinct group with subtle intraspecific grouping representing regional variation between Kenyan and Somali specimens owing to Kenyan specimens having slightly larger leaves and longer petioles. *T. humifusa* leaves are small (2-4.5 cm x 2.2-5.5cm) with conspicuously white to grey hairs compared to the rusty brown hairs present in other species.

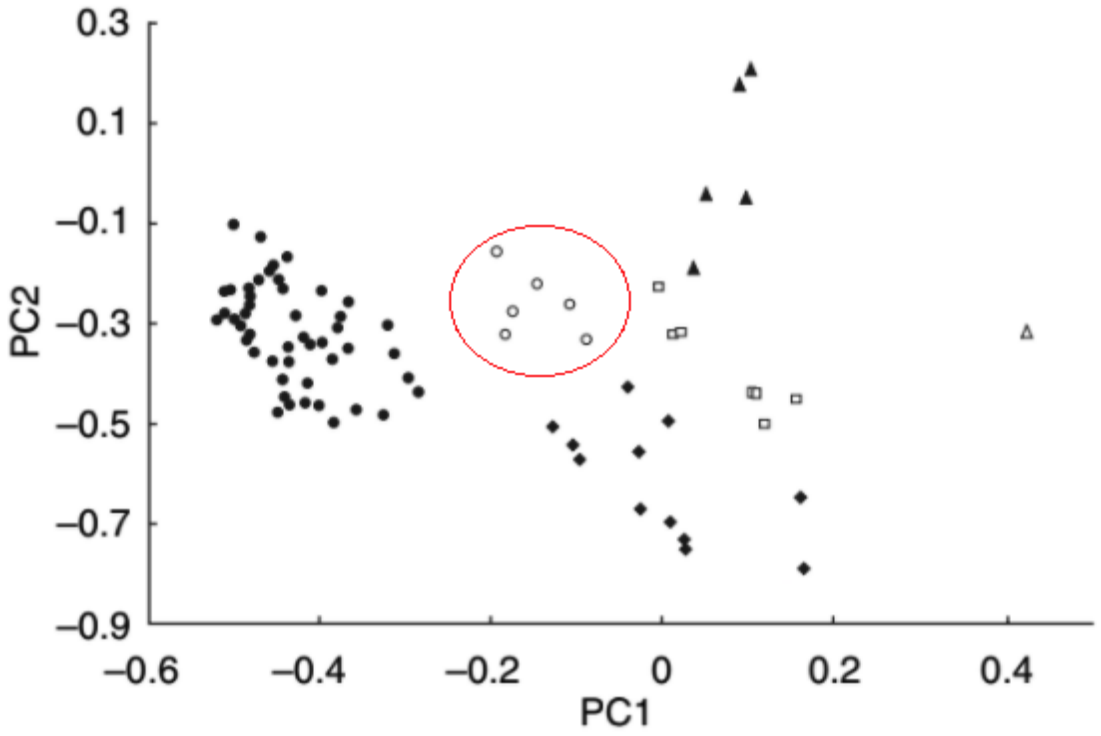

Figure 1 The above Principal Components Analysis (PCA) was produced by Castro et al., (2005). Specimens in the red circle (○) are undescribed Tylosema *taxa that are intermediates of* T. angolense *(□),* T. fassoglense *(●), and* T. esculentum *(◆).* (△) *and* (▲) *are* T. humifusa *and* T. argentea *respectively.*

Tylosema specimens collected in southern Angola clustered relatively distinctly in both the PCA and phenogram (see Figure 2) produced. These findings prompted Castro et al. (2005) to assign these specimens the rank of species, namely *T. angolense*. The unique assemblage of morphological characters putatively present in *T. angolense* include the leaf blade being apical bilobe by 1/6 to 1/2 of its length, microscopic leaf hairs present without a swollen base, and a distinct lack of tendrils. The authors note that *T. angolense* specimens were likely initially identified as *T. esculentum* given that the latter is said to occur in Angola. *T. esulentum* did group

distinctly from *T. angolense* and were confirmed to have far deeper biloded leaves (up to 4/5s of the total leaf length) and had short tendrils present, conforming to morphological descriptions in Coetzer and Ross (1976).

Castro et al. (2005) used 80 *Tylosema* specimens for their phenetic analyses, with 6 specimens grouping somewhere between *T. fassoglense*, *T. angolense*, and *T. esculentum.* All 6 possessed some combination of morphological characteristics rendering them unidentifiable by the diagnostic treatment for *Tylosema* species Castro et al. (2005) produced, rendering them atypical. Two specimens had leaves with very shallow distal sinuses similar to *T. fassoglense,* but the leaf size and the lack of tendrils were reminiscent of *T. angolense*. The location where these specimens were originally sampled was unfortunately not specified.

Castro et al. (2005) also produced a phenogram showing these atypical *Tylosema* specimens clearly grouping with the newly described *T. angolense* (Figure 2). Though, it is true that specimens which could be identified as *T. angolense* clustered distinctly from the atypical specimens within this broader group. Castro et al. (2005) also noted that the *T. angolense* cluster branches off in the phenogram at the same coefficient distance as the other three species clusters. They argued that these two facts further the case for the recognition of *T. angolense* as a species and comment that the atypical specimens cannot be commented on at present given how few there were and that they were sampled from widely dispersed geographical regions (Angola, Zambia, and Mozambique).

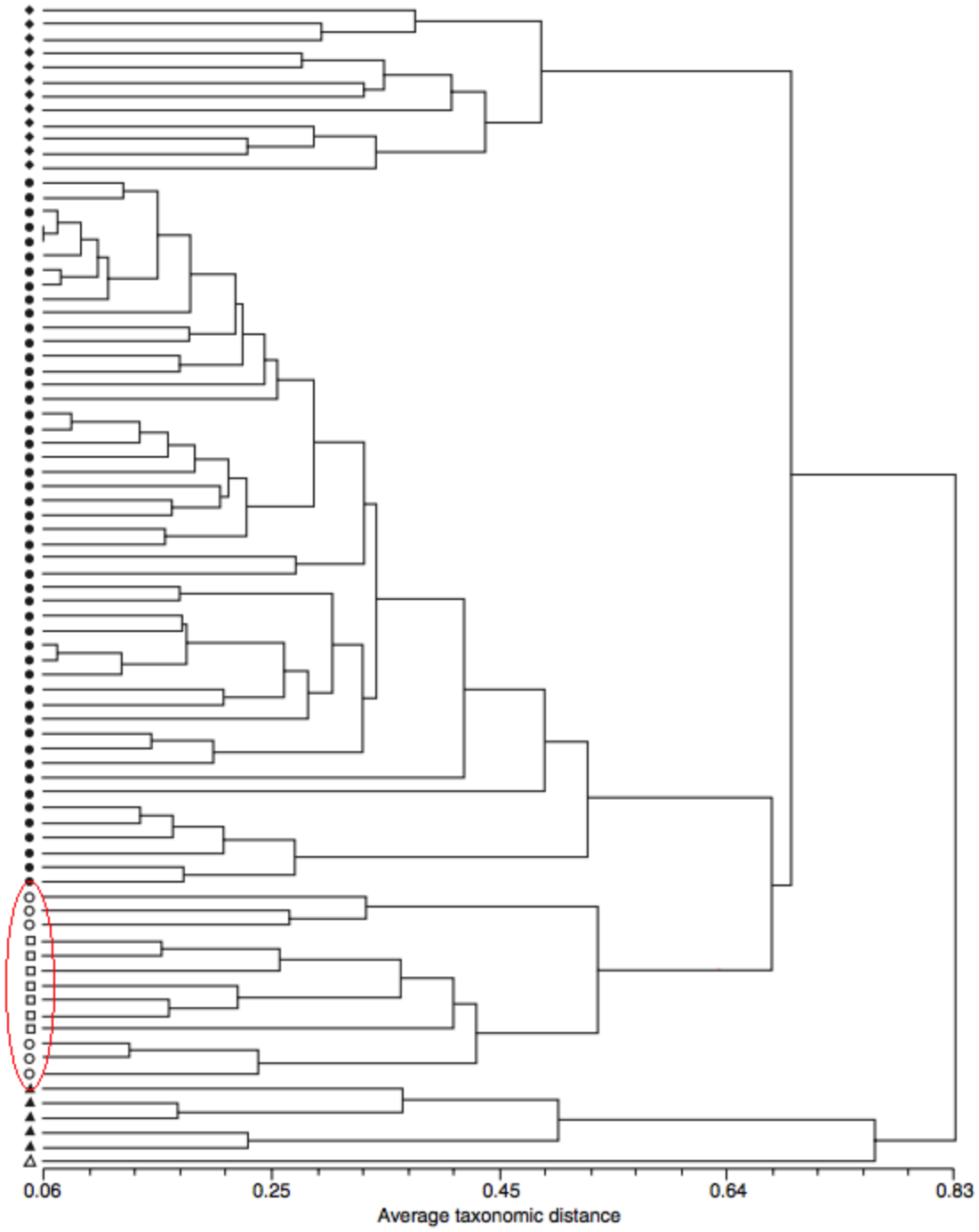

Figure 2 Phenogram produced by Castro et al., (2005). Specimens in the red circle show T. angolense *(□) imbedded between undescribed atypical* Tylosema *specimens (○) found in the study.*

This assessment on the placement of atypical *Tylosema* specimens seems somewhat hopeful as the number of atypical specimens and *T. angolense* specimens found are perfectly comparable (6 and 7 specimens respectively). It is also noted that two of these atypical specimens were collected in southern Angola near the Cubango River where *T. angolense* specimens were sampled. These atypical *Tylosema* specimens were judged not to be *T. esculentum* on account of the absence of tendrils. The evidence for the recognition of *T. angolense* is tenuous at best and the presence of these atypical specimens is a glaring issue in this taxonomic treatment. Compounding the problem is the distinct lack of field information generally present on herbarium specimens seen on these atypical specimens, something which the authors do note.

Phylogenetics and Molecular Taxonomy

To date, a phylogenetic assessment of *Tylosema* using genetic data has yet to be formally conducted. The only molecular sequence data available for *Tylosema* members are the chloroplast markers mentioned previously in *Bauhinia* s.l. treatments (Hao et al., 2003; Sinou et al., 2009). Molecular data have long been integrated into taxonomic practices thanks to the vast amount of information nucleotide sequences offer (Will et al., 2005). Some authors have put forward that a single short DNA sequence could be used to identify species in taxonomically challenging groups, a method known as DNA-barcoding (Herbert et al., 2003; Hollingsworth et al., 2011). The simplicity and ease of use make DNA-barcoding an attractive option for many scientists, but many taxonomists have been critical of this approach: expert taxonomists have been quick to point out that using a single or partial gene region to identify species misses the goal of taxonomy entirely, in that species identification and classification requires holistic assessment of multiple data points

from multiple specimens (Mallet and Willmot, 2003; Meyer and Paulay, 2005; Will et al., 2005). Some have argued that to imply a single gene region could identify species universally would be as problematic as implying a single morphological character could do the same (Will et al., 2005; Wheeler, 2005).

A major conceptual flaw in DNA-barcoding is the idea that a small gene region would have universal application for species identification across a diverse range of taxa. Homologous gene regions from different species (or taxa) can have diverse nucleotide substitution rates, meaning it is unlikely that a proposed universal barcode would produce resolved species level phylogenies across all taxa (Mallet and Willmot, 2003; Meyer and Paulay, 2005). In fact, the goal of finding a single gene of interest to function as a universal barcode for plants has proven fruitless and many scientists have abandoned the effort (Fazekas et al., 2008; Nock et al., 2011). The true benefit of integrating DNA data into taxonomy comes from using multiple gene sequences preferably in conjunction with morphological, ecological, and behavioural (where applicable) data to garner as much information as possible for species delimitation (Dayrat, 2005; Padial et al., 2011; Carsten et al., 2013).

The application of genetic markers in *Tylosema* could help resolve species limits in the genus. In plants markers from the nucleus and chloroplast genome (plastome) are typically utilized whereas mitochondrial DNA is unsuitable due to the typically low nucleotide substitution rates (Shaw et al., 2005; Hollingworth, 2011). As mentioned previously, no single chloroplast markers exists which can universally be used to identify plant species, leading researchers to agree that an array of plastome genes would be more appropriate for phylogenetic applications (Fazekas et al., 2007;

Nock et al., 2010; Kress and Erickson, 2007). However, chloroplast genomes seem to undergo nucleotide substitutions at a slow rate as well, though they remain less conserved than plant mitochondrial DNA (Shaw et al., 2007; CBOL et al., 2009). Fazekas et al. (2009) found that a ceiling on the number of plastome markers utilized exists, and that no significant increase in phylogenetic resolution was seen beyond using three suitable variable chloroplast markers in combination. Plastome markers are still useful to taxonomists since many primers for select markers have been designed for universal amplification across angiosperms (Fazekas et al., 2009; Hollingworth, 2011).

Due to these limitations in plastome markers, researchers have turned to nuclear genes to supplement plant phylogenetic studies. The most popular marker is the Internal Transcribed Spacers region (ITS) of the rDNA cistron, two regions which flank the transcribed 5.8S region of the rDNA array (CBOL et al, 2009; Yao et al., 2010; Chen et al., 2010; Kress et al., 2005). The rDNA cistron is transcribed to produce rRNA, molecules necessary for translating other transcribed RNA molecules into amino acid chains and later proteins (Baldwin et al., 1995). The constant demand for rRNA molecules means that rDNA cistrons are present in hundreds to thousands of tandemly repeated copies across eukaryotic genomes and can even be found spanning multiple chromosomes (Poczai and Hyvonen, 2010). It would be detrimental to organisms if swathes of these rDNA cistrons were to evolve point mutations independently of each other, resulting in potential change or loss of function in rRNA molecules. rDNA cistrons thus undergo a process called concerted evolution, whereby all cistron regions are homogenized through unequal crossing over during recombination events (Arnheim et al., 1980; Dover, 1994; Naidoo et al.,

2013). Concerted evolution thus ensures rDNA cistron arrays are homogeneous both across the genome and between individuals of the same species (Dover, 1994).

Despite this homogenization, the ITS regions of the cistron do accumulate mutations at a stable enough rate that they can be used to infer well resolved phylogenies (Hollingsworth, 2011). ITS regions do not actually form part of the final rRNA molecule following transcription, meaning mutations in nucleotide sequences may accumulate, but because these regions serve a structural function in rRNA transcription their nucleotide substitution rates remain low enough that mutations do not accumulate excessively (Baldwin et al., 1995). Plant ITS regions are thus more informative than chloroplast markers and their inclusion in chloroplast marker datasets often improves resolutions compared to phylogenies where only plastome markers are utilized (CBOL et al., 2009). Primers designed to amplify both ITS regions and the 5.8S region between them are available and have been shown to amplify universally in angiosperms (White et al., 1990). The ITS region is an attractive marker for plant taxonomists and its inclusion in species identification has been strongly recommended (Hollingsworth, 2011).

The ITS marker is, however, not without its caveats as addressed by numerous authors (Alvaraz and Wendel, 2003; Feliner and Rossello, 2007; Poczai and Hyvonen, 2010). The main concern with ITS markers in plant phylogenetics centers around how concerted evolution of the rDNA cistron could be stymied in plants due to a number of factors. These include frequent hybridization and reticulation events between plant species, whole genome duplication or polyploidization, and chromosomal duplication events (Alvaraz and Wendel, 2003).

Plants are notorious for hybridization and a large proportion of plant species carry genomic remnants of historical polyploidy events (Soltis and Soltis, 2009). This is relevant to *Tylosema* and its ploidy levels will be discussed later. Concerted evolution is a gradual process that is characterized by different stages moving towards complete homogenization of the rDNA cistron, meaning rDNA cistrons of plants with multiple genomes or chromosomes are more likely to be at an intermediate stage of homogenization (Buckler et al., 1997). If the ITS region across multiple genomes are not homogenized, an array of heterogeneous ITS sequences could exist between within or between individuals of the same species (Alvaraz and Wendel, 2003). Additionally, the accumulation of mutations in the rDNA cistron regions of polyploid species could outpace concerted evolution resulting in non-functioning ITS paralogues being amplified with functioning orthologues in PCR reactions (Poczai and Hyvonen, 2010). These potentialities, especially the presence of non-homologous paralogues, can drastically reduce the phylogenetically informative nature of the ITS region for plants (Alvares and Wendel, 2003; Feliner and Rossello, 2007).

Despite the drawbacks, ITS remains a useful genetic marker for phylogenetic studies in plants. The region is of manageable length (~700bp), is biparentally inherited, and exists in high copy numbers for ease of amplification (Alvarez and Wendel, 2003). Rather than abandoning the ITS marker, researchers have advised a cautionary approach for its application. Feliner and Rossello (2007) explain that paralogous ITS pseudogene can be differentiated from functional ITS sequences either through redesigning ITS primers or cloning PCR products that have visible length variation. It has even been demonstrated that ITS pseudogenes can be used to infer well resolved phylogenies in plant groups where functioning ITS regions are less informative (Razafimandimbison et al., 2004; Besnard et al., 2007). The weaknesses of ITS may be well

addressed but at present, few nuclear markers are universally available to plant taxonomists that are as taxonomically informative, and it is not entirely clear that alternative low copy nuclear markers would be without paralogues of their own (Feliner and Rossello, 2007).

If molecular phylogenetics is to provide accurate means of species discovery, then the incorporation of as many unlinked gene regions as possible remains the inevitable goal. This means scouring the nuclear genomes of plants for phylogenetically informative regions that can help delimit species and supplement the usage of ITS and chloroplast markers available (Hillis, 1995). The difficulty with using these low copy nuclear genes (LCNG) lies with how universally they could be amplified across taxa and the investment cost needed to find them (Sang, 2002). As mentioned above, paralogues may also be present for LCNGs, contributing to potential homoplasy in trees (Sang, 2002; Feliner and Rossello, 2007). There is also the issue of incomplete lineage sorting, whereby orthologous genes between closely related species retain ancestral polymorphisms, leading to conflicting phylogenies inferred from different markers (Pamilo and New, 1988; Maddison, 1997; Nichols, 2001). The segregation time for nuclear alleles is theoretically four times as long as those from chloroplast or mitochondrial genomes since the latter two are haploid and uniparentally inherited (Moore et al., 1995) making LCNG genes more susceptible to incomplete lineage sorting than genes from plastomes (Sang, 2002). Nonetheless, chloroplast and ITS markers have limitations which can only be overcome if investment into the development of LCNG markers is made (Sang, 2002; Alvarez and Wendel, 2003; Maddison and Knowles, 2006).

A number of LCNG markers have been developed but have rarely been tested outside of the small set of species for which they were designed. Babineau et al. (2013) tested two sets of LCNG developed by Choi et al. (2006) and Li et al. (2008). Li et al. (2008) developed what they called a conserved orthologue set (COS) of single or low copy intron spanning gene markers that were present across a wide range of angiosperm families. Choi et al. (2006) developed a large set of intron spanning nuclear markers specifically for phylogenetic use in the legume family. The benefit of having LCNG markers spanning intron regions is that introns are less conserved than exon regions and can accumulate nucleotide changes at a rate fast enough to render them phylogenetically informative. However, the ideal marker would span two intron regions with an exon in between; since exons are generally well conserved and can serve as a tag for researchers to know their primers have amplified the correct marker (Strand et al., 1997; Choi et al., 2006; Li et al., 2008).

Babineau et al. (2013) tested the phylogenetic utility of a subset of the markers developed by Choi et al. (2006) and Li et al. (2008) (16 in total) in closely related species of the Caesalpinioideae subfamily. It should be noted that the authors tested these markers prior to the reclassification of legume subfamilies by the LPWG (2017), but all species tested remained within the newly classified Caesalpinioideae *sensu stricto*. Phylogenetic analyses conducted by Babineau et al. (2013) showed that four intron-spanning LCNG markers produced well resolved phylogenies at the species level for all samples tested. These markers derived from the following genes: Auxin-indpt growth promoter (AIGP), translation initiation factor 3-like protein (EIF3E), Mg-protoporphyrin IX monomethyl ester cyclase (AT103), and serine hydroxymethyltransferase (SHMT).

Each of the four markers discussed above showed a greater number of parsimony informative characters compared to a set of regularly utilized plastome markers, while being comparable to ITS. Interestingly, the markers chosen from Li et al. (2008) produced better resolved phylogenies at lower taxonomic ranks than those from Choi et al. (2006) despite the latter set being specifically designed for use in legumes. Li et al. (2008) tested their COS markers across 87 species from 67 different plant families with an amplification success rate across all markers being slightly lower but comparable to that of the *rbc*L gene, a chloroplast masker commonly used by plant taxonomists. Babineau et al. (2013) postulate that studies testing LCNG markers across a wide range of distantly related taxa could result in a better understanding of how informative said markers are.

The taxonomic issues present in *Tylosema* may be resolved or at least ameliorated with the inclusion of molecular data for species delimitation. Numerous gene phylogenies would need to be inferred and compared to ensure that correct species limits are reached (Hillis, 1995; Wendel and Doyle, 1998). Although the limitations of chloroplast and ITS markers have been explored it remains the case universal primers are already available for these regions, and that they could still be potentially informative. When considering time constraints and funding it would be illconceived to not apply these markers to *Tylosema* members before exploring the possibility of phylogenies constructed from LCNG. However, LCNG remains crucial for accurate molecular taxonomic work and numerous attempts at developing universal primers in plants for a range of genes have been conducted (Small et al., 2004). If orthologues are identified and applied correctly LCNG genes could help bring clarity to species limits in *Tylosema*.

Cytology and Polyploidy

Goldblatt and Davidse (1977) showed that *T. fassoglense* had a chromosome count of 2n=52, and is likely a tetraploid with a base chromosome number of x=13, further supporting its exclusion from *Bauhinia* which is well reported to have a base of x=14 or multiples thereof in all members studied (Goldblatt and Davidse, 1977). Goldblatt (1981) found that within the Cercideae tribe (later recognised as the Cercidoideae subfamily) *Cercis* is diploid while the remaining members (i.e. *Bauhinia* s.l.) are fundamentally polyploid in origin. Goldblatt and Davidse (1977) also found that *Piliostigma thonningii* is 2n=24, concurring with findings from Mangenot and Magenot (1962) though conflicting with the 2n=26 found by Turner and Fearing (1959). Goldblatt and Davidse elaborate that *P. thonningii* differs from other *Piliostigma* species which are 2n=28, and perhaps more closely related to *Bauhinia* s.s.

Monaghan (1995) analysed isozyme banding patterns in *T. esculenum* and found that the species is a likely autotetraploid, not an allotetraploid. Autopolyploidy occurs when a genome duplication event takes place within a species; 7allopolyploidy occurs if hybridization between different species takes place followed by a genome doubling event (Barker et al., 2016). Monaghan and Hallorand (1996) used randomly amplified polymorphic DNA markers (RAPD) to determine genetic diversity across three *T. esculentum* populations in Botswana. Genetic diversity across the populations was high though 85% of the variation was found within populations, rather than between. Monaghan and Hallorand (1996) explain that this lack of genetic differentiation between populations despite the wide geographic separation further supports *T. esculentum*'s tetraploidy,

since alleles in polyploid populations would become fixed at a slower rate compared to those of diploid populations.

Takundwa et al. (2012) determined the chromosome number in *T. esculentum* in comparison with *Pisum sativum* (pea). *T. esculentum* was found to have 22 haploid chromosomes (2n=44) compared to *P. sativum* with 7 (2n=14), confirming the findings of Murtaza et al., (2005) for the latter. Oddly, Cullis et al. (2019) states that Takundwa et al., (2012) found that *T. escuentum* is hexaploid with a chromosome count of 2n=42, since the base chromosome number for legumes is n=7 as demonstrated in *P. sativum* (Goldblatt 1981; Doyle 2012; Stai et al., 2019). This is untrue as Takundwa et al., (2012) simply states that previous studies (unnamed) found that *T. esculentum* had a chromosome count between 42 and 50. In fact, Takundwa et al. (2012) state that *T. esculentum* is likely tetraploid given evidence of tetraploidy in *T. fassoglense*. The authors then suggest that flow cytometry analysis would be required to confirm ploidy for the entire genus. The polyploidy of *Tylosema* is largely unclear given the studies previously conducted.

