An integrative taxonomic approach using genome-wide SNP data resolves variation in the *Dimorphotheca pluvialis-sinuata* species complex (Calenduleae: Asteraceae) in South Africa

By Thaabiet Parker

Thesis presented in fulfillment of the degree of Bachelor of Science Honours in Biological Sciences

Under supervision of:

Assoc. Prof. G.A. Verboom (Department of Biological Sciences, UCT)

Dr N.G. Bergh (Compton Herbarium, South African National Biodiversity Institute)

Prof A.G. Ellis (Department of Botany and Zoology)

Department of Biological Sciences, University of Cape Town

October 2019

ABSTRACT

Here I present an integrated taxonomic approach utilising genome-wide single nucleotide polymorphism (SNP) data in conjunction with morphological and environmental data to resolve variation in the South African daisy species complex *Dimorphotheca pluvialis-sinuata*. This species has historically been separated based on white vs orange ray colour alone, but field observations have shown widespread variation in ray colour patterns, fruit characters and disk floret sculpturing.

45 populations were sampled from 17 a priori morphs which encompassed the range of variation in the complex; from which DNA of 93 individuals were extracted for genotyping-by-sequencing (GBS) and subsequent creation of a SNP dataset of variant sites across the genome. Structure-like analysis in the form of sMNF algorithms were used to assign populations to independent genetic clusters i.e. hypothesised evolutionary species. Support for genetic cluster assignment was tested using a PCoA of genetic distances, pairwise *F*st^{WC}, LDA and exploratory analysis of morphological and selected environmental variables.

All analysis of SNP data showed support for five independent gene clusters within the species complex which corresponded to five potential evolutionary species. Most species lacked clear distinction in morphology, with most clear differentiation being driven by ray fruit characters. Species assignment was strongly supported by ecological niche partitioning on the basis of differing elevation, precipitation and soil clay levels.

The five species were as follows : (S1) A white-rayed form from the coastal Cape Town region with long, smooth ray fruits and elongate disk appendages; (S2) White-rayed forms which were morphologically similar to S1 but distributed in the Namaqualand region; (S3) Orange/peach-rayed forms from Namaqualand Hardeveld escarpment with small, tubercled ray fruits; (S4) Three variable forms from lowland Darling crossing into Olifants Valley; (S5) Highly morphologically diverse, clay forms from the Cederberg, Vanrhynsdorp area.

INTRODUCTION

Species conceptualization remains a major challenge in taxonomy, with biologists advocating different and partially-incompatible species concepts over the past decade. These range from the phenetic species concept, which uses morphology to delimit species (Sneath, 1976), to the biological species concept (BSC; de Queiroz, 2005), which defines species as reproductively isolated populations. Recently there has been a growing consensus to recognize species as independently evolving metapopulations (i.e. ancestor-descendent lineages) that may remain semi-permeable to gene flow from other such populations (De Queiroz, 2007, Wiley, 1978). This evolutionary species concept (ESC) separates species conceptualization from the operational criteria used to delimit species in practice. The ESC thus recognizes the need for species delimitation based on a combination of different available data sources as separate lines of evidence.

This integrated taxonomic approach is advantageous in many aspects as it serves to resolve much of the current conflict relating to species conceptualization. For example, application of the phenetic species concept - which has historically dominated traditional taxonomy - often fails to recognize evolutionarily independent lineages due to lack of diagnostic phenotypic traits (Jensen, 2009). Other popular species concepts present separate issues; such as the biological species concept being only applicable to sexually reproducing species, or the phylogenetic species concept resulting in potentially excessive species delimitation (Zachos, 2015).The ESC explicitly does not define a cut-off point for when metapopulations should be considered species, but use of multiple operational criteria as supporting lines of evidence for species delimitation presents a more comprehensive approach to understanding variation in taxa (Leaché et al., 2014). Of particular relevance is that many taxonomic revisions now incorporate genetic data as primary support for species delimitation; using morphological and environmental data as supplementary support and to inform on speciation processes (Chesters et al., 2012, Edwards & Knowles, 2014).