It is likely that *Tylosema* has a complex cytological history which may be the case for other genera recognised within *Bauhinia* s.l. Within the context of the Cercidoideae Stai et al. (2019) confirmed that the genus *Cercis* is of diploid origin (n=7), compared to the genera comprising *Bauhinia* s.l., which is unique among legume genera. Stai et al. (2019) also found that chromosome numbers for *Bauhinia* s.l. members (*Piliostigma*, *Griffonia*, and *Adenolobus*) was n=14, further demonstrating that base chromosome numbers in members vary greatly. This variation of chromosome counts may not be unique to members of Cercidoideae as genera within the subfamily Caesalpinioideae deviated greatly from the standard n=12,13, or 14 suggesting that further chromosomal fusions, reductions, or possible ploidy increases took place (Stai et al., 2019).

Polyploidy, or whole-genome duplication (WGD) as it is often called in literature, has been recognised as playing a crucial role in the evolution of land plants, especially angiosperms (Adams and Wendel, 2005; Soltis et al., 2009; Jiao et al., 2011; Soltis et al., 2015; Van de Peer et al., 2017; Landis et al., 2018). Historical polyploidy events have been demonstrated in a number of recognised angiosperm clades (Jiao et al., 2011), and subsequent to these polyploidy events research has shown an explosion in the number and diversity of taxa such in the Asteraceae, Brassicaceae, and of course the legume family, Fabaceae (Schranz et al., 2012; Soltis and Soltis, 2016). In the case of *Tylosema* evidence of polyploidy is present but the type of polyploidy has not been examined, with the exception of the aforementioned likely autopolyploid origins in *T. esculentum* (Monaghan, 1995). Determining cytology, ploidy, and polyploidy categories in *Tylosema* species in future research would aid in clarifying the origins of the genus and its members.

Genetic Diversity

Understanding genetic diversity in plants is foundational to addressing a wide array of problems biologists tackle in fields ranging from agriculture and plant breeding, to ecological issues and conservation (Mondini et al., 2009). From the end of the previous century to the early 2000s a vast array of molecular techniques have been developed to measure genetic diversity in crops and wild plants, replacing conventional methods of the past based on phenotypic and morphological traits as a proxy for genetic variability (Mondini et al., 2009; Kalia et al., 2011). The most noteworthy techniques developed at the time included restriction fragment length polymorphisms (RFLP),

amplified fragment length polymorphisms (AFLP) (Vos et al., 1995), the aforementioned RAPD (Waugh and Powell, 1992) markers, and finally microsatellite markers or simple sequence repeats (SSR). With so many methods being developed around the same time it was inevitable that numerous authors would compare their utilities, strengths, and weaknesses (Pejic et al., 1998; Powell et al., 1996; Nybom, 2004; Russell et al., 1997).

The four methods can be placed in one of two categories: hybridization based and PCR based. Of the four addressed above, only RFLP is hybridization based. RFLP, one of the earliest molecular methods developed, involves a combination of restriction endonuclease enzymes and the hybridization innovation developed by Southern (1975). RFLP entails cleaving DNA into fragments using endonuclease enzymes, bacterial enzymes capable of cutting DNA sequences at known points to produce fragments of known length (Nathans and Smith, 1975; Botstein et al., 1980). Mutations occurring in the fragment or at the enzyme cleavage site can result in length variations or new cleavage sites between different individuals or between restriction fragments from homologous chromosomes (Botstein et al., 1980).

Specific polymorphic fragments can be detected by running restriction enzyme digested DNA on an agarose gel, hybridizing said gel with radioactive chemical probes, and visualizing the labelled fragments using X-Ray film (Southern, 1975). The RFLP method sports many advantages: markers tend to detect high variability, are highly reproducible, and are codominantly inherited meaning they can detect heterozygotes from homozygotes (Pejic et al., 1998; Russell et al., 1997). Unfortunately, the method is complex, time consuming, requires large amounts of DNA, radioactive labels, and specialized lab equipment compared to PCR based methods, rendering

RFLP markers unfavourable in most labs since the advent of PCR marker systems (Mondini et al., 2009; Nybom et al., 2004).

The shortcomings of RFLP markers and their inability to be used in high throughput studies were addressed by the development of the RAPD method by Williams et al., (1990). Essentially, RAPD implements random single primers (not primer pairs) of 10 bases in PCR amplification, resulting in random sequences of varying length across the genome being produced (Williams et al., 1990; Waugh and Powell, 1992). The primer length is essential for the approach, as the shorter than usual primers increase the probability of annealing to homologous sequences, despite the primers being purposefully random (Mondini et al., 2009). The PCR products can then be visualized using a standard electrophoresis system and a DNA fingerprint is produced (Nadeem et al., 2018). RAPD have a near universal application since the primers used are entirely random, but this randomness is also a notable caveat in the method as reproducibility between studies has been a challenge (Wang et al., 1996; Karp et al., 1997; Jones et al., 1997).

AFLP markers were developed by Vos et al. (1995) to provide a more robust method for genetic variability assessment given the realized weaknesses in RAPD. AFLP markers are similar to RAPD markers in that they exploit PCR technology, though they also make use of restriction enzymes like RFLP. It involves producing DNA fingerprints by digesting DNA with restriction enzymes, and then the resulting fragments have nucleotide adapters ligated to their ends to serve as binding sites for specifically designed primers. These fragments are then PCR amplified and visualized on a polyacrylamide gel to produce a DNA fingerprint (Vos et al., 1995; Blears et al., 1998). AFLP markers have many strengths: they are cheap and simple to develop, can be applied

to virtually any and all organisms, and require no initial knowledge of the DNA sequences amplified (Vos et al., 1995). The trouble with AFLP markers is their biallelic (presence/absence) nature of the bands produced, meaning genetic diversity for markers is scored on a presence/absence basis. This makes the differentiation of heterozygous from homozygous individuals unfeasible (Nybom et al., 2004).

The final markers to gain popularity come the end of the 20th century are SSRs. SSRs are short sequences comprising of 1 to 6 nucleotides motifs repeating consecutively; they are found universally in eukaryotes (Tautz and Renz, 1984). These repeat motifs can be classed as perfect, imperfect, or compound (Bull et al., 1999). Perfect repeats consist of nucleotide motifs which repeat consistently (e.g. ACTACTACT) where imperfect repeats contain nucleaotide motifs that alter in their arrangement across repeats (ACTCATACT). Compound repeats contain more than repeat motif in a single region (ATATCGCG) (Bull et al., 1999).

Since their discovery SSR markers have been used extensively to determine genetic diversity in both crops and wild plant populations (Varshney et al., 2005; Selkoe et al., 2006; Kalia et al., 2011). SSRs have gained considerable favour among plant biologists over RFLP, RAPD, and AFLP markers. This is the case despite their major drawback: SSR regions must first be found in genomic data and each marker must have primer pairs designed specifically for it, meaning an initial time-consuming investment must be made if markers aren't already available (Varshney et al., 2005). This issue is avoided in the AFLP method and by the arbitrary markers used in RAPD methods (Pejic et al., 1998, Russell et al., 1997; Nybom 2004).

Comparitive studies have consistently shown that SSR markers detect higher polymorphism per marker than any of the three rival methods. SSRs are also codominantly inherited like RFLPs meaning they're suitable for differentiating homozygotes from heterozygotes, while avoiding the complications that hybridization methods entail (Powell et al., 1996; Russell et al., 1997; Pejic et al., 1998; Nybom et al., 2004). SSRs are highly polymorphic, multiallelic, easily reproducible, abundantly spread across genomes, and easily amenable to high throughput assessments (Kalia et al., 2011). As a result, many biologists feel the advantages SSR markers have over rival marker systems are numerous, and thus outweigh their aforementioned weaknesses. In fact, the rapid development of whole genome sequencing and other technologies have made the weakness of SSR discovery nearly negligible (these developments will be discussed in detail in a later section).

What makes SSRs so highly variable is their unique evolution; how they come into existence and subsequently expand/contract in length repeats. The origins of SSRs are subject to debate, though generally it is accepted that they begin as proto-microsatellite, series of a few repeat motifs (3 or 4) not necessarily tandemly arranged (Buschiazzo and Gemmell, 2006). A number of hypotheses have been proposed to explain how these proto-microsatellites expand and develop into recognizable SSRs, but the principal mechanism accepted in literature is DNA strand slippage during replication (Levinson et al., 1987).

DNA polymerase III slips from the template strand at the repeat motif and is joined again, however with a loop formed in either the template or nascent strand. The resulting loop causes an out of frame alignment of the two reunite strands, and depending on which strand the loop occurs, an extension (nascent strand) or contraction (template strand) of the repeat motif is produced

(Brohede and Ellegren, 1999; Ellegren, 2004; Wang et al., 2009; Kalia et al., 2011). Usually, replication errors like these are removed by exonuclease enzymes through a process of mismatch repair, but loop structures can hide these errors from repair mechanisms (Moore et al., 1999; Ellegran, 2004). This mechanism is supported by the positive correlation between mutations occurring in repair enzymes and SSR length stability (Li et al., 2002). As a result, SSR regions have a significantly higher mutation rate than point mutations occurring elsewhere in the genome (Schlotterer, 2004)

SSR regions have also been utilized in inter-simple sequence repeat (ISSR) fingerprinting methods, whereby instead of designing primers at the conserved flanking regions of SSRs, the SSR repeat regions are used as templates for primers and the internal sequences between identical SSRs are amplified (Gupta et al., 1994; Godwin et al., 1997). These amplified regions are then visualized on a gel (either polyacrylamide or agarose) and the resulting banding pattern is used to determine genetic variability in samples (Godwin et al., 1997; Reddy et al., 2002). ISSR methods are PCR based and have been shown to be more variable than RAPD and RFLP methods (Godwin et al., 1997; Taylor and Barker, 2012). Unlike SSR markers, ISSR does not require prior genomic information, a benefit they share with AFLP markers, which makes them a cheap alternative (Reddy et al., 2002). However, ISSR markers also share similar caveats with AFLPs in that they are dominantly inherited, meaning the detection of heterozygotes is ruled out (Mariette et al., 2002; Bentley et al., 2015). A further limitation in this technique is that since genetic diversity is determined by gel visualization, bands of similar sizes may be of different origins, misleading diversity assessments (Sanchez et al., 1996).

SSR marker usage in plants

The popularity of SSRs over competing markers is not without merit, and as such plant biologists from a wide range of disciplines have adopted them for a number of different studies. Initially, SSRs were commonly used in the agricultural sector, due to the requirement of prior genomic information in the species of interest in order to develop adequately sized primer sets. This genomic data was either not present or would prove costly to obtain in wild plants, and so SSRs were less commonly applied outside of commercial crops (Mondini et al., 2009). However, the versatility of SSR markers is extensive and they can be used to address a wide array of biological questions that interest ecologists and conservationists (Selkoe and Toone, 2006; Guichoux et al., 2011).

Technology has advanced at a canter over the past few decades. The development of next generation sequencing technologies, refinements in co-amplification of multiple markers in a single PCR reaction (multiplexing), and the development of fluorescent dyes to differentiate loci of similar sizes in multiplex reactions have made the pipeline from microsatellite development to usage more streamlined than before (Butler, 2005; Santana et al., 2009; Guichoux et al., 2011; Gardener et al., 2011). A number of computer programs have been designed to filter out repeat motifs from genomic or DNA sequence databases, allowing more traditional SSR discovery techniques to be replaced by *in silico* marker mining (Sharma et al., 2007). A comparison of the myriad of SSR detection programs would be beyond the scope of this review, but the general process these finding tools implement can be summarized into three broad categories: those that bias perfect tandem repeat motifs; those which utilize a two tiered search function, firstly building a preliminary set of repeat motifs which could all potentially be microsatelites, only to further filter

verifiable microsatellites from this dataset; and finally an unrefined method whereby sequence data is aligned to a set of desired motifs, and any successful alignments are pulled out to form a dataset (Sharma et al., 2007).

One program that offers a good amount of flexibility, where search functions can be customized for perfect or imperfect motifs of various repeat lengths is MSATCOMMANDER (Faircloth, 2008). What makes MSATCOMMANDER advantageous over alternative programs is that it can detect SSR repeat motifs while utilizing PRIMER3 as a built-in feature to design primers at the flanking regions. MSATCOMMANDER also allows primers to be designed with 5'-tail for M13 labelling. Though drawbacks to the program are present (SSR statistical data are not generated, requires Python to run, and only FASTA format input files are accepted) the simple graphic user interface (GUI) and versatility MSATCOMMANDER offers make it an attractive choice for microsatellite detection (Faircloth 2008; Sharma et al., 2007; Hodel et al., 2016).

Specialized lab equipment and extensive funding are no longer limitations to many scientists looking to use SSRs in studying natural plant populations. DNA extraction and marker isolation can also be outsourced to a number of companies and university facilities for a minimal cost (Selkoe and Toone, 2006). The number of biological questions SSR markers can be used to answer is extensive. For example, plants are stationary organisms and only disperse via pollen and seed. Traditional methods of tracking plant dispersal have been challenging, but microsatellites allow scientists to overcome these challenges by tracking parentage and relatedness within and between plant populations (Ashley, 2010). SSR markers can be used to investigate initial domestication events in the wild counterparts of important plant crops (Liu et al., 2019). SSRs can also be used

to determine if there is a genetic basis for differences in flowering times within and between different populations of the same species (Selkoen and Toone, 2006).

The impact SSR markers have had on conservation ecology is massive as the information they provide on population structure, genetic diversity, gene flow, and migration events has proven invaluable for the conservation status of many species (Chase et al., 1996; Hendrick, 2001; Selkoe and Toone, 2006; Allendorf et al., 2010). Although a number of new genetic methods and techniques have been developed since the advent of genome sequencing, such as genotyping by sequencing (GBS) and restriction site associated DNA sequencing (RAD-Seq), SSRs remain a relevant option for researchers since they're generally cheaper than newer methods while remaining comparatively informative provided sample sizes are large enough (Hodel et al., 2016).

Genetic Diversity in *Tylosema*

The common trend in the literature shows a bias in research interest towards *T. esculentum* over other *Tylosema* species; this remains the case in genetic studies as only *T. esculentum* has been assessed, though sparsely. As mentioned previously, Monaghan and Hollaran (1996) used RAPD markers to assess genetic diversity across three *T. esculentum* populations in Botswana. Their findings determined that genetic diversity was higher within populations than it was between, and that genetic diversity across the species was high overall. Nepolo et al., (2009) used the Fast Isolation by AFLP of Sequences Containing Repeats (FIASCO) method to develop microsatellite markers, given the advantages SSRs have over RAPD markers, but did not utilize these markers in any genetic diversity study.

Takundwa et al., (2010) developed SSR markers in *T. esculentum* using the same FIASCO method. From the 80 markers developed, polymorphic markers (MARA001, MARA065, MARA068, MARA077) were chosen as a subset to assess genetic diversity from individual plants across 11 Namibia populations. Agarose gel images for markers MARA001 and MARA068 showed that only two bands (alleles) were amplified at those loci. For markers MARA065 and MARA077, up to 6 alleles could be seen on gel images. The presence of multiple alleles was used to confirm high levels of genetic variability in *T. esculentum* populations, which can also be an indication of polyploidy. Takundwa et al. (2010) contrast their findings with Monaghan and Hollaran (1996) stating that diversity both between and within *T. esculentum* populations was likely high.

Chimwamurombe (2010) also used an SSR marker set consisting of 12 primer pairs to assess genetic diversity in 20 plants from a single *T. esculentum* subpopulation in Omitara, Namibia. This marker set was developed by Dr Chris Cullis, Case Western Reserve University, Ohio, and Chimwamurombe (2010) seems to be the first published use of the set. Of the 12 markers offered, only 4 amplified successfully. Similarly to Takundwa et al. (2009), variability was assessed in terms of alleles present/absent in individual samples, but the results from this study showed that intrapopulation diversity was lower than previous estimates.

It would still prove fruitful to invest effort and money into the development of additional SSR markers for *T. esculentum*, and importantly to apply these markers to other *Tylosema* species. It seems likely that *T. fassoglense* and *T. angolens*e are closely related species (Castro et al., 2005). *Tylosema* plants are long lived and reproduce seemingly exclusively through outcrossing, traits

which correlate positively with transferability of SSR markers across species of the same genus (Barbara et al., 2006). The move to conserve genetic material in wild plants has not only strengthened conversation planning, but also provides scientists with a wealth of genetic information and that can utilize in modifying crops or domesticating native plants (Nepolo et al., 2009; Allendorf et al., 2010). *Tylosema* species have been neglected by researchers, and a push to gain more genetic information will help conserve them in the wild for the gain of human use (Nepolo et al., 2009).

Aims

This study aims to address the two issues related to southern Africa *Tylosema* with regard to their potential use as a future crop plant in Africa and globally:

1. Resolve the taxonomic uncertainty surrounding *Tylosema sp*. in southern Africa by using molecular phylogenetic techniques as outlined in the above literature review.
2. Identify SSR regions in *T. esculentum*, develop primer pairs for said SSRs, and test for amplification and polymorphisms within these regions for both *T. esculentum* and *T. fassoglense* populations.

Chapter 2

Phylogenetic analysis of southern African *Tylosema sp.*

Introduction

Species limits in *Tylosema* are poorly understood. *T. esculentum* is a favoured species for domestication (Cullis and Kunert, 2017); however, it spans two distinct environments in the wild: the Kalahari Desert and the Highveld grasslands of South Africa (Coetzer and Ross, 1976). These two biomes differ greatly in average temperature, rainfall, and interspecies competition. The variation in ecological pressures between the two biomes could mean that *T. esculentum* is in fact two different species, or subspecies, occupying each geographical area. *T. angolense* has recently been described by Castro et al. (2006) and is endemic to southern Angola. However, certain *Tylosema* specimens in southern Angola (and across central and southern Africa) are morphological intermediates between *T. angolense* and the widely distributed *T. fassoglense*. *T. fassoglense* is noted for displaying a wide range of morphological variability in vegetative characters across its distribution (Coetzer and Ross, 1976; Coetzer et al., 2011), meaning *T. angolense* could conceivably be an ecotype of *T. fassoglense* and not a distinct species in itself. This lack of phylogenetic diagnosability across the three *Tylosema* species would slow down any domestication effort. The agricultural potential of *Tylosema* species is huge considering their nutritional value, and an accurate account of species limits in the genus would ensure this potential

is reached. No molecular phylogeny exists in literature for *Tylosema* despite the fact that morphological data has proven too inconsistent to accurately delimit species. The aim of this study is to produce a molecular phylogeny for the three *Tylosema* species in question using chloroplast and nuclear DNA. This phylogenetic analysis is intended to better define the species limits between *T. angolense* and *T. fassoglense* and would clarify species limits in *T. esculentum* across its geographical and ecological range.

Method

Sample collection

Localities of *T. fassoglense* and *T. esculentum* in South Africa were found using specimens from the University of Pretoria's H.G.W.J. Schweickerdt Herbarium (PRU), and online databases i.e. the Global Biodiversity Information Facility (GBIF) and South African National Biodiversity Institute's (SANBI) new Plants of Southern Africa database (POSA). For the sake of efficacy only specimens with available GPS points, and/or detailed locality descriptions were used to map a sampling route. Six *T. esculentum* localities were sampled from: three across the North-West province and three in Gauteng, though one locality is on the University of Pretoria's Experimental farm. *T. esculentum* plants on the experimental farm were relocated from an unknown wild population, though they are likely of South African origin. Eleven *T. fassoglense* localities were identified: Nine localities across the Limpopo province and two in Mpumalanga. At each locality for both species, leaf samples were collected in silica gel for rapid desiccation, keeping the leaf

material dry for DNA extraction. An herbarium voucher was collected and pressed at each locality for the H.G.W.J. Schweickerdt Herbarium. Seeds were collected when present and later planted on the University of Pretoria experimental farm. Dr Juan Vorster (Department of Plant and Soil Sciences, University of Pretoria), collected *T. esculentum* seeds in Botswana and Namibia, but unfortunately specific locality information was not documented. Mr Arnold Frisby (Department of Plant and Soil Sciences, University of Pretoria) collected seed and leaf material of *T. angolense* in Angola. These seeds were germinated at the University of Pretoria Experimental Farm and the first leaves to emerge were collected into silica gel for DNA extraction. Additionally, herbarium leaf material was provided by Mr John Burrows from Buffelskloof Private Nature Reserve's Herbarium (BNRH). These samples were individual leaves of *T. fassoglense* from localities in Penge, Limpopo, and Maputo, Mozambique.

DNA extraction

DNA was extracted using a simplified CTAB protocol from Doyle and Doyle (1988). A 1cm square of leaf material for each sample was fragmented into a mortar and ground in 1ml of cetyltrimethylammonium bromide (CTAB), to which 0.2µl of β-mercaptoethanol was added. The solution was poured into a 1.5ml eppendorf tube which was then incubated in a 60°C heating block for 30 minutes. Following incubation, 500µl of chloroform: isoamyl alcohol (24:1) was added to each sample. Samples were then centrifuged for 1 minute at 13000 revolutions per minute (rpm). Once samples were spun, 600µl of the top supernatant layer was removed and pipetted into a new eppendorf tube. 400µ of isopropanol was added to the supernatant which was then placed into a -20°C fridge overnight. The following day, samples were centrifuged for 10 minutes at 13000 rpm and the supernatant was removed, leaving a pellet of DNA. The DNA was washed with 750µl of

70% ethanol. The ethanol was then poured off and pellets were left to fully dry in a 4°C fridge overnight. Once fully dried, the DNA was resuspended in 300µl of double distilled water and was ready for use.

Amplification and Sequencing of Barcoding Markers (*trn*L-F; *psb*A-*trn*H; ITS)

Three universal phylogenetic markers, or barcoding markers (Herbert et al., 2003), were chosen for this study: *trn*L-F intron and spacer, *psb*A-*trn*H intergenic spacer, and the internal transcribed spacer (ITS). The first two genes represented chloroplast markers, where ITS represented nuclear data. Primers for the *trn*L intron and spacer were obtained from Taberlet et al. (1991). The forward primer for *psb*A-*trn*H was obtained from Sang et al. (1997) and the reverse was obtained from Tate and Simpson (2003). ITS forward and reverse primers were obtained from White et al. (1990). In the case of the *trn*L-F marker, the intron and spacer regions were amplified and sequenced separately using primer pairs "C" and "D" and "E" and "F" respectively. The *trn*L-F intron and spacer is known to typically exceed lengths of up to 1000 bp in angiosperms (Taberlet et al., 1991) and the use of internal primers ensured a greater chance of sequence success.

PCR reactions for each marker all contained 10µl of MyTaq Reaction Buffer (comprising final concentrations of 1mM dNTPs and 3mM of $MgCl_2$), 0.5µl of MyTaq (2.5 units) DNA polymerase, 1µl of forward primer and 1µl of reverse primer (final concentration of 0.2µM each), and 5µl of DNA (approximately 100-200ng of genomic DNA). For specimens obtained from the Buffelskloof herbarium, 7µl of DNA was added to the reaction as well as 1µl of 50mM $MgCl_2$ (with a concentration of 1mM making a final concentration of 4mM) to increase taq efficiency and ensure greater amplification success. Distilled water was added to each reaction up to make up a final volume of 50µL. PCR conditions for all chloroplast markers were the same: an initial denaturation

step for 3 minutes at 96°C; followed by 30 cycles of denaturation for 1 minute at 96°C, annealing for 1 minute at 48°, and elongation for 1 minute at 72°C; followed by a final elongation step for 7 minutes at 72°C. PCR conditions for ITS sequences began with a denaturation step for 1 minute at 96°C; then 30 cycles of denaturation for 1 minute at 96°C, annealing for 1 minute at 48°C, and elongation for 3 minute at 72°C; followed by a final elongation step for 7 minutes at 72°C . Herbarium specimens had their cycles increased to 35 in chloroplast and ITS markers. All PCR reactions were conducted using the 2720 Thermal Cycler and the GeneAmp* PCR System 2700, both from Applied Biosystems. Amplified products were run on a 1% borax gel with ethidium bromide using a gel electrophoresis tank. Gels were run for 30 minutes at 120V and then visualized under UV light using a Gel DocTM XRay+.

Once markers were successfully amplified, they were cleaned using the MSB® Spin PCRapace kit from Stratec. PCR samples were mixed with a 250µl binding buffer and incubated at room temperature for 1 minute. Samples were then centrifuged through a filter column for 4 minutes at 13200 rpm. The flow-through was discarded and 25µl of elution buffer was added to the column which was further centrifuged for 1 minute at 11000 rpm resulting in 25µl of cleaned PCR product as the flow-through. Cleaned PCR products were then checked on a 1% agarose gel. Clean DNA was then used for cycle sequencing.

For cycle sequencing separate reactions were set up for forward and reverse primers of each marker. For each sample a reaction contained 5µl of clean PCR DNA, 1µl of Big Dye, 1.5µl of sequencing buffer, and 1µl of the forward/reverse primer of each marker. Distilled water was then added to make up a final volume of 10µl in every reaction. The amount of DNA was reduced in *psb*A and *trn*H primer reactions from 5µl to 1µl to reduce overloading the sequencing machinery at the University of Pretoria's bioinformatic facility. Through sequencing it was found that *psb*A-

*trn*H intergenic spacer is comparatively A/T rich in *Tylosema*, which resulted in frequent DNA polymerase slippage during the sequencing process and low-quality sequencing results. Lowering the DNA concentration per reaction for the *psb*A-*trn*H marker, which solved the issue, resulting in good quality sequences.

The PCR conditions for every cycle sequencing reaction were as follows: an initial denaturation step at temperature of 96°C for 1 minute, followed by 25 cycles of denaturation at 96°C for 10 seconds, an annealing step at 50°C for 5 seconds, and an extension period at 60°C for 4 minutes. No final extension period following the 25 cycles was required. Once cycle sequencing was complete, the PRC products were then precipitated. 2µl of ethylenediaminetetraacetic acid (EDTA) and 30µl of 100% ethanol were added to cycle sequencing products and left to incubate at room temperature for an hour. The solution was then centrifuged at 13000 rpm for 20 minutes. The supernatant was pipetted out and 30µl of 70% alcohol was added to the pellet. The solution was then centrifuged for another 15 minutes at 13200 rpm. The supernatant was removed again, and the DNA pellet remained behind to air dry overnight. Once dry, samples were then taken to the University of Pretoria's DNA sequencing facility for Sanger sequencing.

Sequences were retrieved and were uploaded to the program Sequencher 4.5 (GeneCodesCorp, 2005) for editing. Sequences from the forward and the reverse primers for each marker, in each individual specimen, were combined to form contigs. Contig assembly parameters were set for 'dirty data' with a minimum of 60% match between forward and reverse sequences, and a minimum sequence overlap of 14 bp. Once contigs were assembled sequences were manually edited to ensure that nucleotide bases were scored correctly according to the IUPAC code nomenclature. Edited contigs were then exported from Sequencher as text files to be aligned for phylogenetic analysis.

Low Copy Nuclear Gene selection, amplification, and sequencing

The two chloroplast markers and ITS are all universally available barcoding markers that have standardly been utilized in plant phylogenetic studies over the past two decades; the drawbacks of such markers have been discussed in the previous chapter. These drawbacks motivated the addition of a LCNG marker to the analysis. Babineau et al., (2013) compiled a set of LCNG markers designed by Choi et al., (2006) and Li et al., (2008) and tested their phylogenetic utility in the legume subfamily Caesalpinioideae. We selected eight of the markers Babineau et al. (2013) tested that were variable enough to assess legume taxa at the species level (see Table 1). Of the markers we selected, Babineau et al. (2013) identified three markers with duplication events in the taxa they tested; ATCP, AROB, and CALTL. Since Babineau et al. (2013) infer that these duplications originated in the common ancestor of Caesalpinioideae s.l. and that these three markers showed comparably high parsimony informative characters, we decided to test their feasibility in *Tylosema* for our own study.

Table 1 Low Copy Nuclear Gene (LCNG), tested in Caesalpinioid legumes.

Name	Putative function	Forward sequence	Reverse sequence	Author
AT103	Mg-protoporphyrin IX monomethyl ester cyclase	CTTCAAGCCMAAGTTCATCTTCTA	TTGGCAATCATTGAGGTACATNGTMACATA	**Li et al. (2008)**
AIGP	Auxin-indpt growth promoter	CTGATAGGGCCAGGAGGCAGGGAAGA	GTTTTTTAGCATTTGGACGAATGGTTGGT	**Choi et al. (2006)**
SHMT	Serine hydroxymethytranferase	ACCACAACTCACAAGTCACTTC	TTGCTGAGAACCTGCTCTTGGTATG	**Choi et al. (2006)**
EIF3E	Translation initiation factor 3-like protein	TTTGAATGTGGCAACTAYTCTRGTGCTGC	ACCTCTTCACACTCYYTCATCTT	**Li et al. (2008)**
AroB	3-dehydroquinate synthase	GCATTCTACCAARCWCARTGTGT	GCTTTGTTTTCACATGAWCKCTTDATAGCA	**Li et al. (2008)**
CALTL	Calreticulin	GTGGAAGGCACCATTGATTGACAAC	TCTTCTTCTCAGCCTCTTCAAATGC	**Choi et al. (2006)**
ATCP	Aquaporin-like protein	AACCAATTGGTATTGCAGCTCAGAGCCA	TTCCTTGCCAAGAACAAACCGAATGTCA	**Choi et al. (2006)**
SQD1	Sulfite: UDP-glucose sulfotransferase	CTTGGGACSATGGGTGARTATGG	CCWACAGCAGCYTGMACACAGAACC	**Li et al. (2008)**

Due to financial and time-related constraints, each marker was tested for amplification and sequencing in four *Tylosema* specimens representing each of the species in question. In the case of *T. esculentum*, two specimens were chosen to cover the Kalahari and Highveld regions respectively. This assessment was conducted to determine whether or not the markers chosen would amplify in *Tylosema*, and from those that did which would likely be the most cost effective for phylogenetic use. PCR reactions for each LCNG marker contained the same concentration of reagents and DNA as the markers mentioned above. PCR conditions for markers developed by Li et al., (2008) were as follows: an initial denaturation period of 95℃ for 2 minutes, followed by 35 cycles of denaturation at 94℃ for 40 seconds, 55℃ for 30 seconds of annealing, and an extension

period of 72℃ for 1 minute and 30 seconds; followed by a final extension period of 72℃ for 5 minutes. Conditions for markers developed by Choi et al., (2006) were set at: 94℃ for 3 minutes of initial denaturation followed by 35 cycles of denaturation at 94℃ for 30 seconds, an annealing period of 53℃ for 30 seconds, and an extension period at 72℃ for 2 minutes; followed by a final extension period at 72℃ for 5 minutes.

Visualizing PCR products for each marker showed that AICP, AROB, and CALTL could be discarded from the study. No amplification took place for AICP and AROB under the conditions used, and two bands were present in gel visualization for CALTL. This indicates a duplication event in this gene, echoing the findings of Babineau et al., (2013). This indicated that the event of gene duplication may have occurred earlier than in the common ancestor of Caesalpinioideae s.l. as Babineau et al. (2013) found; or that multiple, independent duplication events occurred in CALTL across the legume family. The remaining five markers all amplified successfully and were subsequently sequenced. It should be noted that in each marker the *T. esculentum* specimen originating from Botswana failed to amplify. Lack of amplification in this specimen was also seen in our chloroplast markers when tested, confirming that there may be a problem with the DNA quality of this extraction. For the remaining specimen, PCR product cleanup, cycle sequencing, and precipitation were conducted using the same protocols as mentioned prior for the chloroplast and ITS markers.

Sequencing results for the remaining five LCNG markers showed that EIF3E, AIGP, AT103 were unfeasible. Chromatogram peaks were unreadable for these three markers, with multiple overlapping nucleotide peaks throughout the sequences. This indicates that either duplication events have taken place in these gene regions or that nonspecific primer binding had occurred during PCR. SHMT and SQD1, the remaining two markers, were sequenced successfully.

However, SHMT showed an insert/deletion (indel) event only in the *T. esculentum* specimen sequenced. Unlike markers EIF3E, AIGP, and AT103 genes which had unreadable chromatograms in all specimens, the SHMT sequence in the *T. esculentum* sequence showed clean nucleotide peaks in the forward and reverse primer strands up till a point. The forward primer showed single peaks from the beginning of the sequence until around the 500 bp region, therein the sequence contained two overlapping peaks at each base pair. Likewise, the reverse strand showed double peaks at each base from the beginning of the sequence until displaying single peaks at the 500th bp, wherein single peaks were then seen (see Figure 3). Due to this consistency, we hypothesized that the SHMT gene is duplicated in *T. esculentum* and the two copies differed in length due to an indel event that did not occur *T. fassoglense* or *T. angolense.*

Sequences lengths for the SQD1 marker were between 290 and 300 bp, and SHMT was approximately 800-850 bp in length. These lengths roughly correspond with results from Babineau et al., (2013), Choi et al., (2006), and Li et al., (2008). SQD1 also had the lowest parsimony informative sites score of any of the markers tested according to Babineau et al. (2013). In contrast, SHMT had a parsimony informative character score greater than that of ITS and was found to be variable enough to provide taxonomic resolution at the species level (Babineau et al., 2013). SHMT had the advantage over SQD1 in terms of length and likely informative characters, however the indel event in *T. esculentum* was an issue. SHMT was thus chosen as the LCNG for our analysis as a methodology for addressing this problem was devised.

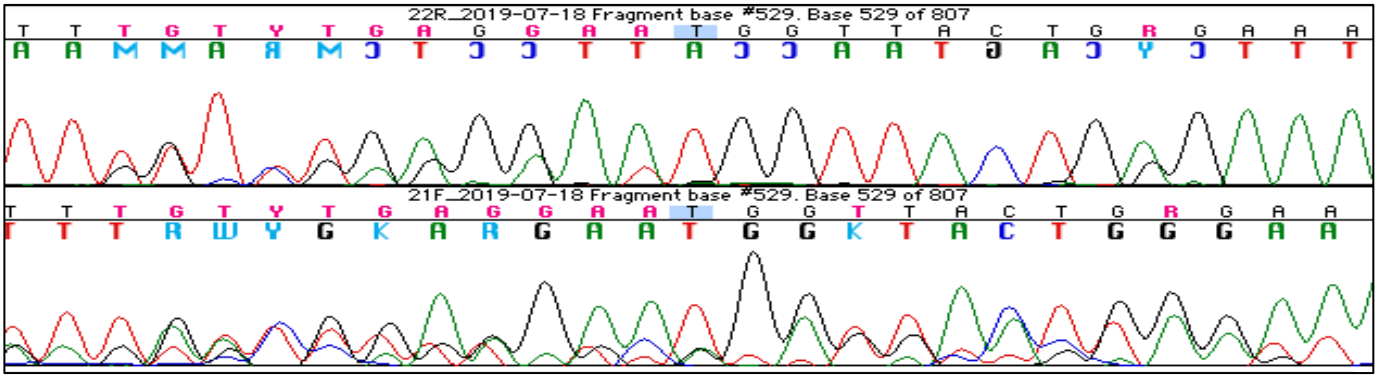

Figure 3 Chromatogram image of SHMT sequence for a T. esclentum specimen. Above: reverse strand of specimen where overlapping peaks can be seen to begin upstream (right). Below: Forward strand where overlapping nucleatides are present due to indel event.

Indel identification in SHMT

Additional *Tylosema* specimens (1 *T. angolense*, 3 *T. fassoglense*, and 3 *T. esculentum*), forming a subset of specimens sampled and extracted, were amplified for the SHMT marker and then sequenced. Sequences were edited and assembled into contigs through the same method as previously stated. This confirmed that the putative indel event had only taken place in *T. esculentum*. No *T. fassoglense* or *T. angolense* specimen showed any evidence of this indel. However, the specific indel event would need to be identified and the two distinct SHMT sequences would need to be separated from each other before the marker could be used for phylogenetic analysis. Instead of cloning *T. esculentum* PCR products to isolate SHMT paralogues, we used the free, online software 'Indelligent' (Dmitriev and Rakitov, 2008) to tease the overlapping sequences apart *in silico.* Indelligent algorithms can detect two or more sequences superimposed on top of each other from chromatogram sequence data.

Four *T. esculentum* specimens, two from South Africa, and one from Namibia and Botswana respectively, were sequenced. In each specimen, the reverse primer was chosen as a guide and coloured peaks in the chromatogram were coded according to the IUPAC codes for nucleotide nomenclature. This system provides codes for overlapping coloured peaks e.g. 'W' for an adenine and thymine base superimposed on top of each other. Coding was completed for each *T. esculentum* specimen and resulting ambiguously coded sequences were submitted to Indelligent for indel event confirmation. An 8 bp indel event was identified in each *T. esculentum* sequence, corresponding to the region wherein overlapping peaks began to appear. Indelligent provided two two sequences per *T. esculentum* specimen: one containing the 8 bp motif and one without it (see Figure 4). These sequences were labelled, indicating the presence/absence of the motif and which specimen they came from. Both sequences per specimen were then used for the SMHT alignment dataset.

Figure 4 Two sequences produced by the online software Indelligent (Dmitriev and Rakitov, 2009) using a single T. esculetum sequence, where overlapping nucleotide peaks were coded using the IUPAC ambiguity codes.

Alignment and Outgroups

Alignments for each marker were produced using the CLUSTALW function (Thompson et al., 1996) in MEGA 6 (Tamura et al., 2013). The two chloroplast markers were thus treated as a single dataset. The *trn*L-F intron and spacer were rejoined in a single alignment. The *psb*A-*trn*H intergenic spacer was aligned separately before being concatenated with the *trn*L-F alignment. ITS and SHMT markers were aligned separately and treated as independent datasets. Deciding on an outgroup for each dataset was challenging considering the taxonomic uncertainty of *Bauhinia* s.l., the clade in which *Tylosema* is situated. Because there is no clarity on which taxa would appropriately constitute a close relative of *Tylosema*, the decision was made to include as many members of Bauhinia s.l. as possible into each alignment. *Bauhinia* has been split into a number of genera spanning much of the southern hemisphere (Wunderlin et al., 1987). NCBI GenBank accessions were acquired for ITS, *trn*L-F, and *psb*A-*trn*H regions of *Bauhinia* s.l. members and utilized as outgroup taxa. Accessions obtained from GenBank included *trn*LF sequences for additional *Tylosema* specimens: two *T. fassoglense* specimens, and one *T. argentea*, *T. humifusa*, and *T. esculentum* each.

Due to the fact the SHMT is a LCNG which has not been utilized in plant phylogenetic studies as extensively as the three universal markers, no GenBank accessions were found for any *Bauhinia* s.l. species. Leaf material for *Bauhinia tomentosa* and *Lysiphyllum hookeri* were provided by Mr Jason Sampson, University of Pretoria, and material from *Piliostigma thonningii* was collected by Dr Bronwyn Egan, University of Limpopo. SHMT was then successfully amplified and sequenced in *L. hookeri* and *P. thonningii*, and added to the corresponding alignment. *B. tomentosa* was unsuccessfully sequenced and could not be used as an outgroup taxon. Once outgroup taxa were

added to each dataset, alignments were produced using the ClustalW (Thompson et al., 1994) function in MEGA6. Alignments still required manual editing to ensure nucleotide positions were biologically sensible. Following this final manual edit, alignments were checked against contig data from Sequencher 4.5 to ensure that no incorrectly coded nucleotides were accidentally included into alignment datasets.

Phylogenetic analysis

Each alignment was exported from MEGA6 in FASTA and NEXUS format respectively. FASTA files for each marker were uploaded to jModelTest version 2.17 to determine the best fit nucleotide substitution rate model. The symmetrical model (SMY) (Zharkikh, 1994) was selected for the Chloroplast marker and SHMT datasets, and the HKY + gamma model (Hasegawa et al., 1985) was selected for ITS. NEXUS files were uploaded to CIPRES (Miller et al., 2010) for phylogenetic analysis using MrBayes 3.2.7 on XSEDE (Huelsenbeck and Ronquist, 2001; Huelsenbeck and Ronquist, 2003). All prior parameter distributions settings were left as default for each dataset. In analyses where datasets for different genes were concatenated, gene regions were partitioned according to the each of the substitution models which had been selected. Posterior probabilities for trees were estimated using Markov Chain Monte-Carlo (MCMC) algorithm with four heated chains for 20000000 generations. Trees were sampled every 2000 generations with a burnin value set for the first 2000 trees. Phylogenies were then viewed on FigTree version 1.4.3. (Rambaut, 2009).

Results

Chloroplast phylogeny

Monophyly in *Tylosema* was well supported in the chloroplast phylogeny (Figure 5). Within the *Tylosema* clade, *T. argentea* and *T. humifusa* formed a well-supported clade with a posterior probability score of 0.98. With only two specimens present (both based on GenBank data), the placement of these two species in one clade requires qualification through the addition of more samples. Attempts to source more material of these species had been unsuccessful to date. *T. fassoglense* was paraphyletic across the tree, as two specimens drawn from GenBank (FJ8011241 and FJ8010911) grouped separately from the remaining *T. fassoglense* specimens. FJ8011241 is a *T. fassoglense* specimen from Tanzania where FJ8010911 is of unknown origin. It is conceivable that the latter was collected from east Africa, meaning the clade formed represents a geographical disjunction within *T. fassoglense*. *T. angolense* was embedded within *T. fassoglense*, which put its status as a separate species in dispute.

T. esculentum was also paraphyletic across the tree with numerous Highveld specimens grouped together with *T. angolense* and South African *T. fassoglense.* Kalahari *T. esculentum* specimens from Botswana and Namibia grouped together in a well-supported clade with a posterior probability score of 0.97, and a separate clade consisting of select Highveld *T. esculentum* was also present with a posterior probability score of 0.9648. These results indicate that Kalahari *T. esculentum* belongs to a separate lineage from their Highveld counterparts. However, the placement of Highveld *T. esculentum* could not be determined from this phylogeny.

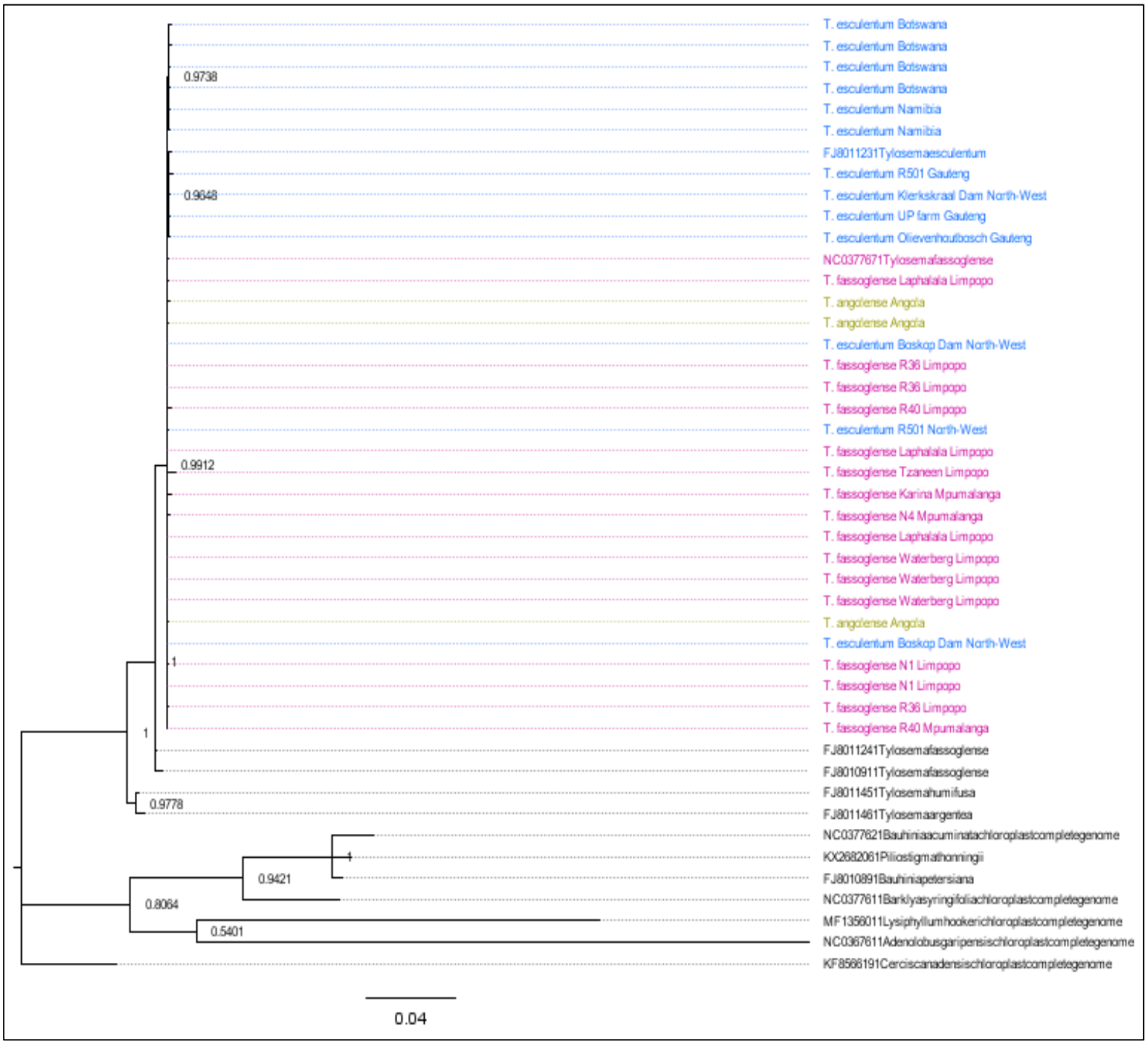

Figure 5 Phylogeny for southern African Tylosema taxa using the trnL-F and psbA-trnH markers. The phylogeny was produced using Bayensian inference method with an HKY + gamma model for nucleotide evolution. Outgroups, selected from GenBank, represent members across the Cercidoideae. Southern African Tylosema species are largely paraphyletic, though T. esculentum from the Kalahari form a well supported clade.

ITS phylogeny

Similarly to the chloroplast phylogeny, the *T. fassoglense* specimen obtained from GenBank (AY258393) grouped separately from the sequences produced from this study (Figure 6). This specimen is the same individual of unknown origin utilized by Sinou et al., (2009). Our assessment remains that sequencing artefacts in this specimen are likely the cause for its placement in our tree topology, or else it is incorrectly identified. The ITS sequence for *T. fassoglense* AY258393 was uploaded to GenBank in 2003 by Hao et al., (2003). The age of these sequences supported our assessment that artefacts were responsible for this specimen's placement in the tree. *T. angolense* specimens formed a well-supported clade with a posterior probability score of 0.9994.

This strongly supported the recognition of *T. angolense* as a separate species within *Tylosema*. However, node support for the split between *T. angolense* and the remaining two species was poor, with a posterior probability score of 0.62. Support for the *T. fassoglense* and *T. esculentum* clade was also inadequate (0.82), and *T. fassoglense* specimens form an unresolved polytomy. *T. esculentum* specimens from the Kalahari and Highveld are grouped together in this tree, though node support was also unsatisfactory at only 0.89. The placement of *T. esculentum* specimens across its distribution remained uncertain in this phylogeny; though there was now some genetic evidence that *T. angolense* is a separate species from *T. fassoglense.*

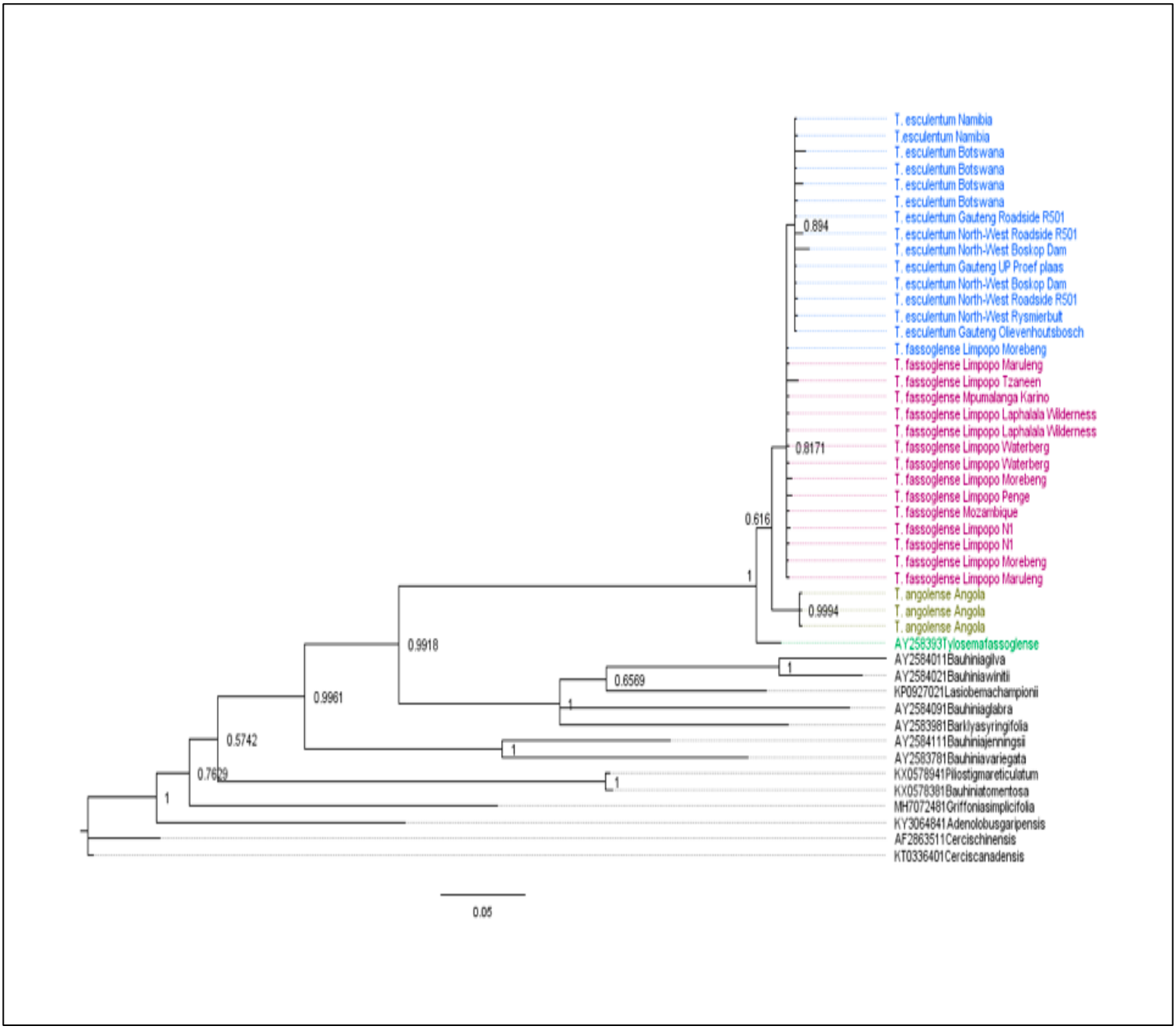

Figure 6 Phylogeny produced for southern African Tylosema taxa using the Internal Transcribed Spacer (ITS) marker. Phylogeny was produced using Bayesian Inference with a SYM model. Outgroup taxa represent members across the Cercidoideae subfamily and were selected from GenBank. T. angolense forms a well supported clade where T. fassoglense and T. esculentum form a clade with poorer support.