The application of DNA data to systematics has generally been used in the context of single-marker DNA barcoding- which operates under the assumption that the evolution of genes

match phylogenetic relationships. Thus this approach does not take into account factors such as introgression, incomplete lineage sorting and hybridisation which result in inferred phylogenies showing discordance with so-called 'gene trees' (Degnan & Rosenberg, 2006). Single nucleotide polymorphisms (SNPs), sampled from across the whole genome, circumvent this issue on account of their multi-locus nature and are thus useful for resolving phylogenetic and consequently taxonomic problems (Spinks et al., 2014). SNP data, as produced by genotyping-by-sequencing (GBS) methods, are also advantageous in that several thousand to several hundred thousand SNPs are typically identified and sequenced; resulting in a significantly higher information content even in non-model organisms (Leaché et al., 2014). In addition, SNP data provides a good trade-off between ease of application, sequencing costs and number of characters versus more cost-intensive processes such as whole genome sequencing (Leaché & Oaks, 2017). Despite the many advantages of SNP datasets, their application in plant taxonomy remains limited, particularly in South African plants.

The genus *Dimorphotheca* is a member of the Cape-centered Calenduleae (family: Asteraceae), a relatively small tribe currently comprising 12 genera and approximately 120 species. The most defining features of Calenduleae include leaf-like involucral bracts in two series and the lack of a pappus at the base of the florets (Nordenstam, 2007:240-245). *Dimorphotheca* in turn is defined by large ray florets with a striking appearance and in several species by dimorphic cypsela; with ray cypsela rod-shaped and disk cypsela flattened and papery with wings (Norlindh, 1943; Figure 2A). *D. sinuata-pluvialis*, a complex within *Dimorphotheca* currently recognized as two species, is a good example of the shortcomings of morphological delimitation, particularly based strictly on a few key characters and discontinuous sampling. The existing classification of the study species complex is by Norlindh (1943), who dedicated much of his life on Calenduleae, and whose work I have used as the baseline taxonomic reference to compare with observed field variation in the species complex.

Dimorphotheca pluvialis and *Dimorphotheca sinuata* are two of the most well-known South African daisies, famous for their contribution to tourism flower displays and cultivation as an ornamental. Currently described by Norlindh (1943) as sister species, *D. pluvialis* is native to the Western Cape, whereas *D. sinuata* is most present in the Namaqualand region. Both *D. sinuata*

and *D. pluvialis* are annual herbs; with slender, usually erect stems, sinuate-dentate leaves and large solitary flowerheads surrounded by a single series of glandular, hairy involucral bracts. Leaf form is variable and environmentally plastic, plants growing in drier environments typically have smaller 1-3 cm long sinuate-dentate leaves while those growing in moist environments have been described to exhibit leaves up to 10 cm long and 3 cm widen (Norlindh, 1943). Norlindh describes this difference as potentially being due to environmental plasticity but also importantly considers that differences are genotypic, including those associated with his 'dwarf sand-dune ecotype'. However, the defining feature currently separating *D. pluvialis* and *D. sinuata* is the colour of the ray florets. In *D. pluvialis* rays are white with violet or purple at the base, whereas rays of *D. sinuata* are orange or reddish-orange. The only other factor considered by Norlindh was the difference in geographical distribution based on his sampling of the species.

Norlindh's existing classification however presents several issues. The separation of the species based on almost exclusively flower colour may be uninformative, as flower color may be affected by factors such as adaption to local pollinator suites (de Jager et al., 2011). Secondly, field observations by A.G. Ellis and John Manning reveal a diversity of ray colours within the complex, outside of white and orange. This division also does not take into account the substantial population-level variation in ray colour patterns, ray size, pollen colour, fruit structuring, disk floret appendages and disk corolla hairs across populations of the species complex (see Figure 1&2). Taken together these issues suggest that the current classification of the complex drastically underestimates the number of evolutionary species present.