SHMT phylogeny

T. angolense and *T. fassoglense* formed a well-supported clade, with a posterior probability score of 1 at this node (Figure 7). This contradicted the result seen in the ITS phylogeny and supported the hypothesis that the two taxa are one species. *T. fassoglense* specimens did group together in a strongly supported polytomy. *T. esculentum* was split across the topology of the tree, with the Kalahari specimens grouped together strongly with a posterior probability support score of 0.9992. The two SHMT copies present in each individual (both for the Namibian and Botswana specimens) grouped together. This suggests that the duplication event of these paralogues is recent. The South African *T. esculentum* specimens grouped separately from their Kalahari counterparts, however in a poorly supported polytomy (0.54 posterior probability score). Node support for the split between Kalahari *T. esculentum* and the remaining *Tylosema* taxa is well supported. However, the placement of South African *T. esculentum* remained uncertain with this dataset.

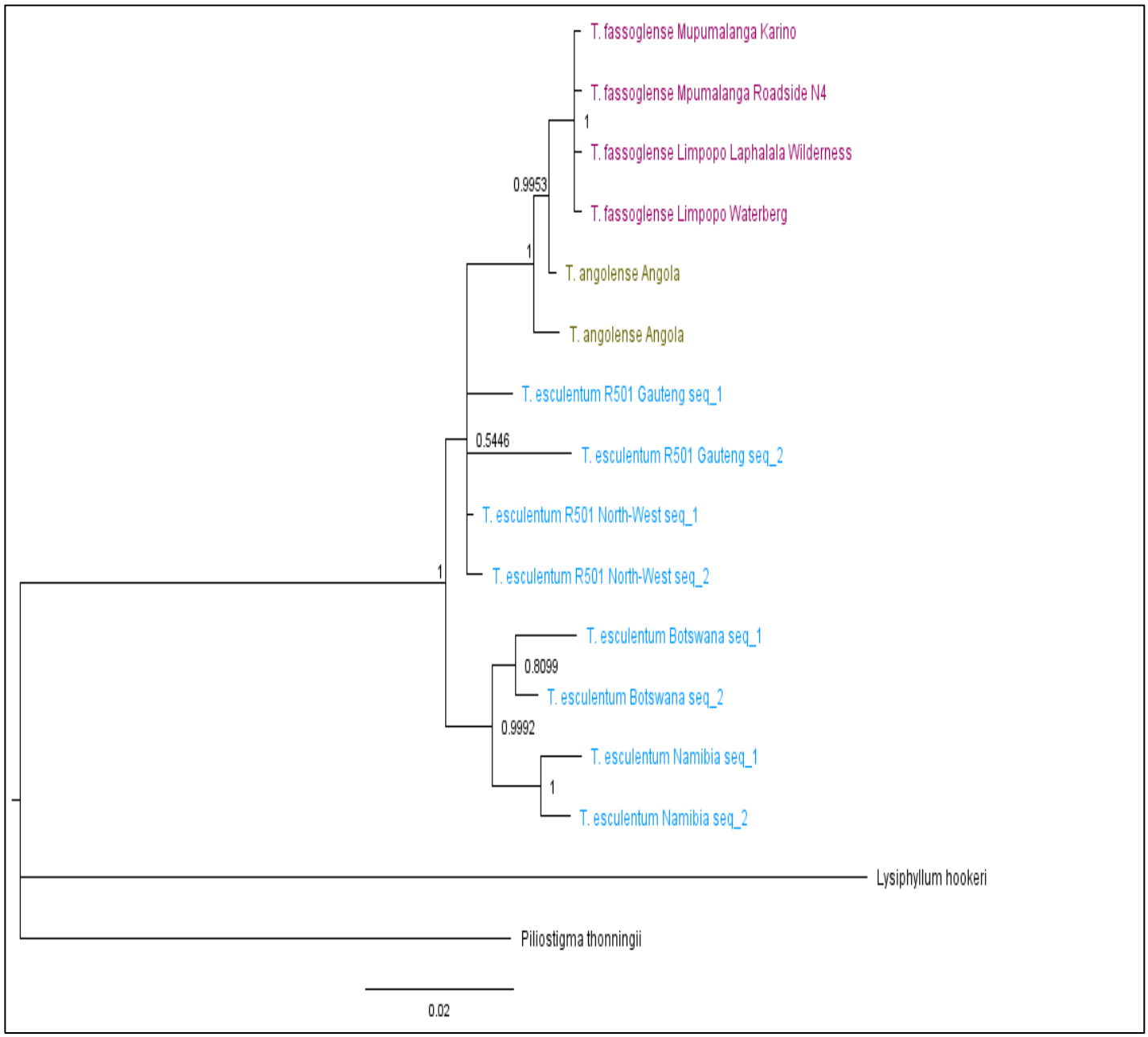

Figure 7 Phylogeny for southern African Tylosema taxa produced using the Sulfite: UDP-glucose sulfotransferase (SHMT) gene, a low copy nuclear gene (LCNG) marker. P. thonningii and L. hookeri were chosen as outgroup taxa because they are members of the Bauhinia s.l. clade (along with Tylosema). Cercidoideae samples were available on GenBank for the SHMT gene. T. angolense and T. fassoglense form a well supported clade where T. esculentum are split between Kalahari and Highveld grassland lineages, though only the former was well supported.

Combined phylogeny

The same subset of specimens used for the SHMT dataset was utilized to form a concatenated alignment for all four molecular markers sequenced (Figure 8). This subset was chosen to ensure the greatest number of taxa were represented by each marker. It remained the case that certain specimens still lacked sequence data due to mixed sequencing success, i.e. *T. esculentum* off the R501 highway in Gauteng did not sequence for the *psb*A-*trn*H intergenic spacer and neither outgroup taxa (*P. thonningii* and *L. hookeri*) successfully sequenced for the ITS marker. However, Bayesian inference methods are not significantly affected by missing data (Wiens and Morrill, 2011). The phylogeny produced from this concatenated dataset clearly showed three lineages across *Tylosema* species within southern Africa. The first was strongly supported with a posterior probability score of 1 and contained *T. fassoglense* and *T. angolense* specimens.

T. fassoglense and *T. angolense* specimens grouped separated within this lineage and were well supported as two distinct species with posterior probability scores of 0.99 and 1 respectively. The second lineage contained two *T. esculentum* specimens from the Highveld grasslands; however, support for this group was tenuous at 0.77 posterior probability score. The remaining lineage contained two specimens from the Kalahari Desert in a strongly supported clade with a posterior probability score of 1. This combined phylogeny supported *T. angolense* as a separate species from *T. fassoglense and* indicated that Kalahari *T. esculentum* belonged to a separate lineage from Highveld *T. esculentum.* The placement of the latter remained uncertain as support for this lineage is weak, a result consistently shown across all phylogenies produced in this study.

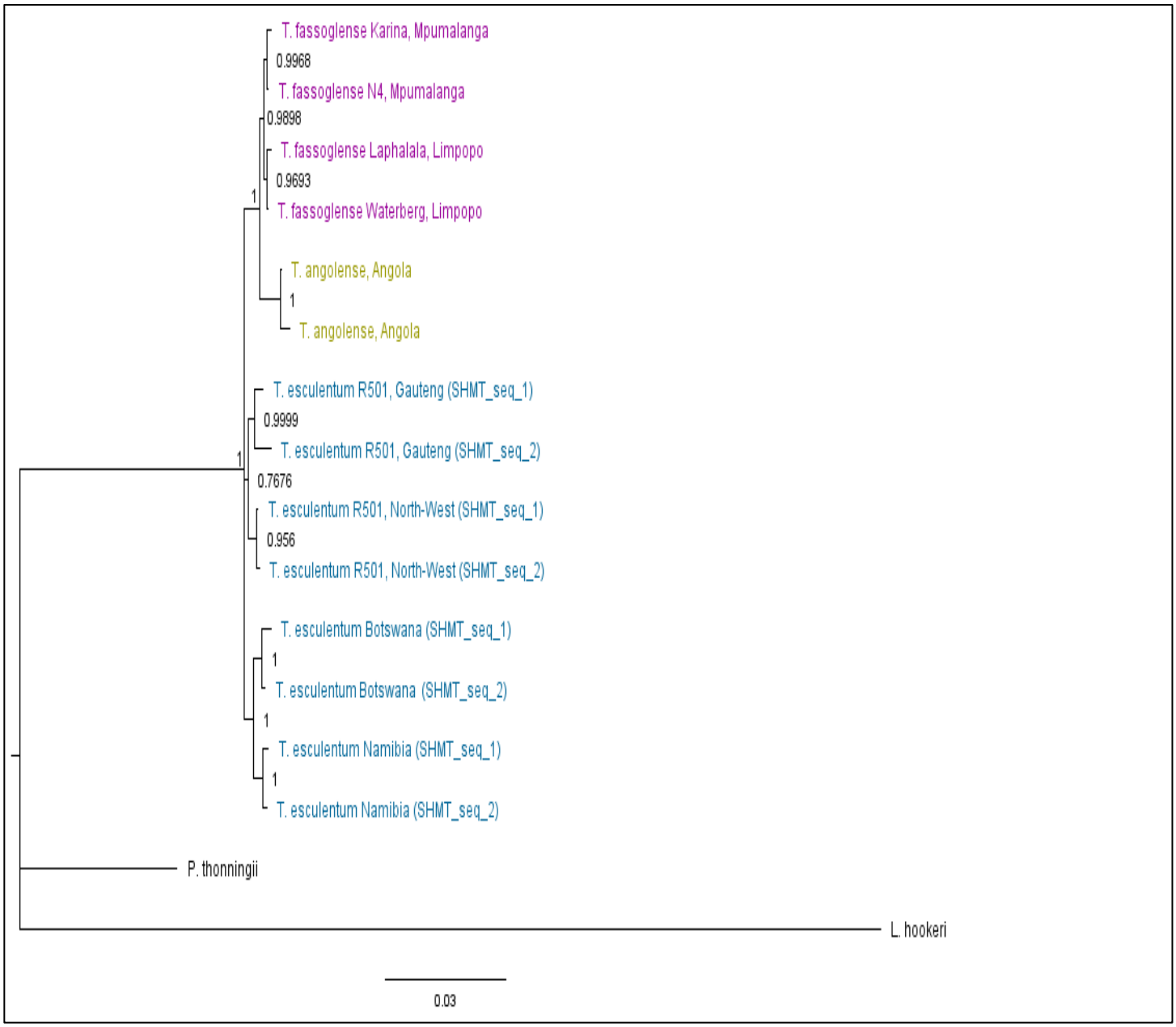

Figure 8 Phylogeny produced for southern African Tylosema taxa using a combined dataset of the follow four markers: trnL-F, psbA-trnH intron and spacer, ITS, and SHMT. Outgroup taxa were P. thonningi and L. hookeri since these taxa were represented by nearly all four markers chosen, except for psbA-trnH intron and spacer. T. angolense and T. fasssoglense belong to the same lineage but remain distinct species. This result is well supported by posterior probability scores. T. esculentum remains split along two lineages: the Kalahari and Highveld grassland lineages. The Kalahari lineage remains well supported where the Highveld grassland lineage remains poorly supported.

Discussion

Phylogenetic placement of *Tylosema sp*.

The aim of this study was to elucidate species limits within southern African *Tylosema*. Clarifying the taxonomy of the genus would not only have economic and agricultural benefits but also provide the basis for ecological and evolutionary studies in future. There is good evidence from our phylogenetic analyses indicating that the recently described *T. angolense* is a distinct species, confirming the result of Castro et al., (2005). The issue of morphological intermediates (atypical *Tylosema* taxa) between *T. angolense* and *T. fassoglense* still remains unresolved, as no such intermediates could be included in this assessment.

The phylogeny from the combined dataset showed that *T. angolense* and *T. fassoglense* belong to the same lineage. It is conceivable that these intermediate *Tylosema* are hybrids of the two species, considering how frequently plant species hybridize and that recently diverging species may not have well evolved reproductive boundaries (Soltis and Soltis, 2009). *T. fassoglense* is known to display a wide variety of morphological characters across its wide distribution (Coetzer et al., 2011). The recent discovery of *T. angolense* warrants a further investigation into the purported morphological variability found in *T. fassoglense*: undescribed *Tylosema sp*. have likely been identified as *T. fassoglense* across sub-Saharan Africa given the taxonomic neglect the genus has received.

T. esculentum was split across two lineages, predictably between the two contrasting biomes the species occupies. Kalahari *T. esculentum* are strongly supported as a distinct clade in the

chloroplast, SHMT, and combined dataset phylogenies. Although this finding was not supported by the ITS tree, this result could be explained by a potential incomplete lineage sorting, where *T. esculentum* from both lineages share polymorphisms in the ITS region from a common ancestor. The placement of Highveld *T. esculentum* is more complicated due to the lack of node support for this lineage across all of our phylogenies. Better support for this lineage would likely be found if more specimens from across the Highveld are added to the combined dataset or if more molecular markers are added to the analysis. The inclusion of the SHMT gene was vindicated given the phylogenetic signal it provided. This provides good evidence that LCNG markers can delimit species where traditional markers alone (ITS and chloroplast) prove insufficient. The inclusion of more LCNG may clarify the placement of the Highveld *T. esculentum* lineage and provide strong support for the recognition of these members as a distinct species.

SHMT in *T. esculentum*

The two SHMT sequences in *T. esculentum* (both lineages) individuals are not the result of a hybridization event between two other *Tylosema* species. This is because the two sequences present in each individual grouped together, as opposed to one grouping with either *T. fassoglense* or *T. angolense. T. esculentum* is also a known polyploid, an autopolyploid according to Mornaghan (1995), and the lack of hybridization evidence from the SHMT marker provides support for this finding. This means that the two SHMT sequences originated in *T. esculentum*, prior to the split between the Highveld and Kalahari lineages. However, the indel event may have occurred early in the evolution of *T. esculentum* considering it is present in both

lineages. As mentioned previously, the incorporation of additional LCNG markers into future phylogenetic studies would clarify the evolutionary history of *T. esculentum*.

Inconsistencies across Phylogenies

Topological inconsistencies were seen across all four phylogenies produced, each one indicating different species level delimitations. For example, *T. angolense* separate from *T. fassoglense* in the ITS phylogeny but lumped together with *T. fassoglense* in the SHMT tree. Both of these topologies were well supported in the respective phylogenies (1 and 0.9994 respectively). A persistent issue taxonomists face when constructing molecular phylogenies is the discrepancy seen between 'gene trees' and 'species trees', or phylogenies that accurately represent species limits (Maddison, 1997). Phylogenies inferred using only a single gene region merely represents the genealogy, or evolutionary history, of the single region in question, not necessarily the taxonomic relationships between taxa (Nichols, 2001; Slowinski and Page, 2008). There are numerous reasons why gene trees may differ topologically. Incomplete lineage sorting may be present between different gene regions, where taxa retain ancestral polymorphisms due to slow mutation rates or deep coalescence of alleles in ancestral populations (Hudson and Turelli, 2003). Ancient hybridization or introgression events may have taken place between taxa resulting in shared orthologues at select gene regions (Petit and Excoffier, 2009). Gene duplication events may also cause discordance between gene tree topologies, though numerous authors have shown that gene duplication can result in phylogenetically informative regions (Rasmussen and Kellis, 2007; Sanderson and McMahon, 2007). This proved to be the case for the SHMT gene in *Tylosema*, as shown in our phylogenies.

These discrepancies between gene trees are further complicated by the fact that chloroplast and nuclear genomes evolve at different rates, leading to the use of chloroplast markers in plant phylogenies at and below the species level to be commonly criticized (Kress and Eriksen, 2007; Fazekas et al., 2009, Hollingworth et al., 2011). Chloroplast genomes are haploid and generally uniparentally inherited in plants, meaning their effective population size is a quarter of that of the nuclear genome (Ross 1990). This smaller effective population size reduces nucleotide substitution rates in chloroplast genomes; however, plant chloroplasts have even lower substitution rates than animal mitochondrial genomes, which are also uniparentally inherited and haploid (Petit and Vendramin, 2007). The unusually slow rate of evolution in chloroplasts partially explains why no universal chloroplast barcode exists for plants.

A common phenomenon seen in plants occurs where plant taxa or populations may have the nuclear genome of one species but the chloroplast (and mitochondrial) genome of another, this is known as 'cytoplasmic capture' (Riesenberg and Soltis; 1991). Several mechanisms have been proposed for how cytoplasmic capture occurs such as greater genetic drift in chloroplast genomes compared to nuclear genomes, male sterility in hybrid plants, or unidirectional inundation of pollen from one species to another (Potts and Reid, 1988; Petit et al., 1997; Petit and Vendramin, 2007). The resulting effect in molecular taxonomy is that shared chloroplast genomes across closely related taxa further reduces the effectiveness by which chloroplast markers can delimit plant species. Riesenberg and Soltis (1991) found that cytoplasmic capture had taken place across multiple plant groups, even ones which are not known to hybridize frequently, meaning cytoplasmic capture cannot be simply ruled out in chloroplast phylogenies.

In the incidence where chloroplast and nuclear phylogenies conflict, the former often reflects geographically clustered taxa where the latter better reflects actual taxonomic relations

(Riesenberg and Soltis, 1991; Inamura et al., 2000; Rautenberg et al., 2010). In the case of *Tylosema* the Kalahari *T. esculentum* grouped separately as geographically distinct in the chloroplast phylogeny, but *T. esculentum* from the Highveld showed no regional grouping, and three such specimens grouped together in a hard polytomy with *T. fassoglense* and *T. angolense*. Two *T. fassoglense* sequences taken from Genbank (FJ8011241 and FJ8010911) are likely of east African origin and did group separately from the southern African taxa. However, these two specimens were sequenced by Sinou et al., (2009), nearly a decade ago. Variation in the FJ8011241 and FJ8010911 *T. fassoglense* sequences could be the result of sequencing artefacts, errors which would have been more common in earlier Sanger sequencing reactions. *T. fassoglense* NC037767 grouped with specimens sequenced in this study and was sequenced and uploaded to GenBank in 2018 by Wang et al., (2019). This specimen was sampled from the Herbarium of the University of Johannesburg, meaning geographical separations within *T. fassoglense* could be present but these results remains inconclusive.

ITS and SHMT phylogenies produced conflicting species limits which could be explained by any of the reasons given above. The concatenated dataset was produced to overcome these issues, allowing each gene to contribute informative nucleotides to the phylogeny (Kluge, 2004). This method proved successful as species limits across all three *Tylosema* taxa were well supported in the combined phylogeny. However, numerous authors have addressed the issues surrounding concatenated datasets, especially when the individual genes comprising the concatenation produce conflicting phylogenies (Kolaczkowski and Thornton, 2004; Mossel and Vigoda, 2005; Felsenstein, 2006; Edwards, 2008).

Concatenated Dataset Phylogeny

Simulation studies have shown that accurate species trees comprising five or more taxa may reflect the topology of certain gene trees, but share short, internal branch lengths with gene trees that conflict with accurate species trees (Degnan and Rosenberg, 2006). Kubatko and Degan (2007) have shown that in concatenated data sets where the above phenomenon is present, species level inference using standard phylogenetic approaches is likely to be positively misleading. These shorter internal branch lengths are due to greater incomplete lineage sorting across loci, which is more likely to occur in taxa where rapid speciation has taken place and/or the effective population size relative to speciation rate is large (Pamilo and Nei, 1988; Rosenberg, 2002). Essentially, for species which have evolved quickly and have not experienced significant genetic drift to fix loci, traditional molecular concatenations do not strictly result in phylogenies which match species trees (Pamilo and Nei, 1988). Concatenated datasets may also suffer from an 'overshadowing' effect, whereby a select few gene partitions provide the entire phylogenetic signal. This means genes with lower phylogenetic signals are effectively masked; meaning the subtle but potentially informative signal they provide is lost through concatenation (Baker et al., 1998).

It is conceivable that (some of) these scenarios are present in the *Tylosema* combined dataset, given the topological incongruences between the different trees each gene produced. Authors have suggested that *Tylosema* species are recently diverging given the extent of interspecific morphological overlaps seen in wild populations (Castro et al., 2005; Coetzer et al, 2011). Various programs have been developed to overcome gene tree vs species tree conflicts, ones which putatively outperform the standard method of phylogeny inference through concatenated datasets

(Yang and Rannala, 2010; Heled and Drummond, 2010; Drummond et al., 2012; Mirarab et al., 2014). These approaches have been proposed to address the issue of incomplete lineage sorting, allowing the lineages of individual genes to be incorporated into the analysis such that signals from each one could infer an accurate species tree (Maddison and Knowles, 2006). Collectively these programs implement multispecies coalescent models, incorporating coalescent theories used in population genetics to infer diverse lineages across closely related species (Yang and Rannala, 2010; Yang, 2015). Assuming no recombination through hybridization or introgression takes place between species, gene lineages should coalesce further back in time than the divergence of species lineages (Camargo et al., 2012). These methods theoretically ensure that the conflation of species trees with gene trees is overcome (Edwards, 2008).

The utilization of coalescence models in phylogenetics has come under criticisms itself. Sukumaran and Knowles (2017) determined that multispecies coalescent models inferred population structure as opposed to actual species limits. This is because assumptions implicit in these models cannot distinguish intraspecific lineages from lineages signifying species boundaries. These findings were confirmed by Leache et al. (2019). Most multispecies coalescent models assume that no gene flow takes/has taken place between taxa in question (Jackson et al., 2017). This assumption is unrealistic in nature (especially for plants), and models excluding gene flow as a parameter overestimate species limits compared with those which account for gene flow across generations (Eckert and Carstens, 2008; Camargo et al., 2012). Although increasing parameters may improve the accuracy of species delimitation, it comes with the drawback of requiring greater computational power for analysis to be conducted (Leache et al., 2019).

The implementation of multispecies coalescent models to *Tylosema* may yet be a fruitful approach, especially if more nuclear genes are incorporated into future assessments. Researchers must

consider the range of programs available and what parameters they offer to ensure that models implemented best fits their study taxa (Carsten and Dewey, 2010; Carsten et al., 2013). Taxonomists should also consider evidence beyond the genomic regions, such as the ecology, morphology, and behaviour (where applicable) of species complexes under investigation (Will et al., 2005; Carsten et al., 2013; Sukumaran and Knowles, 2017). This would help provide a more holistic sense of species limits where multispecies coalescent models conflate population level lineages with species lineages (Carsten et al., 2013; Sukumaran and Knowles, 2017; Larsen et al., 2017).

It should also be noted that while concatenated datasets may theoretically not provide species trees, the phylogenies they produce may still give some indication towards what the true species tree may be (Edwards, 2008). Concatenated datasets also perform better at retrieving accurate species trees where gene trees do not display greatly conflicting topologies (low amounts of incomplete lineage sorting) (Johnson et al., 2019). As discussed above, this may not be the case for *Tylosemsa*. However, the combined dataset phylogeny may still provide a good indication of species lineage history in southern Africa considering the ecological and morphological evidence it supports.

Tylosema sampling

A notable limitation in our analysis is the under-sampling present at various levels. The decision to only test *Tylosema* species across southern Africa was justifiable given the immediate agricultural potential these members could have in the region. However, a full taxonomic review, or monograph, of all *Tylosema* taxa would prove even more useless in both this regard and in terms of addressing long standing evolutionary questions. Very little is known about *Tylosema* species

and east African species such as *T. argentea* and *T. humifusa* may also be edible. With *Tylosema*, as with many African plants, under-sampling is a persistent issue that interferes with taxonomic and conservation efforts, as was seen in Castro et al. (2005). The lack of sufficient herbarium vouchers for *Tylosema* species contributed to the inability to identify atypical (morphologically intermediate) *Tylosema* specimens. Debates in the past have raged on among taxonomists as to whether increased taxon sampling significantly improved the accuracy of phylogenies, or whether researchers should focus on a few select taxon while utilizing multiple genes (Cummings et al., 1995; Graybeal, 1998; Townsend and Lopez-Giraldez, 2010; Nabhan and Sarkar, 2011).

Constructing phylogenies using a wide array of gene regions is now a standard practice among molecular taxonomists (as discussed previously), but a wealth of evidence, both from simulation studies and real data, shows that insufficient sampling across a groups distribution can result in misleading phylogenies and underestimations of species diversity (Zwickl and Hillis, 2002; Nabhan and Sarkar, 2011; Smith et al., 2013). The concatenated tree produced for *Tylosema* contained fewer specimens than either the ITS or chloroplast phylogenies, but the same number as that of SHMT. As stated, this was to ensure that the SMHT gene at least was included in every specimen used in the combined tree. The drawback faced was thus a reduction in specimens present. This lack of taxa representation does not mean the lineages shown in our study are incorrect, but the addition of more specimens could reveal other lineages not present in our tree. The specimens which were present did cover a wide distribution (Angola, Namibia, Botswana, and South Africa) and included all three putative species in question, and thus gives an accurate account of *Tylosema* evolution in southern Africa.

The SHMT duplication event in *T. esculentum* specimens did pose a time-consuming challenge during analysis, even if Indelligent was able to tease co-occurring sequences apart in the end. This reduced the number of *T. esculentum* specimens utilized, as fewer specimens were easier to process and resulted in a more accurate result. Future studies would do well to incorporate more *T. esculentum* specimens, and look to invest in more sophisticated methods for the detection of nuclear markers. Hollingsworth et al. (2016) proposed the use of target enrichment sequencing to find high volumes of nuclear markers for plant phylogenetic sequences. Johnson et al. (2019) implemented this technique and developed a library of 353 nuclear gene probes across over 600 angiosperm species. These probes could be applied to *Tylosema*, and the results from this study indicate that the risk of including previously developed markers can be fruitful.

Conclusion

This study has provided support for the recognition of *T. angolense* as a distinct species separate to *T. fassoglense* and has provided some evidence that South African *T. esculentum* belong to a separate lineage from the Kalahari Desert *T. esculentum* (Namibia and Botswana). The inclusion of the LCNG, SHMT, proved fruitful considering that the barcoding markers (*trn*L-F, *psb*A-*trn*H, and ITS) were insufficient in delimiting species on their own. These findings not only demonstrate that LCNG genes are invaluable for phylogenetic analysis, but that the genus is likely more diverse than currently recognized i.e. there could be more *Tylosema* species than the five currently described. Researchers have long suspected this to be the case considering the wide variation both at a morphological and genetic level seen in *Tylosema* members.

Addressing the issue of species diversity in *Tylosema* warrants sampling taxa across sub-Saharan Africa, a task which has not adequately been done before to our knowledge. Insufficient sampling has hampered taxonomic treatments for *Tylosema* in the past and will remain a persistent bottleneck until it is thoroughly addressed. Future researchers should also consider applying a large number of nuclear genes and multispecies coalescent models in order to detect species limits within *Tylosema*. For example, this study benefited greatly from the inclusion of the SHMT gene, however uncertainties surrounding the placement of the Highveld *T. esculentum* lineage remained. Additional nuclear genes could address this issue and provide clarity on species limits within the genus as a whole. *Tylosema* species have a complex evolutionary history, and researchers will have to contend with polyploidy and incomplete lineage sorting in order to resolve the current taxonomic issues.

Chapter 3

Microsatellite marker development

Introduction

Tylosema species across sub-Saharan Africa are critically understudied, despite being potential sources of food for countries across this region and globally (Van der Maesen, 2006; Jackson et al., 2010). Little is known about the conservation status, seed dispersal, genetic variability, and general ecology of members of these potentially important species. If *Tylosema* species are to be successfully domesticated it must first be ensured that wild populations are protected and managed, and that their existence as genetic resources is well understood and cared for by scientists. Microsatellite markers remain a popular genetic tool among ecologists used to address the issues mentioned above in wild plant populations (Selkoe and Toone, 2006). As stated previously, SSR markers have been developed in *T. esculentum* using the notoriously challenging FIASCO method by Nepolo et al., (2009) nearly a decade ago. These markers have also never been tested in other *Tylosema* species.

The aims of this study were to develop new SSR markers using genomic data from *T. esculentum* with the intention to test for amplification in *T. fassoglense* specimens too. Additionally, we aimed to determine whether or not the markers we designed were in fact polymorphic (had multiple alleles present of varying repeat length), as only polymorphic markers can reliably be used in future genetic variability studies.

Methods

Mining for SSR regions

Genomic scaffold data from *T. esculentum* was produced by Dr Chris Cullis, Case Western Reserve University, Cleveland, Ohio, using Illumina and PacBio sequencing platforms. With Dr Cullis' permission we downloaded the scaffold data in FASTA format to mine for microsatellite regions using the program MSATCOMMANDER (Faircloth, 2008). Parameters for SSR marker mining were set for perfect repeat regions only with repeat motifs set for trimers, tetramers, pentamers, and hexamers. Monomers and dimers were purposefully excluded since small length variations in these motifs are more difficult to detect than in SSRs of longer repeat motifs.

MSATCOMMANDER has the primer designing software, PRIMER3, built-in, which was utilized to design primers in the flanking regions of SSR markers identified where possible. GC content for primers was set at 40-60%, a total amplicon size was set between 150 to 450 base pairs. The microsatellite and primer pair data were then imported into Microsoft Excel where it was filtered for quality assurance and to select potential microsatellite regions. MSATCOMMANDER overestimates the number of microsatellite regions in genomic data, meaning duplicates of many identified regions were present, and many regions found lacked primers pairs given the strict parameters set for primer design.

Once we had removed duplicate SSRs and those lacking primer data, further filtering was conducted to ensure only regions with sufficient repeat counts were used. For tetramer, pentamers, and hexamers, a minimum repeat count of 7 was chosen. A minimum repeat count of 10 was chosen for trimers. This filtered dataset was then reorganized by repeat motif type e.g. trimers to

hexamers. This left us with a primer pair dataset of 46 markers consisting of 35 trimers, 9 tetramers, and 2 hexamer microsatellite regions (see Supplementary data). No pentamer markers were left in the dataset following quality pruning. From this final dataset three trimer markers, one tetramer marker, and one hexamer marker, all of varying amplicon lengths, were then randomly chosen for amplification and detection of polymorphisms (Table 2).

Table 2 Five primers pairs for the microsatellite regions tested for polymorphism in T. esculentum and T. fassoglense.

Locus	Forward primer	Reverse primer	Motif	Repeat count	Amplicon size
MA1	TACTCGTCGTACTGCACTGG	TGGTGGTGGAGTTGAGGAAG	AAGGAG	8	339
MA6	CCAAGTTGGCAGAGTGGTTG	TCATCAACAGCTAGCCTCCG	ACAT	8	346
MA8	AGATGCTGCGTTACCTTGAAG	TCCAGGGCCCACAATTGATG	ATC	13	298
MA9	GGGATCCTCATTGCTGACAG	CCTCTGAAATTTCTCGCCAGG	ATC	13	330
MA10	CAACGCTGAGGTTGTGTCTG	AGAATCCAAGTTGAGGTCATGG	ATC	11	409

Microsatellite Amplification

SSR marker amplification was tested on four or five specimens from two *T. esculentum* and two *T. fassoglens*e localities: The *T. esculentum* localities of LM14 and LM5, and *T. fassoglense* localities of LM11 and LM24 (see Figure 9). These populations broadly represent the distribution of the two species across South Africa. LM14 is along the R501 highway in the North-West Province and LM5 is found on a plot of rural land in Olievenhoutbosche, Centurion in Gauteng. LM11 is found along the R40 highway in Mpumalanga, and LM24 is found in the Lapalala Nature Reserve in the Waterberg. Leaf material from each locality was collected in silica gel. Each marker

was amplified across all samples with each reaction containing 10µl sequencing buffer from Bioline (comprising final concentrations of 1mM dNTPs and 3mM of $MgCl_2$), 0.5 MyTaq (2.5 units at final concentration), 1 µl forward and reverse primer solution respectively (final concentration of 0.2µM each), 32.5µl distilled H2O, and 5µl of DNA (approximately 150-200ng). PCR conditions were set at an initial denaturation temperature of 94C for 4 minutes, 30 cycles of 94C for 30 seconds, 55C for 60 seconds, and 72C for 2 minutes, followed by a final extension period at 72C for 5 minutes. Amplified PCR products were visualized using gel electrophoresis on an agarose gel (3% TBE) run at 80V for 3 hours in order to assess banding patterns for each marker in both *Tylosema* species. If either multiple bands or bands of varying sizes were detected in the visualized PCR products then it was inferred that the marker in question was polymorphic.

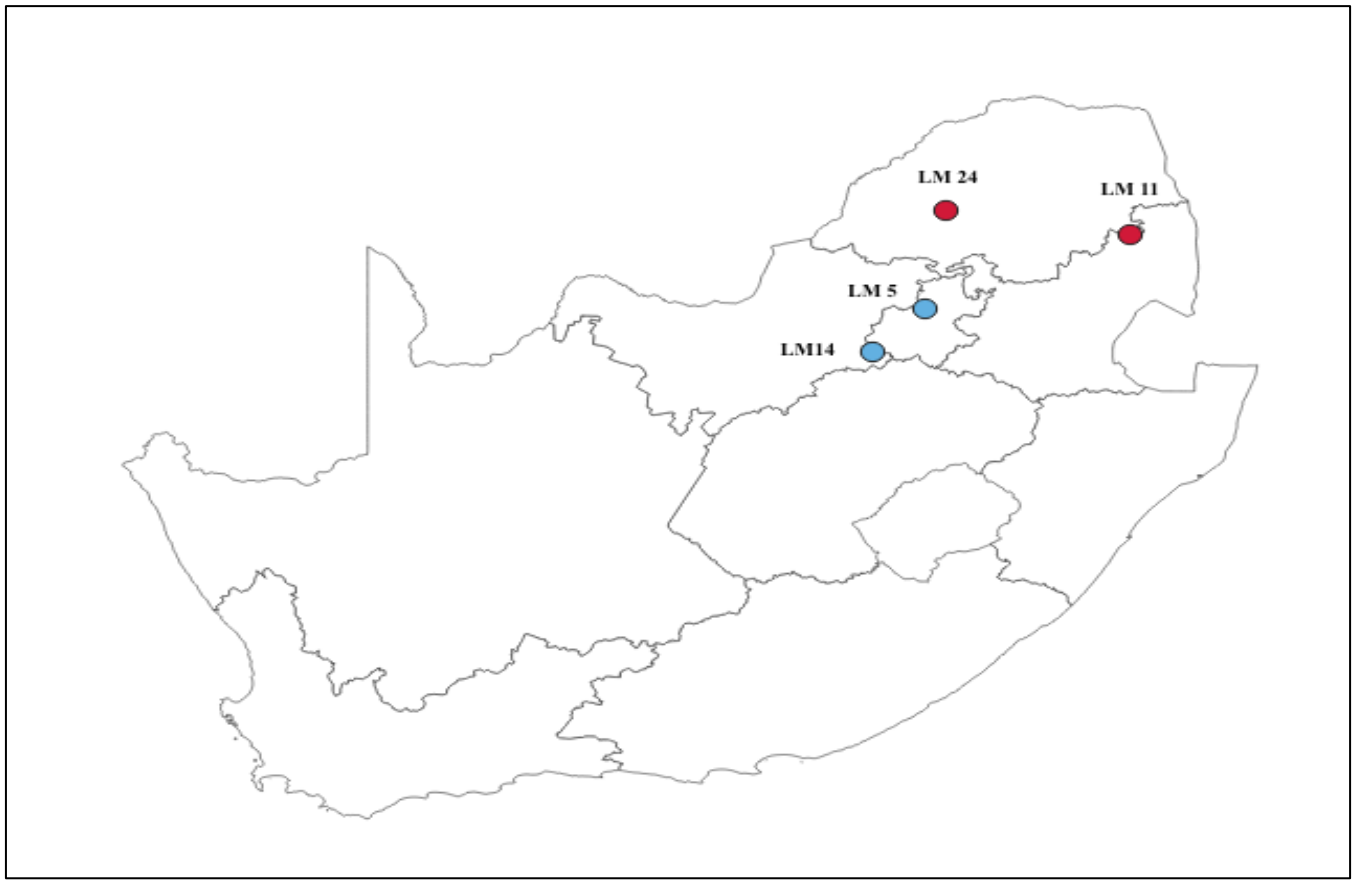

Figure 9 Map of South Africa showing the two T. esculentum (LM5 and LM14 [ ]) and two T. fassoglense (LM11 and LM24 [ ]) populations used to test polymorphisms in the five SSR markers selected.

Results

Amplification and Characterization of MA1 primer pair

Electrophoresis on agarose gels showed patterns of polymorphisms based on size in samples 14.8; 14.14; 11.7; 5.15; 24.6; and 24.8 (Figure 10). Not only was there at least one polymorphic individual at that locus in each population, but that the MA1 marker can be used in both *Tylosema* species.

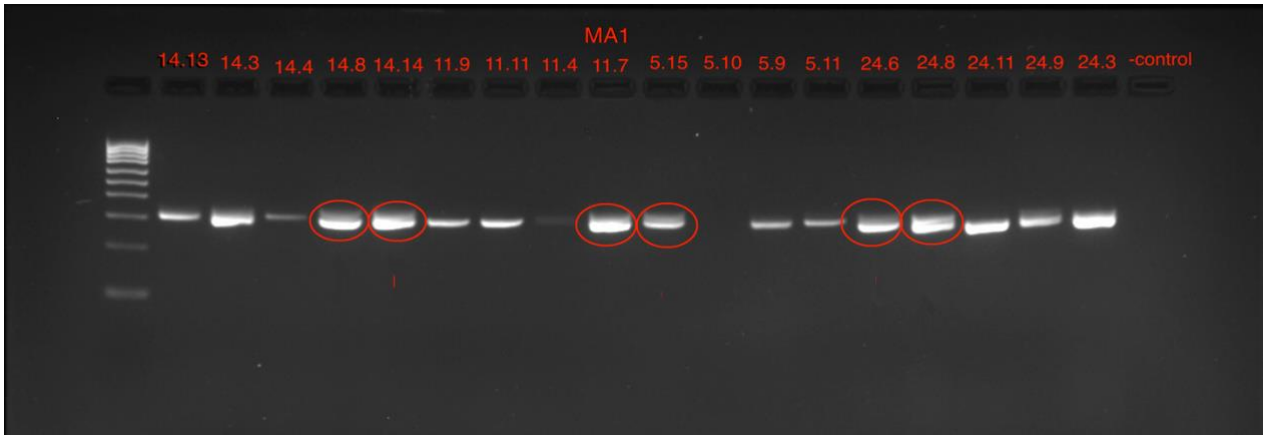

Figure 10 Gel image for MA1 primer pair marker detailing various polymorphic specimens. Polymorphic specimens are 14.8 and14.14 (T. esculentum from the North-West Province), 11.7 (T. fassoglense from Mpumalanga), 5.15 (T. esculentum from Centurion, Gauteng), and 24.6 and 24.8 (T. fassoglense from the Waterberg, Limpopo).

Amplification and Characterization of MA6 primer pair

Amplification was successful in both *Tylosema* species and across all four populations; however, no visible polymorphism could be detected from the agarose gel after electrophoresis (Figure 11). Increasing electrophoresis time yielded no further separation in the migrating bands. Band brightness was high in a number of samples for the MA6 primer pair, indicative of a high DNA concentration in certain samples, which could hinder band separation of closely sized alleles. An

x10 dilution was made for certain samples PCR products and rerun on an agarose gel for 3 hours. The diluted products still showed no detectable polymorphisms. Size discrepencies between samples from populations LM14 and LM5 and LM24 seem to be revealed (the former being slightly smaller than the latter two). However, this is likely the result of uneven migrartion across the gel given the relative positions of sample bands to the molecular ladders used. Size polymorphisms for MA6 between the different Tylosema populations are thus not present.

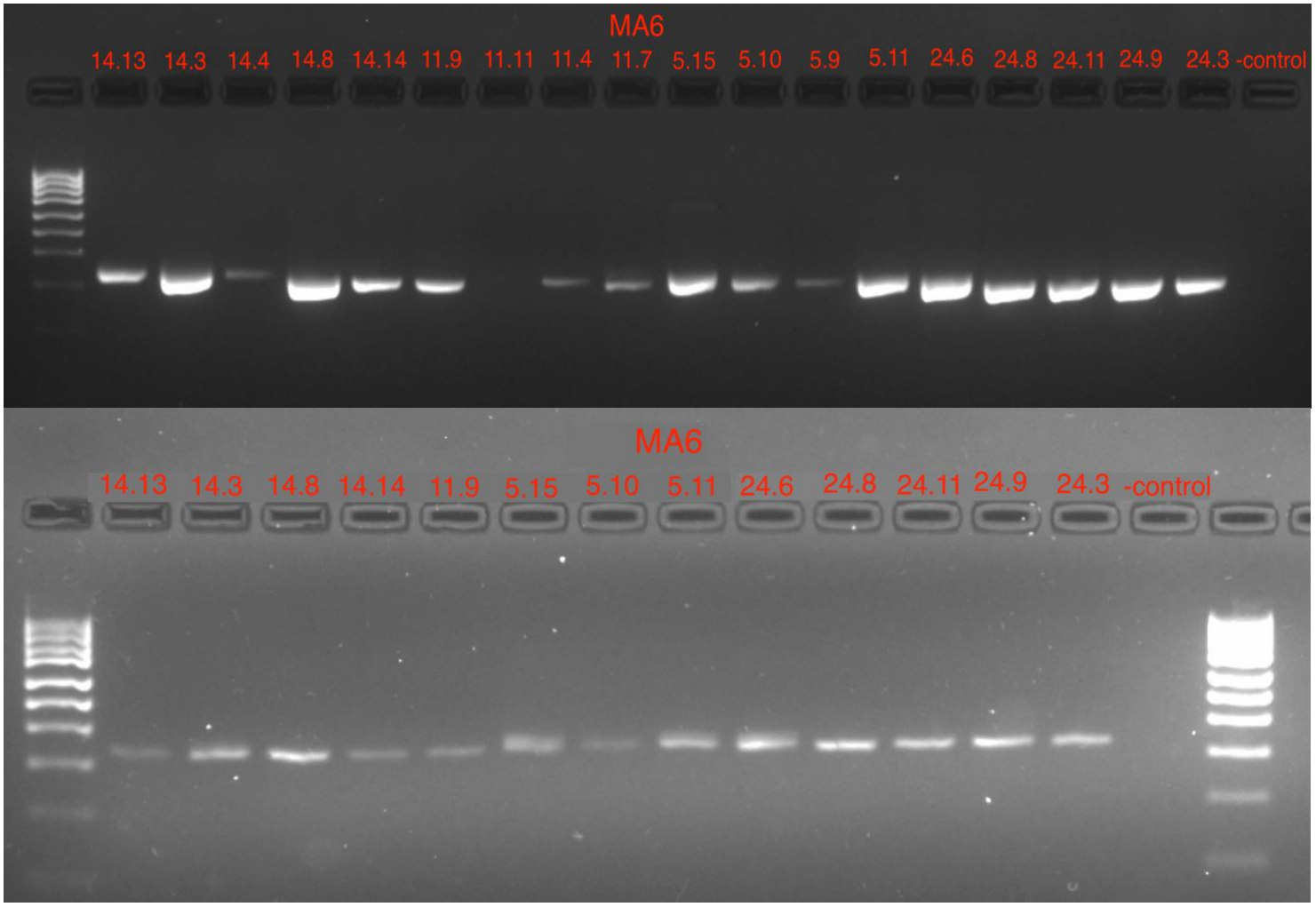

Figure 12 Gel images for MA6 primer pair marker. Above: original PCR products. Below: Diluted PCR products. Lopsided migration in diluted samples gives the impression of size polymorphisms between populations LM14 and LM5 and LM24. Using the molecular ladders on either side of the sample runs shows this isn't the case.

Amplification and Characterization of MA8 primer pair

Samples 14.3; 14.8; 14.14; and 11.7 showed clear length polymorphism with multiple bands present in the gel visualization (Figure 12). As with MA6, Marker MA8 showed bright bands especially in the LM5 and LM24 populations. As this likely hindered band separation, PCR products for these samples were diluted to x10 the original concentration and rerun on a gel. The dilution showed polymorphisms for samples 24.8 and 24.9 (Figure 13).

Amplification and Characterization of M9 primer pair

Marker M9 was not highly polymorphic as only specimens 14.4 showed obvious multiple bands in the gel visualization (Figure 12). However, specimen 14.4 showed five bands at this locus, the most bands for any specimen using any marker. Similarly to marker MA8, a number of samples showed bright bands in populations LM5 and LM24. A x10 dilution was also made for samples from these populations and rerun on a gel. The dilution allowed for the detection of polymorphisms in samples 5.11 and 5.15 (Figure 13).