Thus, considering the currently exclusively morphological classification of *Dimorphotheca pluvialis-sinuata,* which does not reveal taxa in the complex, I use an integrated taxonomic approach to reassess species limits in the complex. Towards this objective, I sampled genetic evidence in the form of SNP data, from numerous populations of *Dimorphotheca pluvialis-sinuata* which encompassed the observed variation in morphology. These data are then used to hypothesise the existence of multiple genetically-divergent species whose distinctness is then evaluated in terms of morphology as well as environmental differentiation.

Figure 1: An illustration of the main a priori morphs recognized within the *Dimorphotheca pluvialis-sinuata* complex (A) white sand (B) tricolor (C) orange (D) white hairy (E) peach loeries (F) red hairy (G) mulberry (H) peach north (I) dp white cape town. A list of morph names and descriptions can be found in Appendix B. Photos by A.G Ellis

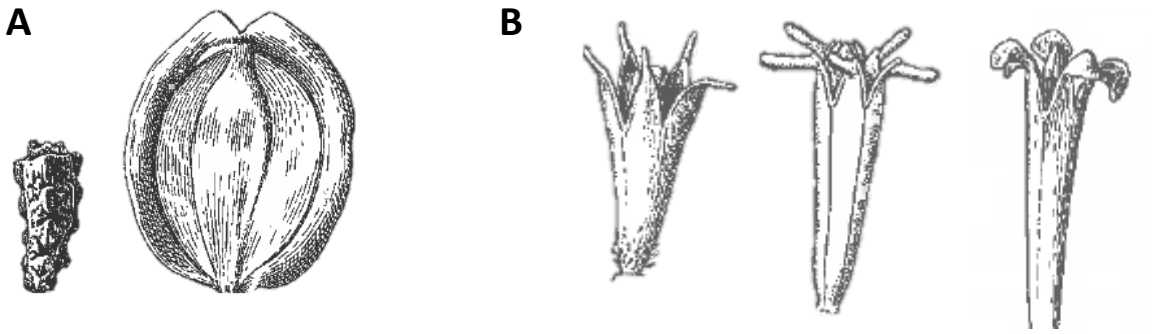


Figure 2: Excerpt of monograph from Norlindh (1943); (A) Differentiation between an example of ray (left) and disk (right) achenes in *Dimorphotheca*; (B) Differentiation in disk corolla lobe appendage structures in the *Dimorphotheca pluvialis-sinuata* complex, right showing 'teardrop' appendage, middle and left showing varying length of elongate appendage type.

METHODS

Field sampling

D. pluvialis and *D. sinuata* populations were sampled in September-November 2018. The field observations of A.G. Ellis lead to a priori classification of the complex into 17 morphs based on ray colour, central disk corolla appendages and other notable floral traits (see Appendix B for names and descriptions). Genetic material was collected from 45 populations, ranging from south of Namibia as far east as Koensrust, and encompassed the range of variation in traits. For each population, leaf cuttings were collected for DNA extraction from each of 5-10 individuals and placed on silica-gel in an airtight container. Flower heads were also collected from individuals in each population and preserved in FAA (60% ethanol, 25% distilled water, 10% formalin, 5% glacial acetic acid). Pressed voucher specimens for each individual will be properly accessioned at the Compton Herbarium (NGB) in Kirstenbosch. In addition, ray and disk fruits from multiple individuals in each population were sampled where possible. Multiple individuals (4-9) per morph were selected for DNA extraction. A single individual per population was used when multiple populations were available, and for wide-ranging species populations were

selected to represent the distribution limits. For range restricted morphs, multiple individuals (3-4) were sampled per population.