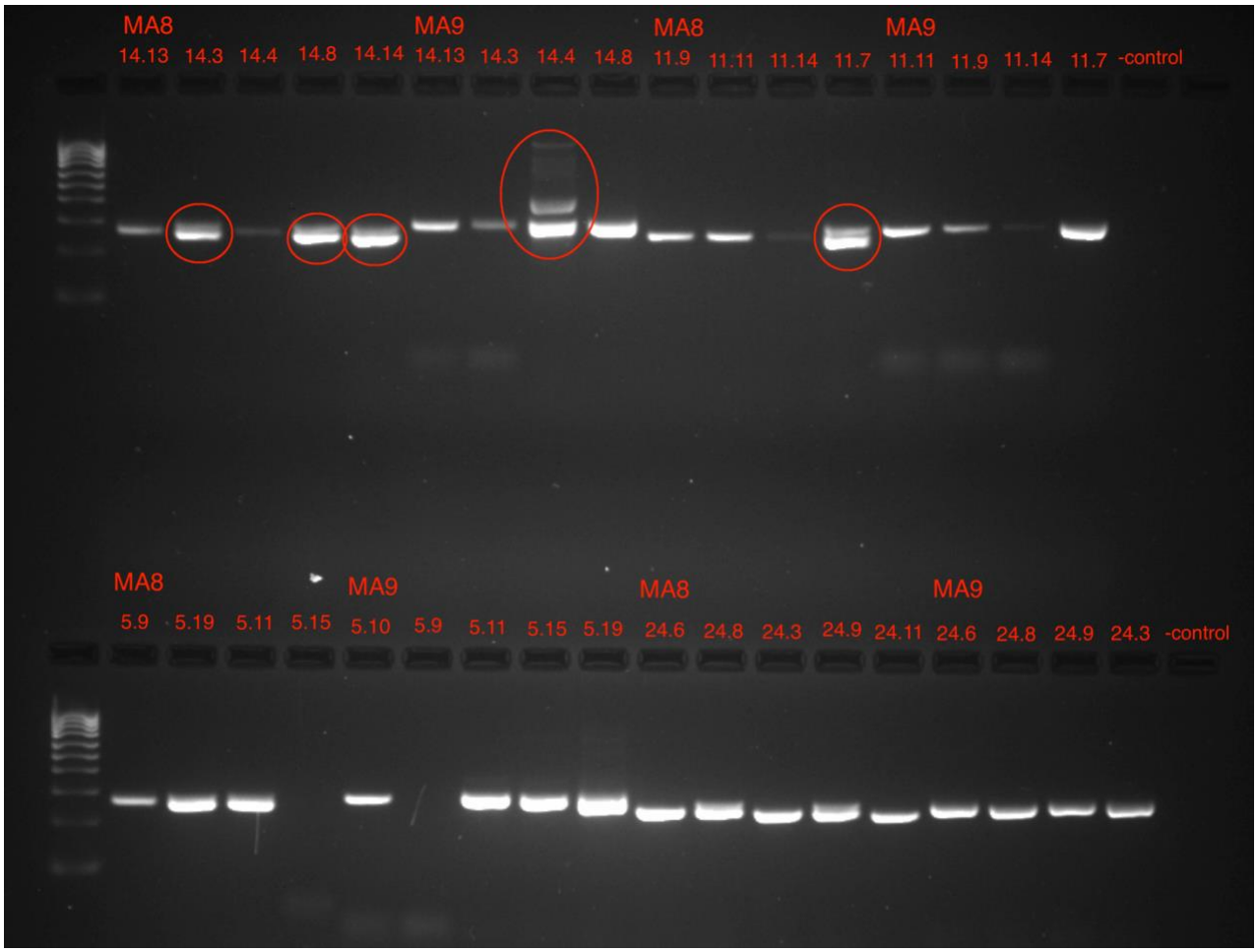

Figure 12 Gel image for MA8 and MA9 primer pairs. Above: Specimens showing polymorphisms for marker MA8 are 14.3, 14.8 and 14.14 (T. esculentum from the North-West Province) and 11.7 (T. fassoglense from Mpumalanga). Polymorphic specimens for marker MA9 are 14.4 (T. esculentum from the North-West Province). Below: Visualized bands were too bright which obscure the detection of polymorphisms in certain specimens.

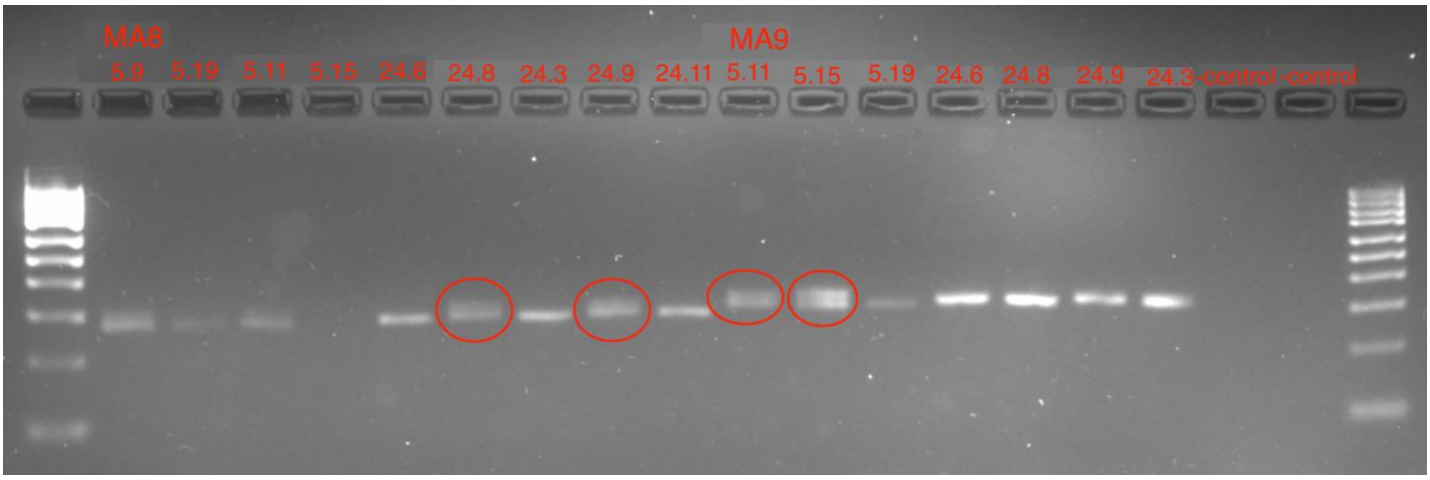

Figure 13 gel image of markers MA8 and MA9 where PCR products for certain specimens had been diluted (x10). Polymorphisms can now be seen in specimens 24.8 and 24.9 (T. fassoglense from the Waterberg, Limpopo) for marker MA8, and in specimens 5.11 and 5.19 (T. esculentum from Centurion, Gauteng) for marker MA9.

Amplification and Characterization of M10 primer pair

Amplification was successful in all *Tylosema* populations and in all specimens bar 5.16 (Figure 14). Primer pair M10 was the only marker that clearly showed length variation between *T. esculentum* and *T. fassoglense*. The M10 marker is almost 100 base pairs smaller in *T. esculentum* samples compared to *T. fassoglense*.

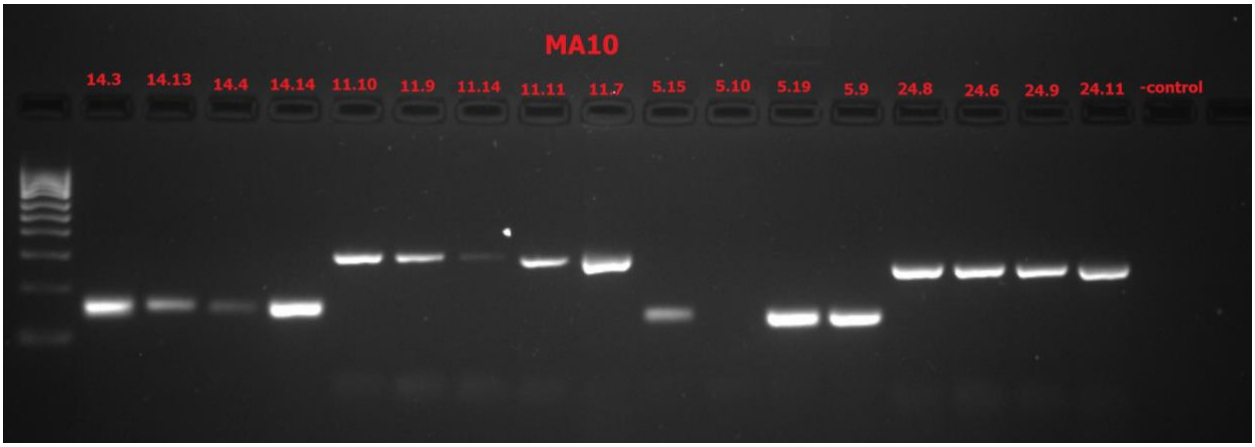

Figure 13 Gel image for MA10 primer pair marker. Clear size polymorphism between T. esculentum and T. fassoglense specimens.

All of the above results have summarized in table fomat (see Table 3 below).

Table 3 Summary of microsatellite amplification and polymorphism in four Tylosema *populations in South Africa (two* T. esculentum *and two* T. fassoglense*)*

	T. esculentum (LM5)		T. esculentum (LM14)		T. fassoglense (LM11)		T. fassoglense (LM24)	
	Amplification	**Polymorphism**	**Amplification**	**Polymorphism**	**Amplification**	**Polymorphism**	**Amplification**	**Polymorphism**
MA1	✓	✓	✓	✓	✓	✓	✓	✓
MA6	✓	×	✓	×	✓	×	✓	×
MA8	✓	×	✓	✓	✓	✓	✓	✓
MA9	✓	✓	✓	✓	✓	×	✓	×
MA10	✓	✓	✓	✓	✓	✓	✓	✓

Discussion

Cross species amplification

Since all five SSR markers were amplified successfully in *T. fassoglense*, given they were designed using *T. esculentum* genomic data, it is conceivable that they are amplifiable across the entire genus or beyond in various genera within Bauhinia s.l. Barbara et al. (2008) found that unsuccessful cross species amplification of SSR markers was greater in plants that are primarily self-fertilizing or annuals. Selfing plants have a lower effective population size (N_e), meaning they accumulate mutations at a higher rate than outcrossing plants (Higgins and Lynch, 2001). *Tylosema* species are likely obligate out crossers given that they are heterostylous (Hartley et al., 2003), and speciation events across the genus could have taken place relatively recently (Castro et al., 2006; Coetzer et al., 2011). Thus, the annealing sites for SSR primers are likely to be conserved in *Tylosema* species. The greater the evolutionary distance between taxa the more likely it is that cross amplification of markers will be unsuccessful, or that successfully amplified products are the result of homoplasy or convergent evolutionary events (Rossetto, 2001). Homoplasy of length, or size, in microsatellites can occur either when non- repeat motif point mutations or insertion/deletion events accumulate in loci or flanking regions; or when two microsatellites are identical in repeat length and sequence but do not share a genealogical origin (Selkoe and Toonen, 2006; Anmarkrud et al., 2008).

Detecting size homoplasy would require sequencing SSR markers, provided said homoplasy is the result of accumulated non-motif mutations (Garza and Freimer, 1996; Angers and Bernatchez, 1997). Convergent SSR alleles displaying homoplasy by size and sequence are harder to detect. However, size homoplasy in microsatellites is more likely to occur in repeat regions containing

compound or interrupted repeat motifs (Adams et al., 2004). The decision to only allow for the detection of perfect repeat regions in our marker selection parameters should mean that size homoplasy is unlikely to be present in the markers we have developed. Numerous authors have argued that the effects of homoplasy in SSR markers is negligible given the amount of genetic diversity homologous SSRs offer (Estoup et al., 2002; Li et al., 2002; Chapuis and Estoup, 2007), though for comparative genetic diversity studies between *Tylosema* species using SSR markers, the presence or absence of size homoplasy should be assessed on an locus to locus basis (Selkoe and Toone, 2006). Sequencing the SSR regions produced in this study could not be achieved considering the time and money it would require, and the rapid shut down of laboratory facilities in the wake of the COVID-19 outbreak.

Polyploidy in *Tylosema* and SSR analysis

As mentioned in Chapter 1, polyploidy in *Tylosema* species is present but not well understood. Various levels of ploidy have been found in *T. esculentum* e.g. tetraploids and supposed hexaploids (Managhan, 1995; Takundwa et al., 2012; Culls et al., 2019). However, conflicting chromosome counts and reports of ploidy level in *T. esculentum* are present in the literature. These uncertainties would require clarification. *T. esculentum* is a likely autopolyploid according to Monaghan and Hallorand (1996), but this too would require confirmation. *T. fassoglense* was described as a tetraploid by Goldblatt and Davidse (1977), though the extensive diversity within the species across its distribution range could mean that populations of varying ploidy levels are also present in this species (Catro et al., 2006; Coetzer et al., 2011).

This is important because the consequences of polyploidy on microsatellite analyses are significant and are influenced by the origins of said ploidy events, how much time has elapsed since

polyploidy has originated, and the exact level of ploidy present (Dufresne et al., 2014). In allopolyploids, recombination likely does not usually take place between loci from the two parental species that have hybridized, meaning alleles from each species segregate separately (Soltis and Soltis, 2009; Buggs et al., 2012). This is known as disomic inheritance, and the resulting effect is that one SSR marker can amplify for two independent loci, one from each parent species. Autopolyploids experience polysomic inheritance where recombination events can occur between all copies of a given locus (Ramsey and Schemske, 2002). As a result, the issue of partial heterozygotes is seen in autopolyploids. Partial heterozygotes can best be explained through an example: an autotetraploid at a given locus with three alleles can be heterozygous with three potential genotypes; AABC, ABBC, and ABCC. In each case the exact heterozygous genotype, or dosage of alleles, will be difficult to determine since only three bands would be detectable (Dufresne et al., 2014).

Issues of inheritance and allele number in polyploids can reduce the potential applications SSR markers have in population genetic analyses. One of the strengths of microsatellites over rival marker systems is that they are codominantly inherited, allowing for the detection of heterozygotes (Nybom et al., 2004). Polyploidy removes this ability to detect complete heterozygotes and the comparative strength SSR markers have over systems such as AFLPs and RAPDs is lost (Dufresne et al., 2014). However, SSR markers remain much cheaper, simpler, and more repeatable to apply than rival marker systems and have remained the favoured marker even in genetic diversity studies of polyploid organisms. For example, Saltonstall (2003) used SSR markers to detect a hexaploid lineage of the polyploid reed species *Phragmites australis* (Cav.) Steud. that is outcompeting indigenous lineages of the same species in North America. This was achieved by sequencing alleles and analysing genotypes based on the presence/absence of alleles in individuals. Garcia-

Verdugo et al. (2013) used SSR markers to show that colonization and spread of the octoploid tree *Prunus lucitanica* L. was predominantly mediated through clonal events. This was done by assessing the number of alleles present across populations sampled, and then determining the number of different genotypes per population by assessing the presence/absence of unique alleles. Additionally, various programs and R packages have been designed to analyse SSR markers in polyploids for the purpose of population genetics such as POLYSAT, PODIST, and TETRASAT, though many of these programs operate under the assumption of known ploidy level and inheritance (Markwith et al., 2006; Tomiuk et al., 2009; Clark and Jasieniuk, 2011).

Conclusion

We have shown in this study that SSR markers can easily be developed from genomic data using a simple program like MSATCOMMANDER. The markers we tested were mostly polymorphic and were all amplifiable in both *Tylosema* species. Although only five markers were tested, it is likely that polymorphic markers within the remaining 41 primer sets that were identified are present. Researchers should feel encouraged to test for further polymorphic markers in our set using gel electrophoresis as demonstrated in this study. This initial step is simple and relatively cost effective. The utilization of these markers for population or conservation genetics assessments in *Tylosema* does come with two caveats mentioned above: Size homoplasy and polyploidy. Although these issues can be worked around to some degree, the extent to which they would interfere with population genetic analysis in *Tylosema* is presently unclear. What these caveats do highlight is the amount of work still required in order to clarify the evolutionary history and species limits with the genus.

Chapter 4

Synopsis

The amount of work required to domesticate any one *Tylosema* species will be immense, a task which would only be complicated by the diversity seen within and between taxa as found in the wild. The threats placed upon African agriculture by climate change, such as reductions in crop yields, will have to be dealt with as swiftly as possible. Speeding up the domestication process of *Tylosema* could provide a solution to some of these issues, but a deep understanding of the evolutionary histories and current condition(s) of these species is essential. It would be naive to claim that those issues were sufficiently addressed in this dissertation, given the sheer size of the task at hand. However, the first steps to providing a molecular basis to the phylogenetic history of *Tylosema* have been taken, and an even more foundational effort to assess population level dynamics has been provided with the development of our SSR marker set.

Species limits between southern African *Tylosema* as they are currently recognized have been supported by our results. The recognition of *T. angolense* as a distinct species from *T. fassoglense* has significant knock-on effects on both the agricultural and conservation prospects of the two species. Though neither are currently considered candidates for domestication by any research group or nations (to our knowledge), their potential as crop plants is likely as high as *T. esculentum*.

We urge that researchers take note of this potential and consider the value these species could have in their agricultural sectors. Future taxonomic research should also take note of the high diversity in *T. fassoglense*, as it spans the length of sub-Saharan Africa and encompasses a wide range of

habitats and microclimates. We predict that hidden diversity in *T. fassoglense* is present and that five currently described species in *Tylosema* is an underestimation of diversity. The case for *T. esculentum* is complex given the results of this study. Kalahari and Highveld lineages within the species are present, which is somewhat intuitive given the dramatically different ecological pressures these two biomes exert. However, the nature of this lineage split and how it affects the taxonomy of the species is unclear. At present, we consider that the two lineages remain within a single species, though we stress that this conclusion is a tentative one. In what is becoming a common theme in *Tylosema*, we predict that greater species level diversity is present, potentially as a result of polyploidy, and that a deeper dive into genomic data will elucidate these relations.

T. esculetum remains the most popular candidate for domestication at present and this may remain the case for many researchers. The species' longstanding history as a popular edible plant, and its ability to grow in water stressed environments (at least the Kalahari lineage) has resulted in a bias in scientific investigation towards *T. esculentum* over other closely related species. It would likely then remain a popular potential crop in arid regions around the world. If this is to be the case, then the findings of this study suggest that South African *T. esculentum* be considered genetically distinct from its Namibian and Botswana counterparts. The two lineages may contain subtly distinct properties relevant not only to crop breeders, but also to food scientists aiming to determine alternative uses for the species beyond a direct source of food.

The microsatellite markers developed in this study could prove invaluable to population level studies undertaken in the future. Domestication projects will require means of tracking both genetic diversity and phenotypically favourable traits in *Tylosema* plants. SSR markers have standardly been used in agricultural practices for the past two decades, and although their popularity has waned somewhat in recent years, they remain an effective tool for determining

intraspecific diversity (Hodel et al., 2016). Domestication projects should consider the conservation status of *Tylosema* plants in the wild, considering these populations constitute the genetic source materials which they will utilize for cultivar development. Microsatellites have strong utility in this regard as well, and the set which we have developed could prove invaluable to future research.

Finally, it is our opinion that agricultural and naturalist interests in *Tylosema* need not be antagonistic, and that a great deal of synergy between the two approaches would likely yield benefits for both. A genus as poorly understood as *Tylosema* requires ecological and evolutionary attention in order to be protected. The anthropogenic benefits *Tylosema* species possess offers a means to rectify the neglect the genus has received over the decades. As the taxonomy of *Tylosema* becomes better understood, so too will the evolutionary events underpin these relations. This will in turn provide the framework from which plant breeders and agriculturalists will work in order to develop *Tylosema* crops to match the climates they are intended to grow in, and the markets they will eventually be sold in. Forthcoming research into *Tylosema* should contain elements of these two approaches and the understanding that the improvement of our understanding in one regard will be to the benefit of the other.

References

Adams, K.L., Wendel, J.F., 2005. Polyploidy and genome evolution in plants. Current Opinion in Plant Biology 8, 135–141.

Adams, R.I., Brown, K.M., Hamilton, M.B., 2004. The impact of microsatellite electromorph size homoplasy on multilocus population structure estimates in a tropical tree (*Corythophora alta*) and an anadromous fish (*Morone saxatilis*). Molecular Ecology 13, 2579–2588.

Allendorf, F.W., Hohenlohe, P.A., Luikart, G., 2010. Genomics and the future of conservation genetics. Nature Reviews Genetics 11, 697–709.

Álvarez, I., Wendel, J.F., 2003. Ribosomal ITS sequences and plant phylogenetic inference. Molecular Phylogenetics and Evolution 29, 417–434.

Angers, B., Bernatchez, L., 1997. Complex evolution of a salmonid microsatellite locus and its consequences in inferring allelic divergence from size information. Molecular Biology and Evolution 14, 230–238.

Anmarkrud, J.A., Kleven, O., Bachmann, L., Lifjeld, J.T., 2008. Microsatellite evolution: Mutations, sequence variation, and homoplasy in the hypervariable avian microsatellite locus *HrU10*. BMC Evolutionary Biology 8, 1–10.

Arnheim, N., Krystal, M., Schmickel, R., Wilson, G., Ryder, O., Zimmer, E., 1980. Molecular evidence for genetic exchanges among ribosomal genes on nonhomologous chromosomes in man and apes. Proceedings of the National Academy of Science USA 77, 7323–7327.

Ashley, M. V., 2010. Plant parentage, pollination, and dispersal: How DNA microsatellites have altered the landscape. Critical Reviews in Plant Sciences 29, 148–161.

Babineau, M., Gagnon, E., Bruneau, A., 2013. Phylogenetic utility of 19 low copy nuclear genes in closely related genera and species of caesalpinioid legumes. South African Journal of Botany 89, 94-105.

Baker, R.H., Yu, X., DeSalle, R., 1998. Assessing the relative contribution of molecular and morphological characters in simultaneous analysis trees. Molecular Phylogenetics and Evolution 9, 427-436.

Baldwin, B.G., Sanderson, M.J., Porter, J.M., Wojciechowski, M.F., Campbell, C.S., Donoghue, M.J., 1995. The ITS region of nuclear ribosomal DNA: A valuable source of evidence on angiosperm phylogeny. Annals of the Missouri Botanical Garden 82, 247–277.

Banks, H., Forest, F., Lewis, G., 2013. Palynological contribution to the systematics and taxonomy of *Bauhinia* s.l. (Leguminosae: Cercideae). South African Journal of Botany 89, 219-226.

Barbará, T., Palma-Silva, C., Paggi, G.M., Bered, F., Fay, M.F., Lexer, C., 2007. Cross-species transfer of nuclear microsatellite markers: Potential and limitations. Molecular Ecology 16, 3759–3767.

Barker, M.S., Arrigo, N., Baniaga, A.E., Li, Z., Levin, D.A., 2016. On the relative abundance of autopolyploids and allopolyploids. New Phytology 210, 391–398.

Bentley, L., Barker, N.P., Dold, A.P., 2015. Genetic diversity of the endangered *Faucaria tigrina* (Aizoaceae) through ISSR "fingerprinting" using automated fragment detection. Biochemical Systematics and Ecology 58, 156–161.

Besnard, G., Rubio De Casas, R., Vargas, P., 2007. Plastid and nuclear DNA polymorphism reveals historical processes of isolation and reticulation in the olive tree complex (*Olea europaea*). Journal of Biogeography 34, 736–752.

Blears, M.J., De Grandis, S.A., Lee, H., Trevors, J.T., 1998. Amplified fragment length polymorphism (AFLP): A review of the procedure and its applications. Journal of Industrial Microbiology and Biotechnology 21, 99–114.

Botstein, D., White, R.L., Skolnick, M., Davis, R.W., 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. American Journal of Human Genetics 32, 314-331

Bower, N., Hertel, K., Storey, R., 1988. Nutriniola evaluation of marame bean (*Tylosema esculentum*, Fabaceae): Analysis of the seed. Economic Botany 42, 533-540.

Brohede, J., Ellegren, H., 1999. Microsatellite evolution: Polarity of substitutions within repeats and neutrality of flanking sequences. Proceedings of the Royal Society B: Biological Sciences 266, 825–833.

Brummitt, R.K. and Ross, J.H., 1976. A note on a *Tylosema fassoglense* (Leguminosae-Caeselpinioideae) from southern Africa. Kew bulletin 31, 219-220.

Bruneau, A., Mercure, M., Lewis, G.P., Herendeen, P.S., 2008. Phylogenetic patterns and diversification in the caesalpinioid legumes. Botany 86, 697-718.

Buckler, E.S., Ippolito, A., Holtsford, T.P., 1997. The evolution of ribosomal DNA: Divergent paralogous and phylogenetic implications. Genetics 145, 821–832.

Buggs, R.J.A., Renny-Byfield, S., Chester, M., Jordon-Thaden, I.E., Viccini, L.F., Chamala, S., Leitch, A.R., Schnable, P.S., Bradley Barbazuk, W., Soltis, P.S., Soltis, D.E., 2012. Next-generation sequencing and genome evolution in allopolyploids. American Journal of Botany 99, 372–382.

Bull, L.N., Pabón-Peña, C.R., Freimer, N.B., 1999. Compound microsatellite repeats: Practical and theoretical features. Genome Research 9, 830–838.

Burchell, W.J., 1822. Travels in the interior of southern Africa Vol 2. Reprinted 1953. London, Batchworth Press, 589

Buschiazzo, E., Gemmell, N.J., 2006. The rise, fall and renaissance of microsatellites in eukaryotic genomes. BioEssays 28, 1040–1050.

Butler, J.M., 2005. Constructing STR multiplex assays. Methods in Molecular Biology 297, 53–66.

Camargo, A., Morando, M., Avila, L., Sites, J., 2012. Species delimitation with abc and other coalescent-based methods: a test of accuracy with simulations and an empirical example. Evolution 66, 2834–2849.

Carstens, B.C., Dewey, T.A., 2010. Species delimitation using a combined coalescent and information-theoretic approach: An example from North American myotis bats. Systematic Biology 59, 400–414.

Carstens, B.C., Pelletier, T.A., Reid, N.M., Satler, J.D., 2013. How to fail at species delimitation. Molecular Ecology 22, 4369–4383.

Castro, S., Silveira, P., Coutinho, A.P., Figueiredo, E., 2005. Systematic studies in *Tylosema* (Leguminosae). Botanical Journal of the Linnean Society 147, 99-115.

CBOL Plant Working Group, 2009. A DNA barcode for land plants. Proceedings of the National Academy of Science USA 106, 12794–12797.

Chapuis, M.P., Estoup, A., 2007. Microsatellite null alleles and estimation of population differentiation. Molecular Biology and Evolution 24, 621–631.

Chase, M., Kesseli, R., Bawa, K., 1996. Microsatellite markers for population and conservation genetics of tropical trees. American Journal of Botany 83, 51–57.

Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., Zhu, Y., Ma, X., Gao, T., Pang, X., Luo, K., Li, Y., Li, X., Jia, X., Lin, Y., Leon, C., 2010. Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. PLoS One 5, 1–8.

Chimwamurombe, C., 2011. Domestication of [*Tylosema esculentum* (Burchell) Schreiber] (Marama bean): A work in progress in Namibia. Bioscience Biotechnology Research Asia 8, 549-556.

Chimwamurombe, P.M., 2010. Domestication of *Tylosema esculentum* (marama bean) as a cropfor Southern Africa: Genetic diversity of the Omitara marama subpopulation of Namibia. Aspects of Applied Biology 96, 37-43.

Chingwaru, W., Faria, M.L., Saravia, C., Cencis, A., 2007. Indigenous knowledge of health benefits of morama plant among respondents in Ghantsi and Jwaneng of Botswana. African Journal of Food, Agriculture, Nutrition and Development 7, 1-3.

Chingwaru, W., Majinda, R.T., Yeboah, S.O., Jackson, J.C., Kapewangolo, P.T., Kandawa-Schulz, M., Cencic, A., 2011. *Tylosema esculentum* (Marama) tuber and bean extracts are strong antiviral agents against rotavirus infection. Evidence-Based Complementary and Alternative Medicine 1, 1- 11.

Choi, H.K., Luckow, M.A., Doyle, J., Cook, D.R., 2006. Development of nuclear gene-derived molecular markers linked to legume genetic maps. Molecular Genetics and Genomics 276, 56–70.

Clark, L. V., Jasieniuk, M., 2011. polysat: An R package for polyploid microsatellite analysis. Molecular Ecology Resoures 11, 562–566.

Coetzer L.A., Robbertse, P.J., Grobbelaar, N.,1981. Morfologie van die Sporoderm van *Tylosema esculentum* en *T. fassoglense*. Journal of South African Botany 47, 769–781.

Coetzer, L.A., Ross, J.H., 1977. *Tylosema*. Flora of southern Africa 16, 61–64.

Coetzer, L.A., Van Wyk, A.E., Buitendag, E., 2011. *Tylosema fassoglense*. Flowering Plants of Africa 62, 70-79.

Cullis, C., Kunert, K., 2017. Unlocking the potential of orphan legumes. Journal of Experimental Botany 68, 1895-1903.

Cullis, C., Lawlor, D.W., Chimwamurombe, P., Bbebe, N., Kunert, K., Vorster, J., 2019. Development of marama bean, an orphan legume, as a crop. Food and Energy Security 3, 1-3.

Cummings, M.P., Otto, S.P., Wakeley, J., 1995. Sampling properties of DNA sequence data in phylogenetic analysis. Molecular Biology and Evolution 12, 814–822.

Dakora, F.D., Lawlor, D.W., Sibuga, K.P., 1999. Assessment of symbiotic nitrogen nutrition in marama bean (*Tylosema esculentum* L.), a tuber-producing underutilized African grain legume. Symbiosis 27, 269-277.

Dayrat, B., 2005. Towards integrative taxonomy. Biol. J. Linn. Soc. 85, 407–415.

De Frey, H.M., Coetzer, L.A., Robbertse, P.J., 1992. A unique anther-mucilage in the pollination biology of *Tylosema esculentum.* Sexual Plant Reproduction 5, 298-303.

De Wit, H.C.D. 1956. A revision of the Malesian Bauhinieae. Reinwardtia, 3, 381–541.

Degnan, J.H., Rosenberg, N.A., 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2, 762–768.

Dmitriev, D.A., Rakitov, R.A., 2008. Decoding of superimposed traces produced by direct sequencing of heterozygous indels. PLoS Computational Biology 4, e1000113.

Dover, G.A., 1982. Molecular drive: A cohesive mode of species evolution. Nature (London) 299, 111–117.

Doyle, J.J. and Doyle, J.L., 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemistry Bulletin 19, 11–15

Doyle, J.J., 2012. Polyploidy in legumes, in: Polyploidy and Genome Evolution. Springer-Verlag Berlin Heidelberg, pp. 147–180.

Drummond, A.J., Suchard, M.A., Xie, D., Rambaut, A., 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution 29, 1969–1973.

Dufresne, F., Stift, M., Vergilino, R., Mable, B.K., 2014. Recent progress and challenges in population genetics of polyploid organisms: An overview of current state-of-the-art molecular and statistical tools. Molecular Ecology 23, 40–69.

Eckert, A.J., Carstens, B.C., 2008. Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow. Molecular Phylogenetics and Evolution 49, 832–842.

Edwards, S. V., 2008. Is a new and general theory of molecular systematics emerging? Evolution 63, 1–19.

Ellegren, H., 2004. Microsatellites: Simple sequences with complex evolution. Nature Reviews Genetics 5, 435–445.

Eric Schranz, M., Mohammadin, S., Edger, P.P., 2012. Ancient whole genome duplications, novelty and diversification: The WGD Radiation Lag-Time Model. Current Opinions in Plant Biology 15, 147–153.

Estoup, A., Jarne, P., Cornuet, J., 2002. Homoplasy and mutation model at microsatellite loci and. Molecular Ecology 11, 1591–1604.

Faircloth, B.C., 2008. MSATCOMMANDER: Detection of microsatellite repeat arrays and automated, locus-specific primer design. Molecular Ecology Resources 8, 92–94.

Faria, M., Mabaya, E., Jordaan, D., 2011. Markets for marama beans in southern Africa: Linking sustainable products with sustainable livelihoods. Development Southern Africa 28, 477-492.

Fazekas, A.J., Burgess, K.S., Kesanakurti, P.R., Graham, S.W., Newmaster, S.G., Husband, B.C., Percy, D.M., Hajibabaei, M., Barrett, S.C.H., 2008. Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. PLoS One 3, e2802.

Fazekas, A.J., Kesanakurti, P.R., Burgess, K.S., Percy, D.M., Graham, S.W., Barrett, S.C.H., Newmaster, S.G., Hajibabaei, M., Husband, B.C., 2009. Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? Molecular Ecology Resourses 9, 130–139.

Feliner, G.N., Rosselló, J.A., 2007. Better the devil you know? Guidelines for insightful utilization of nrDNA ITS in species-level evolutionary studies in plants. Molecular Phylogenetics and Evolution 44, 911–919.

Felsenstein, J., 2006. Accuracy of coalescent likelihood estimates: Do we need more sites, more sequences, or more loci? Molecular Biology and Evolution 23, 691–700.

García-Verdugo, C., Calleja, J.A., Vargas, P., Silva, L., Moreira, O., Pulido, F., 2013. Polyploidy and microsatellite variation in the relict tree *Prunus lusitanica* L.: How effective are refugia in preserving genotypic diversity of clonal taxa? Molecular Ecology 22, 1546–1557.

Gardner, M.G., Fitch, A.J., Bertozzi, T., Lowe, A.J., 2011. Rise of the machines - recommendations for ecologists when using next generation sequencing for microsatellite development. Molecular Ecology Resources 11, 1093–1101.

Garza, J.C., Freimer, N.B., 1996. Homoplasy for size at microsatellite loci in humans and chimpanzees. Genome Research 6, 211–217.

Gbetibouo, G.A., Ringler, C., Hassan, R., 2010. Vulnerability of the South African farming sector to climate change and the variability: An indicator approach. Natural Resource Forum 34, 175-187.

GBIF.org (20 April 2018) GBIF Occurrence Download https://doi.org/10.15468/dl.ycg764

Gene Codes Corp, 2005. Sequencher [computer program]. Version 4.5. Ann Arbor, MI: Gene Codes Corporation. Available from: http://www.genecodes.com/.

Godwin, I.D., Aitken, E.A.B., Smith, L.W., 1997. Applications of inter simple sequence repeat (ISSR) markers to plant genetics. Electrophoresis 18, 1524-1528.

Goldblatt P., 1981. Cytology and phylogeny of Leguminosae. Advances in Legume Systematics 427–463.

Goldblatt P., Davidse G., 1977. Chromosome numbers in legumes. Annals of the Missouri Botanical Garden 64,121–128.

Graybeal, A., 1998. Is it better to add taxa or characters to a difficult phylogenetic problems? Systematic Biology 47, 9–17.

Guichoux, E., Lagache, L., Wagner, S., Chaumeil, P., Léger, P., Lepais, O., Lepoittevin, C., Malausa, T., Revardel, E., Salin, F., Petit, R.J., 2011. Current trends in microsatellite genotyping. Molecular Ecology Resources 11, 591–611.

Gupta, M., Chyi, Y.S., Romero-Severson, J., Owen, J.L., 1994. Amplification of DNA markers from evolutionarily diverse genomes using single primers of simple-sequence repeats. Theoretical and Applied Genetics 89, 998–1006.

Hao, G., Zhang, D., Zhang, M., Guo, L., Li, S., 2003. Phylogenetics of *Bauhinia* subgenus *Phanera* (Leguminosae: Caesalpinioideae) based on ITS sequences of nuclear ribosomal DNA. Botanical Bulletin of Academia Sinica 44, 223-228.

Hartley, M.L., Tshamekeng, E., Thomas, S.M., 2002. Funtional Heterostyly in *Tylosema esculentum* (Caesalpinioideae). Annals of Botany 89, 67-76.

Hasegawa, M., Kishino, H., Yano, T. aki, 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution 22, 160–174.

Hassan, R., Nhemachena, C., 2008. Determinants of African farmers' strategies for adapting to climate change: Multinomial choice analysis. Relational Responsibility: Resources for Sustainable Dialogue 2, 83-104.

Herbert, P.D.N., Cywinska, A., Ball, S.L., DeWaard, J.R., 2003. Biological identifications through DNA barcodes. Proceedings of the Royal Society B: Biological Sciences 270, 313–321.

Hedrick, P.W., 2001. Conservation genetics: Where are we now? Trends in Ecology and Evolution 16, 629–636.

Heled, J., Drummond, A.J., 2010. Bayesian Inference of Species Trees from Multilocus Data. Molecular Biology and Evolution 27, 570–580.

Hickey, L.T., Hafeez, A.N., Robinson, H., Jackson, S.A., Leal-Bertioli, S.C.M., Tester, M., Gao, C., Godwin, I.D., Hayes, B.J., Wulff, B.B.H., 2019. Breeding crops to feed 10 billion. Nature Biotechnology 37, 744-754.

Higgins, K., Lynch, M., 2001. Metapopulation extinction caused by mutation accumulation. Proceedings of the National Academy of Science USA 98, 2928–2933.

Hillis, D.M., 1995. Approaches for assessing phylogenetic accuracy. Systematic Biology 47, 3-8.

Hodel, R.G.J., Segovia-Salcedo, M.C., Landis, J.B., Crowl, A.A., Sun, M., Liu, X., Gitzendanner, M.A., Douglas, N.A., Germain-Aubrey, C.C., Chen, S., Soltis, D.E., Soltis, P.S., 2016. The Report of My Death was an Exaggeration: A Review for Researchers Using Microsatellites in the 21st Century. Applications in Plant Sciences 4, 1600025.

Hollingsworth, P.M., Graham, S.W., Little, D.P., 2011. Choosing and using a plant DNA barcode. PLoS One 6, 1-13.

Hollingsworth, P.M., Li, D.Z., Van Der Bank, M., Twyford, A.D., 2016. Telling plant species apart with DNA: From barcodes to genomes. Philosophical Transactions of the Royal Society B: Biological Science 371.

Holse, M., Husted, S., Hansen, A., 2010. Chemical composition of macrame bean (*Tylosema esculentum*)- A wild African bean with unexploited potential. Journal of Food Composition and Analysis 23, 648-657.

Hudson, R.R., Turelli, M., 2003. Stoachsticty overrules the "three-times rule": genetic drift, genetic draft, and coalescence times for nuclear loci versus mitochondrial DNA. Evolution 57, 182–190.

Huelsenbeck J.P. and Ronquist, F., 2001. MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754-755.

Huelsenbeck, J.P. and Ronquist, F., 2003. MyBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572-1574.

Inamura, A., Ohashi, Y., Sato, E., Yoda, Y., Masuzawa, T., Ito, M., Yoshinaga, K., 2000. Intraspecific sequence variation of chloroplast DNA reflecting variety and geographical distribution of *Polygonum cuspidatum* (Polygonaceae) in Japan. Journal of Plant Research 113, 419–426.

Jackson, J.C., Duodo, K.G., Holse, M., Lima de Faria, M.D., Jordaan, W.C., Hansen, A., Cencic, A., Kandawa-Schultz, M., Mpotokwane, S.M., Chimwamurombe, P., de Kock, H.L., Minnaar, A., 2010. The Morama Bean (*Tylosema esculentum*): A Potential Crop for Southern Africa. Advances in Food and Nutrition Research, 61, 187-246.

Jackson, N.D., Carstens, B.C., Morales, A.E., O'Meara, B.C., 2017. Species delimitation with gene flow. Systematic Biology 66, 799–812.

Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., Landherr, L., Ralph, P.E., Tomsho, L.P., Hu, Y., Liang, H., Soltis, P.S., Soltis, D.E., Clifton, S.W., Schlarbaum, S.E., Schuster, S.C., Ma, H., Leebens-Mack, J., Depamphilis, C.W., 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473, 97–100.

Johnson, M.G., Pokorny, L., Dodsworth, S., Botigué, L.R., Cowan, R.S., Devault, A., Eiserhardt, W.L., Epitawalage, N., Forest, F., Kim, J.T., Leebens-Mack, J.H., Leitch, I.J., Maurin, O., Soltis, D.E., Soltis, P.S., Wong, G.K.S., Baker, W.J., Wickett, N.J., 2019. A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering. Systematic Biology 68, 594–606.

Jones, C.J., Edwards, K.J., Castaglione, S., Winfield, M.O., Sala, F., Van De Wiel, C., Bredemeijer, G., Vosman, B., Matthes, M., Daly, A., Brettschneider, R., Bettini, P., Buiatti, M., Maestri, E., Malcevschi, A., Marmiroli, N., Aert, R., Volckaert, G., Rueda, J., Linacero, R., Vazquez, A., Karp, A., 1997. Reproducibility testing of RAPD, AFLP and SSR markers in plants by a network of European laboratories. Molecular Breeding 3, 381–390.

Kalia, R.K., Rai, M.K., Kalia, S., Singh, R., Dhawan, A.K., 2011. Microsatellite markers: An overview of the recent progress in plants. Euphytica 177, 309–334.

Karamanos, J., and Travlos, I.S., 2012. The Water Relations and Some Drought Tolerance Mechanisms of the Marama Bean. Agronomy Journal 104, 65-72.

Karp, A., Edwards, K.J., Bruford, M., Funk, S., Vosman, B., Morgante, M., Seberg, O., Kremer, A., Boursot, P., Arctander, P., Tautz, D., Hewitt, G.M., 1997. Molecular technologies for biodiversity evaluation: Opportunities and challenges: New technologies for detecting variation in dna complement traditional methods in biodiversity. Nature Biotechnology 15, 625–628.

Kayitesi, E., de Kock, H.L., Minnaar, A., Duodu, K.G., 2012. Nutritional quality and antioxidant activity of maraca-sorghum composite flours and porridges. Food Chemistry 131, 837-842.

Keagan, A., Von Staden, J., 1981. Marama bean, *Tylosema esculentum*, a plant worthy of cultivation. South African Journal of Science 77, 87.

Keith, M.E., Renew, A., 1975. Notes of some edible wild plants found in the Kalahari. Koedoe 18, 1-12.

Ketshajwang, K.K., Holmback, J., Yeboah, S.O., 1998. Quality and compositional studies of some edible leguminosae seed oils in Botswana. Journal of the American Oil Chemists' Society 75, 741-743.

Kim, Y., Cullis, C., 2017. A novel inversion in the chloroplast genome of marama (*Tylosema esculentum*). Journal of Experimental Botany 68, 2065–2072.

Kluge, A.G., 2004. On total evidence: for the record. Cladistics 20, 205-207.

Kolaczkowski, B., Thornton, J.W., 2004. Performance of maximum parsimony and likelihood phylogenetic when evolution is heterogeneous. Nature 431, 980-984.

Kress, W.J., Erickson, D.L., 2007. A Two-Locus Global DNA Barcode for Land Plants: The Coding rbcL Gene Complements the Non-Coding *trn*H-*psb*A Spacer Region. PLoS One 2, e508.

Kress, W.J., Wurdack, K.J., Zimmer, E.A., Weigt, L.A., Janzen, D.H., 2005. Use of DNA barcodes to identify flowering plants. Proc. Natl. Acad. Sci. U. S. A. 102, 8369–8374.

Kubatko, L.S., Degnan, J.H., 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Systematic Biology 56, 17–24.

Landis, J.B., Soltis, D.E., Li, Z., Marx, H.E., Barker, M.S., Tank, D.C., Soltis, P.S., 2018. Impact of whole-genome duplication events on diversification rates in angiosperms. American Journal of Botany 105, 348–363.

Larsen, B.B., Miller, E.C., Rhodes, M.K., Wiens, J.J., 2017. Inordinate fondness multiplied and redistributed: The number of species on earth and the new pie of life. Quarterly Review of Biology 92, 229–265.

Leaché, A.D., Zhu, T., Rannala, B., Yang, Z., 2019. The Spectre of Too Many Species. Systematic Biology 68, 168–181.

Levinson, G., Gutman, G.A., 1987. High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. Nucleic Acids Research 15, 5323–5338.

Lewis, G.P., and Forest, F. 2005. Cercideae. In Legumes of the world. Edited by G. Lewis, B. Schrire, B. Mackinder, and M. Lock. Royal Botanic Gardens, Kew, UK, 57–67.

Li, M., Wunder, J., Bissoli, G., Scarponi, E., Gazzani, S., Barbaro, E., Saedler, H., Varotto, C., 2008. Development of COS genes as universally amplifiable markers for phylogenetic reconstructions of closely related plant species. Cladistics 24, 727–745.

Li, Y.C., Korol, A.B., Fahima, T., Beiles, A., Nevo, E., 2002. Microsatellites: Genomic distribution, putative functions and mutational mechanisms: A review. Molecular Ecology 11, 2453-2465.

Liu, S., Cornille, A., Decroocq, S., Tricon, D., Chague, A., Eyquard, J.P., Liu, W.S., Giraud, T., Decroocq, V., 2019. The complex evolutionary history of apricots: Species divergence, gene flow and multiple domestication events. Molecular Ecology 28, 5299–5314.

LPWG, 2017. A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny. Taxon 66, 44-77

Maddison, W., Knowles, L., 2006. Inferring phylogeny despite incomplete lineage sorting. Systematic Biology 55, 21–30.

Maddison, W.P., 1997. Gene trees in species trees. Systematic biology 46, 523-536.

Mahgoub, S.E.O., Mthombeni, F.M., Maswabi, E.B., Jackson, J.C., 2013. Consumers' knowledge and perception on utilization of the Morama bean (*Tylosema esculentum*) in Botswana. International Journal of Consumer Studies 3, 265-270.

Mallet, J., Willmott, K., 2003. Taxonomy: renaissance or Tower of Babel? Trends in Ecology and Evolution 18, 63-65.

Mangenot S., Mangenot, G.,1958. Deuxieme list de nombres chromosomiques nouveaux chez diverses dicotyledones et monocotyledones d'Afrique Occidentale. Bulletin Jardin Botanique de l'État 28, 315-329.

Mariette, S., Le Corre, V., Austerlitz, F., Kremer, A., 2002. Sampling within the genome for measuring within-population diversity: Trade-offs between markers. Molecular Ecology 11, 1145–1156.

Markwith, S.H., Stewart, D.J., Dyer, J.L., 2006. TETRASAT: A program for the population analysis of allotetraploid microsatellite data. Molecular Ecology Notes 6, 586–589.

Maruatona, G.N., Duodu, K.G., Minnaar, A., 2010. Physiochemical, nutrinional and functional properties of marama bean flour. Food Chemistry 121, 400-405.

Meyer, C.P., Paulay, G., 2005. DNA barcoding: Error rates based on comprehensive sampling. PLoS Biology 3, 1–10.

Miller, M.A., Schwartz, T., Pickett, B.E., He, S., Klem, E.B., Scheuermann, R.H., Passarotti, M., Kaufman, S., Oleary, M.A., 2015. A RESTful API for access to phylogenetic tools via the CIPRES science gateway. Evolutionary Bioinformatics 11, 43–48.

Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., S. Swenson, M., Warnow, T., 2014. ASTRAL: Genome-scale coalescent-based species tree estimation. Bioinformatics 30, 541–548.

Mitchell, R.A., Keys, A.J., Madgwick, P.J., Parry, M.A., Lawlor, D.W., 2005. Adaptationof photosynthesis in marama bean *Tylosema esculentum* (Burchell A Scheiber) to a high temperature, high radiation, drought-prone environment. Plant Physiology and Biochemistry 43, 969-976.

Mmonatau, Y., 2005. Flour from the morama bean: composition and sensory properties in a Botswana perspective. MSc thesis. University of Stellenbosch, South Africa.

Mohamad, A., Alhasnawi, A.N., Kadhimi, A.A., Isahak, A., Yusoff, W.M.W., Radziah, C.M.Z.C., 2017. DNA isolation and optimization of ISSR-PCR reaction system in *Oryza sativa* L. Internation Journal on Advanced Science, Engineering and Information Technology 7, 2264–2272.

Monaghan, B.G., 1995. Genetic variation in momma bean (*Tylosema esculentum*). MSc. (Agric.) thesis University of Melbourne, Melbourne.

Monaghan, B.G., and Halloran, G.M., 1996. RAPD variation within and between natural populations of morama [*Tylosema esculentum* (Burchell) Schreiber] in Southern Africa. South African Journal of Botany 62, 287-291.

Mondini, L., Noorani, A., Pagnotta, M.A., 2009. Assessing plant genetic diversity by molecular tools. Diversity 1, 19-35.

Moore, H., Greenwell, P.W., Liu, C.P., Arnheim, N., Petes, T.D., 1999. Triplet repeats form secondary structures that escape DNA repair in yeast. Proceedings of the National Academy of Science of the USA 96, 1504–1509.

Moore, W. S. 1995. Inference of phylogenies from mtDNA variation: mitochondiral gene tree versus nuclear-gene trees. Evolution 49: 718–726.

Mossel, E., Vigoda, E., 2005. Evolution: Phylogenetic MCMC algorithms are misleading on mixtures of trees. Science 309, 2207–2209.

Müller, C., Cramer, W., Hare, W., Lotze-Campen, H., 2011. Climate change risk for African agriculture. Proceedings of the National Academy of Sciences of the United States of America 108, 4313-4315.

Müseler, D., Schönfeld, H.C., 2006. The nutrient content of the marama bean (*Tylosema esculentum*), an underutilized legume from southern Africa. Agricola, 2-8.

Müseler, D.L., 2005. Evaluation of the quality characteristics of the marama bean (*Tylosema esculentum*), an underutilized grain and tuber producing legume in southern Africa. MSc thesis. University of Namibia, Namibia.

Nabhan, A.R., Sarkar, I.N., 2012. The impact of taxon sampling on phylogenetic inference: A review of two decades of controversy. Briefings in Bioinformatics 13, 122–134.

Nadeem, M.A., Nawaz, M.A., Shahid, M.Q., Doğan, Y., Comertpay, G., Yıldız, M., Hatipoğlu, R., Ahmad, F., Alsaleh, A., Labhane, N., Özkan, H., Chung, G., Baloch, F.S., 2018. DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. Biotechnology and Biotechnological Equipment 32, 261–285.

Naidoo, K., Steenkamp, E.T., Coetzer, M.P.A., Wingfield, M.J., Wingfield, B.D., 2013. Concerted Evolution in the Ribosomal RNA Cistron. PLoS One 8, e59355.

Nathans, D., Smith, H., 1975. Restriction endonucleases in the analysis and restructuring of DNA molecules. Annual Review of Biochemistry 44,273- 293.

National Research Council, 2006. Marama. In Lost crops of Africa: Volume II: Vegetables. National Academy Press, Washington, DC, 234-244.