Morphological measurements

The morphology of each of the field-sampled specimens was characterized by measuring floral characters from the FAA-preserved material under a dissecting microscope, except in the case of two populations for which pressed specimens were used. Flower heads from two individuals per population were measured. Measurements were done on florets at or after antheses wherever possible. Characters were selected on the basis of observed field variation (see Appendix C). Notable inclusions were continuous measurements to characterize differences in inner disk floret appendage structures (elongate vs 'teardrop', Figure 2B), in the form of appendage length and appendage diameter. The character list also included 'capitulum curvature, which was a proportion of the total inner to outer disk floret length (ovary + corolla tube + lobe), in order to capture differences in height of inner disk florets not enforced by disk floret appendages. Fruit characters were measured using collections from two individuals of each population, except seven populations where no fruits were found. Ray colour was categorized by three binary variables at the morph level: presence of a thick dark band at the ray base, un-pigmented vs pigmented dominant ray colour and presence of two ray colours.

Environmental measurements

Environmental variables were selected on the basis of factors which typically influence vegetation structure in the GCFR and Namaqualand as well as the ecology of the two species as winter annuals (Cowling et al., 1999, Gremer & Venable, 2014). All variables were sourced at GPS localities of sampling populations, one population (Population 3, Appendix A) had no GPS co-ordinates recorded and was thus excluded from subsequent environmental analysis. Climate variables were sourced from WorldClim v 2.0 at 30 second resolution (Fick & Hijmans, 2017) and were as follows: Temperature of wettest quarter (TWQ), mean annual precipitation (MAP) and precipitation of wettest quarter (PWQ). Soil variables were sourced from a recently released dataset by Cramer et al. (2019) which models soil layers in the GCFR from a network of analysed samples. The following soil variables were compared: pH, extractable phosphorous

(mg/kg soil), total nitrogen (% N weight/soil weight) and average clay (% clay weight/soil weight). Elevation was sourced from the Shuttle Radar Topography Mission (SRTM;Van Zyl, 2001) data.

Genomic DNA extraction and purification

DNA extraction from silica material was done using the 2% CTAB extraction protocol of Doyle & Doyle (1990) with a few modifications. Twenty milligrams of dried leaf material was pulverized using a Retsch™ Mixer-Mill tissue grinder at 250 rpm for 15 minutes. Polyvinylpyrrolidone (PvP) was added in the grinding stage in order to absorb polyphenols which could co-precipitate with DNA. The chloroform: isoamyl (24:1) wash step was performed twice to remove persistent organic matter. DNA precipitation via addition of a 2/3 volume of isopropanol was extended to 12 hr and samples were kept at -20 ^{o}C. Washing of the precipitated DNA pellet was done using 750 ul of 70% ethanol after centrifugation. This step was repeated twice to improve DNA purity, and pellets were left to dry for approximately 3 hr to allow excess ethanol to evaporate. The pellet was eluted in 60 ul TE buffer, heated at 65^{o}C for 30 minutes, and mixed by pipetting to allow for DNA resuspension. Where DNA pellets did not resuspend; additional TE volume was added and extracts were reheated and re-vortexed. Subsequent volumes associated with purification were adjusted proportionally.

Three microlitres of RNase were added and the samples incubated for 15 minutes at 37 ^{o}C. Samples were purified using Beckman Coulter™ Agencourt Ampure XP magnetic beads. For each sample, 1.5 volume of magnetic bead solution was added and mixed by pipetting. After allowing the magnetic beads to bind to DNA, the samples were placed on a magnetic separating rack (Invitrogen-Dynal™, ThermoFisher Scientific, USA) to allow the beads to separate and the remaining solution, containing contaminants, was discarded. The beads were then washed twice with 70% ethanol as in the CTAB extraction protocol and dried. Ethanol drying time varied based on the humidity on the day but was generally kept to approximately 30 minutes to reduce DNA loss. The magnetic beads were re-eluted and separated on the magnetic rack, after which purified DNA solution was transferred to new tubes. DNA concentration and quality were assayed using a NanoDrop ND-1000 spectrophotometer (ThermoFisher Scientific, USA).

Extracts were also examined on 1% agarose gels to visualize DNA degradation. Samples with concentrations > 50 ng/ul and 260/280 absorbance ratios between 1.7 and 2 that exhibited minimal degradation on gels were used for sequencing.

Sequencing and SNP filtering

A total of 96 individuals were selected for genotyping-by-sequencing (GBS). Extractions were stored at -20 ^{o}C except during transit from Cape Town to Novogene Genome Sequencing Company Ltd. in Beijing, China. Here samples were re-assayed using a Qubit fluorometer before GBS according to Novogene's protocol outlined at https://en.novogene.com/. The GBS workflow generates a reduced representation library which is then high-throughput sequenced. Sample genomes were fragmented using the enzymes MseI, HaeIII, and MspI at restriction sites on the chromosomes to generate 350 base-pair (bp) fragments which were then ligated with P1 and P2 barcodes complementary to fragment overhangs and amplified by PCR. Paired-end sequencing was performed on PCR-enriched adaptor-ligated DNA fragments on the Illumina HiSeq platform, generating approximately 200,000 (RAD loci) of 150 bp reads. Raw sequence data were filtered to exclude paired reads which contain Illumina library construction adaptors, paired reads where uncertain nucleotides constitute more than 10 % of the read, and paired reads where low quality nucleotides (quality score ≤ 5, or a 10 % error rate) constitute more than 50 % of either read. The filtered reads were then used to assemble a *de novo* reference genome (a reference consisted of essentially consensus of SNPs across all individuals sequenced) as there is no available reference genome for any close relatives. At the time of writing, only 93 individuals had been sequenced so data from these were used for subsequent analysis.

A concatenated variant dataset across all 93 individuals was produced by quality-trimming and mapping all reads to the *de novo* reference genome, calling variants (SNPs and indels), and finally filtering the variants. All the above processing was conducted using a high-performance computing unit facility with 40 processing cores and 384 GB of memory (University of Cape Town ICTS High Performance Computing team). Trimming, alignment and basic filtering were implemented using the *dDocent* pipeline (Puritz et al., 2014); and variant calling and additional

filtering steps using the NGSEP platform (Perea et al., 2016) in *R* version 3.5.1 (R Development Core Team, 2018). Quality trimming of cleaned reads was performed using Trimmomatic (Bolger et al., 2014) which removes any remaining adaptor contamination and low-quality bases (below a quality score of Q20, or a 1% error rate) from the beginnings and ends of reads where read quality tends to be poor. Trimmed reads were mapped to the *de novo* genome using the Burrows-Wheeler Aligner MEM algorithm (Li & Durbin, 2010), with default conservative mapping parameter settings [-A 1 -B 4 -O 6]. Alignment files created using BWA were combined using the merge function from SAMtools v 1.9 (Li, 2011) with default parameters. Variant calling was performed using NGSEPs multi-sample variant detector function with 100 alignments allowed to start at the same site.

SNP filtering was conducted with two sets of parameters, corresponding to strict and lenient filtering, as a form of sensitivity testing for downstream analysis. In both datasets, variants which were successfully genotyped in at least 90% of individuals were kept (as opposed to default 50 % by dDocent). The parameters which were changed between datasets were genotype call rate (10 % across five populations in in lenient, 10 % across 1 population in strict) and minimum allele depth (three in lenient, five in strict). All other parameters were identical between datasets. Individuals missing 40% of data were removed. Only biallelic SNPs were retained and indels were filtered as well. Finally, scaffolds containing more than 20 SNPs were excised, as these are most likely paralogs.