Nepolo, E., Takundwa, M., Chimwamurombe, P.M., Cullis, C.A., and Kunert, K., 2009. A review of geographical distribution of marama bean [Tylosema esculentum (Burchell) Schreiber] and genetic diversity in the Namibian germplasm. African Journal of Biotechnology 8, 2088-2093.

Nhemachena, C., Hassan, R., 2007. Micro-level analysis of farmers' adaptation to climate change in southern Africa. In International Food Policy Research Institute.

Nichols, G., 2001. Gene trees and species trees are not the same. Trends in ecology and evolution 16, 358-364.

Nock, C.J., Waters, D.L.E., Edwards, M.A., Bowen, S.G., Rice, N., Cordeiro, G.M., Henry, R.J., 2011. Chloroplast genome sequences from total DNA for plant identification. Plant Biotechnology Journal 9, 328–333.

Nybom, H., 2004. Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. Molecular Ecology 13, 1143–1155.

Nyembwe, P., Minnaar, A., Duodu, K.G., de Kock, H.L., 2015. Sensory and physiochemical analyses of roasted marama beans [*Tylosema esculentum* (Burchell) A. Schreiber] with specific focus on compounds that may contribute to bitterness. Food Chemistry 178, 45-51.

Nyembwe, P.M., de Kock, H.L., Taylor, J.R.N., 2018. Potential of defatted marama flour-cassava starch composites to produce functional gluten-free bread-type dough. Lwt - Food Science and Technology 92, 429–434.

Padial, J.M., Miralles, A., De la Riva, I., Vences, M., 2010. The integrative future of taxonomy. Frontiers in Zoology 7, 1–14.

Pamilo, P. and M. Nei. 1988. Relationships between gene trees and species trees. Molecular Biology and Evolution 5, 568–583.

Pejic, I., Ajmone-Marsan, P., Morgante, M., Kozumplick, V., Castiglioni, P., Taramino, G., Motto, M., 1998. Comparative analysis of genetic similarity among maize inbred lines detected by RFLPs, RAPDs, SSRs, and AFLPs. Theoretical and Applied Genetics 97, 1248–1255.

Petit, R.J., Excoffier, L., 2009. Gene flow and species delimitation. Trends in Ecology and Evolution 24, 386–393.

Petit, R.J., Pineau, E., Demesure, B., Bacilieri, R., Ducousso, A., Kremer, A., 1997. Chloroplast DNA footprints of postglacial recolonization by oaks. Proceedings of the National Academy of Science USA 94, 9996–10001.

Petit, R.J., Vendramin, G.G., 2007. Plant phylogeography based on organelle genes: an introduction. In Phylogeography of southern European refugia: Evolutionary perspectives on the origins and conservation of European biodiversity. Edited by Weiss, S., Ferrand, N., Springer, 23-97.

Poczai, P., Hyvönen, J., 2010. Nuclear ribosomal spacer regions in plant phylogenetics: Problems and prospects. Molecular Biology Reports 37, 1897–1912.

Potts, B.M., Reid, J.B., 1988. Hybridization as a Dispersal Mechanism. Evolution 42, 1245.

Powell, A.M., 1987. Marama bean (*Tylosema esculentum*, Fabaceae) seed crop in Texas. Economic Botany 41, 216-220.

Powell, W., Morgante, M., Andre, C., Hanafey, M., Vogel, J., Tingey, S., Rafalski, A., 1996. The comparison of RFLP, RAPD, AFLP, and SSR (microsatellite) markers for germplasm analysis. Molecular Breeding 2, 225-238.

Rambaut, A., 2009. Figtree. Available at: http://tree.bio.ed.ac.uk/ software/figtree/.

Ramsey, J., Schemske, D.W., 2002. Neopolyploidy in flowering plants. Annual Review of Ecology and Systematics 33, 589–639.

Rasmussen, M.D., Kellis, M., 2007. Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. Genome Research 17, 1932–1942.

Rautenberg, A., Hathaway, L., Oxelman, B., Prentice, H.C., 2010. Geographic and phylogenetic patterns in *Silene* section *Melandrium* (Caryophyllaceae) as inferred from chloroplast and nuclear DNA sequences. Molecular Phylogenetics and Evolution 57, 978–991.

Razafimandimbison, S.G., Kellogg, E.A., Bremer, B., 2004. Recent origin and phylogenetic utility of divergent ITS putative pseudogenes: A case study from *Naucleeae* (Rubiaceae). Systematic Biology 53, 177-192.

Reddy, M.P., Sarla, N., Siddiq, E.A., 2002. Inter simple sequence repeat (ISSR) polymorphism and its application in plant breeding. Euphytica 128, 9–17.

ribosomal RNA genes for phylogenetics. In M. Innis, D. Gelfand, J. Sninsky, and T. White [eds.], PCR protocols: a guide to methods and applications, 315–322. Academic Press, San Diego, California, USA

Riesenberg, L.H., Soltis, D.E., 1991. Phylogenetic Consequences of Cytoplasmic. Evolutionary Trends in Plants 5, 65–84.

Rosenberg, N.A., 2002. The Probability of Topological Concordance of Gene Trees and Species Trees. Theoretical Population Biology 61, 225–247.

Ross, M., 1990. Sexual asymmetry in hermaphroditic plants. Trends in Ecology and Evolution 5, 43-47.

Rossetto, M., 2001. Sourcing of SSR markers from related plant species. In Plant genotyping: The DNA fingerprinting of plants. Edited by Henry, R.J. CABI Publishing, Wallingford, 211-224.

Russell, J.R., Fuller, J.D., Macaulay, M., Hatz, B.G., Jahoor, A., Powell, W., Waugh, R., 1997. Direct comparison of levels of genetic variation among barley accessions detected by RFLPs, AFLPs, SSRs and RAPDs. Theoretical and Applied Genetics 95, 714–722.

Saltonstall, K., 2003. Microsatellite variation within and among North American lineages of *Phragmites australis*. Molecular Ecology 12, 1689–1702.

Sanchez de la Hoz, M.P., Davila, J.A., Loarce, Y., Ferrer, E.,1996. Simple sequence repeat primers used in polymerase chain reaction amplifications to study genetic diversity in barley. Genome 39, 112-117.

Sanderson, M.J., McMahon, M.M., 2007. Inferring angiosperm phylogeny from EST data with widespread gene duplication. BMC Evolutionary Biology 7, 1–14.

Sang, T., 2002. Utility of low-copy nuclear gene sequences in plant phylogenetics. Critical Reviews in Biochemical Molecular Biology 37, 121–147.

Sang, T., Crawford, D.J., Stuessy, T.F., 1997. Chloroplast DNA phylogeny, reticulate evolution, and biogeography of *Paeonia* (Paeoniaceae). American Journal of Botany 84, 1120-1136.

Santana, Q.C., Coetzer, M.P.A., Steenkamp, E.T., Mlonyeni, O.X., Hammond, G.N.A., Wingfield, M.J., Wingfield, B.D., 2009. Microsatellite discovery by deep sequencing of enriched genomic libraries. Biotechniques 46, 217–223.

Schery, R.W. 1951. Leguminosae. Part 2. In Flora of Panama. Edited by R.E. Woodson, R.E. Schery and Collaborators. Annals of the Missouri Botanical Garden 38, 301–394.

Schlotterer, C., 2000. Evolutionary dynamics of microsatellite DNA. Chromosoma 109, 365–371.

Schmitz A., 1973. Contribuition palynologique à la taxonomie des Bauhinieae (Caesalpiniaceae). Bulletin du Jardin Botanique National de Belgique 43, 369–423.

Selkoe, K.A., Toonen, R.J., 2006. Microsatellites for ecologists: A practical guide to using and evaluating microsatellite markers. Ecology Letters 9, 615-629.

Sharma, P.C., Grover, A., Kahl, G., 2007. Mining microsatellites in eukaryotic genomes. Trends in Biotechnology 25, 490–498.

Shaw, J., Lickey, E.B., Beck, J.T., Farmer, S.B., Liu, W., Miller, J., Siripun, K.C., Winder, C.T., Schilling, E.E., Small, R.L., 2005. The tortoise and the hare II: Relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. American Journal of Botany 92, 142–166.

Shaw, J., Lickey, E.B., Schilling, E.E., Small, R.L., 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: The Tortoise and the hare III. American Journal of Botany 94, 275–288.

Sinou, C., Forest, F., Lewis, G.P., Bruneau, A., 2009. The genus Bauhinia s.l. (Leguminosae): A phylogeny based on the plastid *trn*L-*trn*F region. Botany 87, 947-960.

Slowinski, J.B., Page, R.D.M., 2008. How should species phylogenies be inferred from sequence data? Systematic Biology 48, 814-825.

Small, R.L., Cronn, R.C., Wendel, J.F., 2004. Use of nuclear genes for phylogeny reconstruction in plants. Australian Systematic Botany 17, 145–170.

Smith F.G., 1964. Some pollen grains in the Caesalpiniaceae of East Africa. Pollen et Spores 6, 85– 98.

Smith, B.T., Ribas, C.C., Whitney, B.M., Hernández-Baños, B.E., Klicka, J., 2013. Identifying biases at different spatial and temporal scales of diversification: A case study in the Neotropical parrotlet genus *Forpus*. Molecular Ecology 22, 483–494.

Soltis, D.E., Albert, V.A., Leebens-Mack, J., Bell, C.D., Paterson, A.H., Zheng, C., Sankoff, D., DePamphilis, C.W., Wall, P.K., Soltis, P.S., 2009. Polyploidy and angiosperm diversification. American Journal of Botany 96, 336–348.

Soltis, P.S., Marchant, D.B., Van de Peer, Y., Soltis, D.E., 2015. Polyploidy and genome evolution in plants. Current Opinions in Genetics and Development 35, 119–125.

Soltis, P.S., Soltis, D.E., 2009. The Role of Hybridization in Plant Speciation. Annual Review of Plant Biology 60, 561–588.

Soltis, P.S., Soltis, D.E., 2016. Ancient WGD events as drivers of key innovations in angiosperms. Current Opinions in Plant Biology 30, 159–165.

South African National Biodiversity Institute, 2016. Botanical Database of Southern Africa (BODATSA) [dataset].

Southern, E.M., 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. Journal of Molecular Biology 98, 503-517.

Stai, J.S., Yadav, A., Sinou, C., Bruneau, A., Doyle, J.J., Fernández-Baca, D., Cannon, S.B., 2019. Cercis: A non-polyploid genomic relic within the generally polyploid legume family. Frontiers in Plant Science 10, 1-18.

Strand, A.E., Leebensmack, J., Milligan, B.G., 1997. Nuclear DNA- based markers for plant evolutionary biology. Molecular Ecology 6, 113– 118.

Sukumaran, J., Knowles, L.L., 2017. Multispecies coalescent delimits structure, not species. Proceedings of the National Academy of Science USA 114, 1607–1611.

Taberlet, P., Gielly, L., Pautou, G., Bouvet, J., 1991. Universal primers for amplification of three non-coding regions of chloroplast DNA. Plant Molecular Biology 17, 1105-1109.

Takundwa, M., Chimwamurombe, P.M., Cullis, C.A., 2012. A Chromosome Count in Marama Bean (*Tylosema esculentum*) by Feulgen Staining using Garden Pea (*Pisum sativum* L.) as a Standard. Research Journal of Biology 2, 177–181.

Takundwa, M., Chimwamurombe, P.M., Kunert, K., Cullis, C.A., 2009. Isolation and characterization of microsatelllite repeats in marama bean (*Tylosema esculentum*). African Journal of Agricultural Research 5, 561-566.

Takundwa, M., Nepolo, E., Chimwamurombe, P.M., Cullis, A.C., Kandawa-Schulz, M.A., Kunert, K., 2010. Development and use of microsatellites markers for genetic variation analysis, in the Namibian germplasm, both within and between populations of marama bean (*Tylosema esculentum*). Journal of Plant Breeding and Crop Science 2, 233-242.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: Molecular evolutionary genetic analysis. Molecular Biology and Evolution 30, 2725-2729.

Tate, J.A., Simpson, B.B., 2003. Paraphyly of *Tarasa* (Malvaceae) and diverse origins of the polyploidy species. Systematic Botany 28, 723-737

Tautz, D., Renz, M., 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Research 12, 4127–4138.

Taylor, C.L., Barker, N.P., 2012. Species limits in *Vachellia (Acacia)* karroo (Mimosoideae: Leguminoseae): Evidence from automated ISSR DNA "fingerprinting." South African Journal of Botany 83, 36–43.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acid Research 22, 4673-4680.

Tomiuk, J., Guldbrandtsen, B., Loeschcke, V., 2009. Genetic similarity of polyploids: A new version of the computer program POPDIST (version 1.2.0) considers intraspecific genetic differentiation. Molecular Ecology Resoures 9, 1364–1368.

Torre A.R., Hillcoat, J.O.D., 1955. *Tylosema* (Schweinf.) Torre & Hillcoat, gen. nov. (Leguminosae). In: Exell AW, Mendonça FA,eds. Boletim da Sociedade Broteriana, série 2 29, 38.

Townsend, J.P., Lopez-Giraldez, F., 2010. Optimal selection of gene and ingroup taxon sampling for resolving phylogenetic relationships. Systematic Biology 59, 446–457.

Travlos, I.S., Economou, G., Karamanos, A.I., 2006. Germination and emergence of the hard seed coated *Tylosema esculentum* (Burch) A. Schreib in response to different pre-sowing seed treatments. Journal of Arid Environments 68, 501-507.

Travlos, I.S., Economou, G., Karamanos, J., 2007. Effects of heat and soil texture on seed germination and seedling emergence of marama bean, *Tylosema esculentum* (Burch.) A. Schreib. Journal of Food, Agriculture, and Environment 5, 153-156.

Turner, B. L., Fearing, 0. S.,1959. Chromosome numbers in the Leguminosae. II. African species, including phyletic interpretations. American Journal of Botany 46, 49-57.

Van De Peer, Y., Mizrachi, E., Marchal, K., 2017. The evolutionary significance of polyploidy. Nature Reviews Genetics 18, 411–424.

Van der Maesen, L. J. G., 2006. *Tylosema esculentum* (Burch.) A.Schreib. In ‘‘PROTA 1:Cereals and pulses/Ce´ re´ ales et le´ gumes secs’’, (M. Brink and G. Belay, Eds). PROTA,Wageningen, Netherlands.

Van Wyk., B., Gericke, N., 2000. People’s Plants: A guide to useful plants of southern Africa. Briza, Pretoria, 22.

Varshney, R.K., Graner, A., Sorrells, M.E., 2005. Genic microsatellite markers in plants: Features and applications. Trends in Biotechnology 23, 48–55.

Verdoorn I.C., 1959. *Bauhinia esculenta*. In: Dyer RA, ed. The flowering plants of Africa 33: Plate 1311.

Vos, P., Hogers, R., Bleeker, M., Reijan, M., van de Lee, T., Hornes, M., Firners, A., Pot, J., Peleman, J., Kuiper, M., Zabeau, M., 1995. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Research 23, 4407-4414.

Wang, X., Rinehart, T.A., Wadl, P.A., Spiers, J.M., Hadziabdic, D., Windham, M.T., Trigiano, R.N., 2009. A new electrophoresis technique to separate microsatellite alleles. African Journal of Biotechnology 8, 2432–2436.

Wang, X.Q., Zou, Y.P., Zhang, D.M., Zhang, Z.X., Hong, D.Y., 1996. Problems in the use of RAPD to the study of genetic diversity and systematics. Acta Botanica Sinica 38, 954-962.

Wang, Y.H., Wicke, S., Wang, H., Jin, J.J., Chen, S.Y., Zhang, S.D., Li, D.Z., Yi, T.S., 2019. Plastid genome evolution in the early-diverging legume subfamily cercidoideae (Fabaceae). Frontiers in Plant Science 9, 1–12.

Waugh, R., Power, W., 1992. Using RAPD markers for crop improvement. Trends in Biotechnology 10, 186–191.

Wendel, J.F., Doyle, J.J., 1998. Phylogenetic incongruence: Window into genome history and molecular evolution. In: Soltis, P., Soltis, D., Doyle, J. (Eds.), Molecular Systematics of Plants II. Kluwer Academic, Dordrecht, pp. 265–296.

Wheeler, Q.D., 2005. Losing the plot: DNA “barcodes” and taxonomy. Cladistics 21, 405-407.

White, T. J., Bruns, T., Lee, S., Taylor, J., 1990. Amplification and direct sequencing of fungal

Wiens, J.J., Morrill, M.C., 2011. Missing data in phylogenetic analysis: Reconciling results from simulations and empirical data. Systematic Biology 60, 719–731.

Will, K.W., Mishler, B.D., Wheeler, Q.D., 2005. The perils of DNA barcoding and the need for integrative taxonomy. Systematic Biology 54, 844–851.

Williams, J.G.K., Kubelik, A.R., Livak, K.J., Rafalski, J.A., Tingey, S. V., 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Research 18, 6531–6535.

Wojciechowski, M.F., Lavin, M., Sanderson, M.J., 2004. A phylogeny of legumes (Leguminosae) based on analysis of the plastid matK gene resolves many well-supported subclades within the family. American Journal of Botany 91, 1846–1862.

Wunderlin, R., Larsen, K., and Larsen, S.S. 1987. Reorganization of the Cercideae (Fabaceae: Caesalpinioideae). Biologiske Skrifter 28, 1–40.

Yang, Z., 2015. The BPP program for species tree estimation and species delimitation. Current Zoology 61, 854-865.

Yang, Z., Rannala, B., 2010. Bayesian species delimitation using multilocus sequence data. Proceedings of the National Academy of Science USA 107, 9264–9269.

Yao, H., Song, J., Liu, C., Luo, K., Han, J., Li, Y., Pang, X., Xu, H., Zhu, Y., Xiao, P., Chen, S., 2010. Use of ITS2 region as the universal DNA barcode for plants and animals. PLoS One 5, e13102.

Zhang, D.X. 1995. A cladistic analysis of *Bauhinia* L. (Leguminosae). Chinese Journal of Botany 7, 55–64.

Zharkikh, A., 1994. Estimation of evolutionary distances between nucleotide sequences. Journal of Molecular Evolution 39, 315-329.

Zwickl, D.J., Hillis, D.M., 2002. Increased taxon sampling greatly reduces phylogenetic error. Systematic Biology 51, 588–598.

Supplementary Data

Table S Microsatellite markers developed using scaffold data from the T. esculentum genome provided by Dr Chris Cullis, Case Western Reserve Univesity, Ohion. Highlighted markers have been tested for polymorphisms in T. esculentum and T. fassoglense.

Label	Motif	Count	Motif type	Leff Sequence	Right Sequence	PCR Product Size
LM1	AAGGAG	8	6	TACTCGTCGTACTGCACTGG	TGGTGGTGGAGTTGAGGAAG	339
LM6	ACAT	8	4	CCAAGTTGGCAGAGTGGTTG	TCATCAACAGCTAGCCTCCG	346
LM8	ATC	13	3	AGATGCTGCGTTACCTTGAAG	TCCAGGGCCCACAATTGATG	298
LM9	ATC	13	3	GGGATCCTCATTGCTGACAG	CCTCTGAAATTTCTCGCCAGG	330
LM10	AAG	11	3	ACCACTGAATAATGGCCAACG	ACGTGGACCAAGATCAGAGC	409
LM2	AAG	11	3	GCTGGAAATGGTGGAAGAGC	TTGTTGCGGTGAGGAGGAAG	377
LM3	AAG	14	3	CTTCGAGCTGCCACAAGAAG	ACTAAGAGGAGGATGGCAAATG	303
LM4	AAG	12	3	TCTAAGGAAGCGGCTGGAAG	TCCCTTTCCAGTTCTCCCTC	399
LM5	AAG	12	3	GGAGGACCTAGGAACCTTGTC	TCCATTGCCAACCCAACTTG	240
LM7	AAC	11	3	CTTTCGAAGGCCACTGGAAC	GCCAACATAGAGAAGAGGCG	403
LM11	AAG	11	3	AGGACATTGGAGCTTGGGAG	CCCATGAATGTTGTCCACGG	428
LM12	AAG	11	3	GAGAGACGTTACTTTGGCCG	TTCATTTCAGCTTCGGTGCC	320
LM13	AAG	12	3	ACCACTGAATAATGGCCAACG	ACGTGGACCAAGATCAGAGC	409
LM14	AAG	12	3	GAGGCACGGCTCTAGACTC	AGAATCTCTCAAGGGCAGCC	268
LM15	AAG	12	3	GGTGAAGCATGCACAATTCC	AACAAAGGGAGAGGCTGAGG	325
LM16	ATC	11	3	CAACGCTGAGGTTGTGTCTG	AGAATCCAAGTTGAGGTCATGG	409
LM17	AAG	11	3	TCCAGCGAGAATAGGCCTTG	GGAGACGTGAACACAAACCC	246
LM18	AAG	11	3	TGTCTCATTCATGCTGTCCC	TTGAGGGTACTTGTCTCGGG	389
LM19	AAG	13	3	ACGAAATGGCTCTGCTTTGG	GCACCGAACAATCTGAACCC	347
LM20	AAG	11	3	GCAGCCACGGTACAAGAAAG	AACTAGAGAAAGGGCTGCCG	324
LM21	AAG	12	3	GGTGAAGCATGCACAATTCC	AACAAAGGGAGAGGCTGAGG	325
LM22	AAG	12	3	GCACAGCGATGTTCCTTCTG	TGTGGTGAGGGAAGGGAAAG	418
LM23	AAG	11	3	ATGGCAGTCACAGTACATGC	TGGAGTGAGGGTTGTTGAGG	174
LM24	AAG	12	3	ATAGGGCGTGAGGGTAGTTG	AAGTGCTCAAACCGTCAAGC	390
LM25	AAG	14	3	ATAGGGCGTGAGGGTAGTTG	AAGTGCTCAAACCGTCAAGC	390
LM26	AAG	13	3	TAGCTATTCTACTGCCGCCG	ACGAAATGGCTCTGCTTTGG	400
LM27	AAG	12	3	CACACTTCCACCTTCGACTC	ACCCTGCAACTCCACTTCTC	447
LM28	AAG	11	3	AGAGTGGGTGGTTGAAAGGG	GGCTGCCTCAACAATTGGTC	423
LM29	AAG	12	3	TGTCTGTGCTCTGGACTTCC	AGCCCAGAGAAGGAATCAGTC	332
LM30	AAG	11	3	TTGAGGGTACTTGTCTCGGG	TGTCTCATTCATGCTGTCCC	389
LM31	AAG	11	3	CGGAGAGAACGATGTTGCTG	TTCGGTGAATCAGTGCAAGC	317
LM32	AAG	11	3	TGGTTGCTGTCTAGTGGTCC	ATCACTTGCATCTGGGTTGC	303
LM33	AAG	11	3	GCCTGTCTAGATCCTGAGGC	GCCATCGGTATTCTTACTTGCC	174
LM34	AAG	18	3	TGGTCGTGGAAGTGGAAGAG	TTGCTAGCACCGGTGTCAAG	361
LM35	AAG	13	3	CCTCTCACCACAGCAGTTTC	CGATGGTCTTCACTGTTGCC	436
LM36	AAG	11	3	GATGTCGTAACCTTGGGCAC	ACGAAGATTTGCTGGACCAAC	442
LM37	AAG	13	3	AGAGAAAGGGCTGCCACTAG	CCACGCCAAGAGAGGAATTC	384
LM38	AGAT	8	4	CAGCTGCAACAAATCACGC	TTGCTGAGTTGCCTGAAAGC	381
LM39	AGAT	10	4	ATTCCTTGTGGCTAACGTGC	ACTTGTGCATGCATGGTAGG	369
LM40	ACTC	9	4	GTCCATTCCTTTGCCCAACC	ATGAAAGCCGGCACAGATTG	392
LM41	AAAG	8	4	CTGCCTCTCTCCCACTAACG	GCAATCAGATAATGCAGGCTG	143
LM42	ACAT	8	4	ACCGCAAACACCAAACTTCC	AACAGGCCAGGTCACAACTC	306
LM43	AAAG	8	4	ACGCACTAGTAGACCAACCC	TGGAATTTGAGTTGGCGGTG	290
LM44	AGAT	9	4	CCCGTGGAAAGAGTCTCTGC	TCGATTTGAACGACAGTGCC	375
LM45	AGAT	9	4	CCCGTGGAAAGAGTCTCTGC	TCGATTTGAACGACAGTGCC	375
LM46	AACCCT	10	6	ATCTGCTTCTGGACCGTCTC	AGTCCGGTGGCTATTCAGAC	405