Identification of genetic clusters

Analysis integrated assessment of separation within the complex using genetic, morphological and environmental variables. Initial genetic delimitation hypotheses were formulated by assigning individuals to estimated number (K) of ancestral gene pools using sparse Non-Negative Matrix Factorization(sNMF) algorithms developed by Frichot et al. (2014) applied to the SNP data set as implemented in the *R* package *LEA* v 2.2.0 (Frichot & François, 2015). The sNMF algorithm computes individual ancestry coefficients, which is typically interpreted as the probability of an individual coming from a particular gene pool (Frichot et al., 2014) using least-squares minimisation of allele frequencies with similar results as in other likelihood-based

methods such as STRUCTURE (Pritchard et al., 2000). One hundred ancestry coefficient matrices (runs) were computed for each value of K between K = 2 and K = 10, where K represents the number of ancestral gene pools or putative species. The best K solutions was selected using the lowest entropy criterion computed during sMNF as recommended by Frichot & François (2015). The ancestry coefficient estimates generated by sNMF for the best-supported value of K were summarised using CLUMPAK (Kopelman et al., 2015), which produces major and minor solutions for each value of K based on the number of runs supporting the solution (i.e. the best-supported mode of clustering gene pools, followed by next-best-supported modes).

The best supported genetic clusters from sNMF were then explored using principal coordinate analysis (PCoA), with individuals as entities, and biallelic SNP loci as attributes, implemented in the *R* packages *adegenet* (Jombart, 2008) and *dartR* (Gruber et al., 2018). Differentiation between hypothesized genetic clusters was assessed using a modification of Wright's pairwise *F*st by Weir & Cockerham (1984) implemented in the *R* package *stamPP* (Pembleton et al., 2013). Pairwise Fst^{WC} comparisons were implemented using genetic clusters as populations.

Assessment of morphological and environmental support

Discrete morphological variation was assessed at the a priori morph level by simple diagrammatic visualization of relevant character states for each morphotype. Continuous morphological variation between sNMF genetic clusters was assessed using linear discriminant analysis (LDA) in the *R* package *MASS* (Ripley et al., 2013) . LDA attempts to find the best linear combination of continuous variables, in this case morphological variables, that discriminates pre-defined classes of objects (in this case sNMF genetic clusters). LDA was conducted on the best genetic cluster hypotheses produced by sNMF, after which highly uninformative and autocorrelated morphological characters were dropped from the analysis and the LDA rerun. As fruit measurements were included in the LDA, the seven populations (14 individuals) from which no fruits were sampled were excluded from this analysis, but representatives of all a priori morphs were present. Potentially informative morphological characters between genetic clusters were also explored using boxplots to assess within-group variation in single variables. An identical analytical methodology was used with regards to environmental variables.

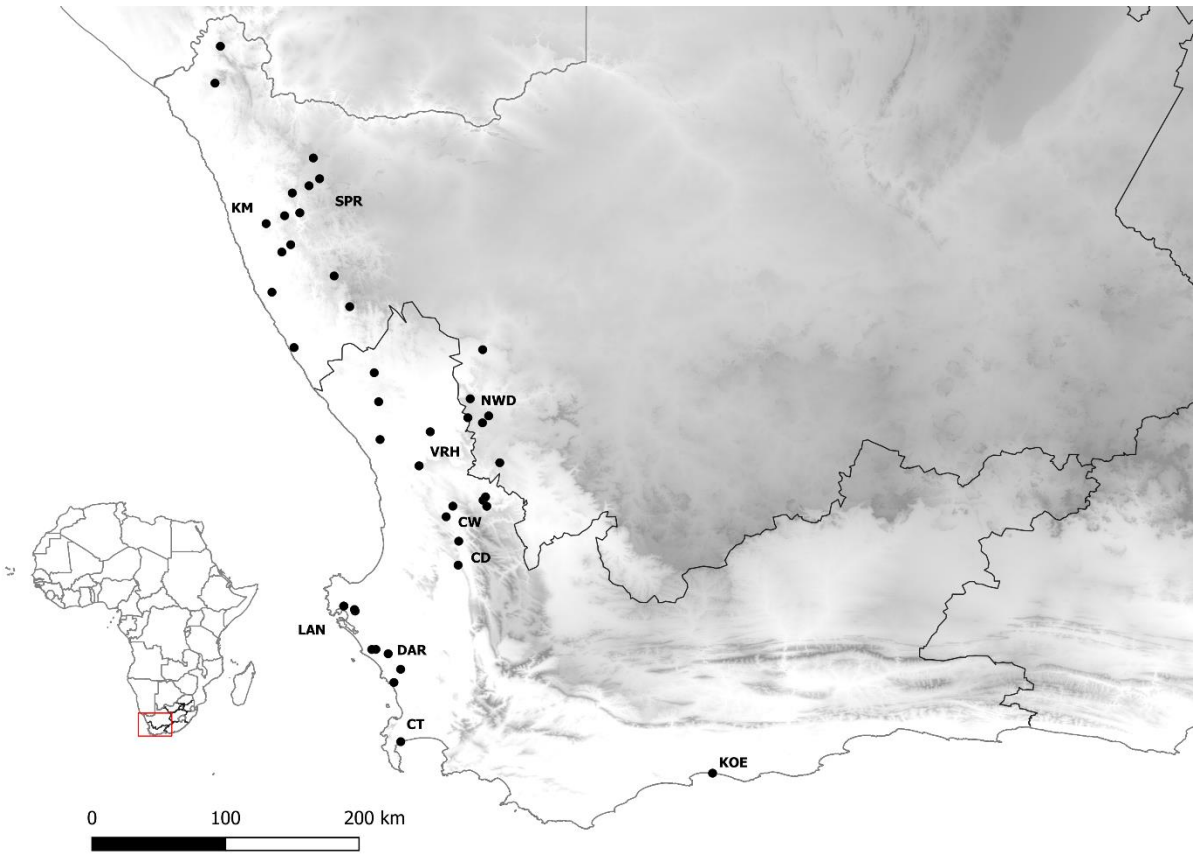


Figure 3: Sampling distribution map showing localities of populations. Elevation overlayed with darker areas indicated land more than 300 m above sea level. Letters showing approximate locations of important towns: CT – Cape Town, LAN – Langebaan, DAR – Darling, CD – Citrusdal, CW - Clanwilliam, VRH – Vanrhynsdorp, NWD – Nieuwoudtville, KM – Komaggas, SPR – Springbok.

RESULTS

Filtering of the SNP dataset with lenient parameters kept 13 771 out of 38 979 variant sites after filtering for variants which had been successfully genotyped in 90% of individuals, as well as filtering for genotype call rate and minimum allele depth. After removing three and four-state SNPs and indels and excising scaffolds containing more than 20 SNPs, the final dataset consisted of 8506 variants. In contrast, strict filtering produced a dataset with 5025 variants after initial filtering, and 2925 variants as the final dataset. Preliminary analysis of the 2925 SNP

dataset showed similar patterns in the PCoA to the 8506 SNP dataset (Appendix E), thus the latter was used for subsequent analysis as it encompassed more data which could potentially be informative.

CLUMPAK output of ancestry coefficient from sMNF analysis indicated the presence of four to five ancestral gene pools/genetic clusters. The cross-entropy criterion reached a minimum of 0.3301 at K=6 (Figure 4C). Due to K=6 having two minor modes comprising of 21/100 and 22/100 runs respectively which all presented markedly different hypotheses for genetic cluster assignment (Appendix F), this hypothesis was deemed not well-supported and discarded for further analysis. K=5 had nearly identical cross-entropy to K=6 (0.3302, Figure 4C), with the major mode supported by 77/100 runs. The K=4 assignment had marginally higher cross-entropy (0.3317) and was supported in all 100 runs, making K=5 and K=4 the best supported genetic cluster assignments.

Under the K=5 assignment (Figure 4A), the complex can be separated into the following genetic clusters which can be visualized in Figure 5: (i) S1, one white a priori morph ('white cape town') from the Cape Town area as far east as Koensrust; (ii) S2, two white morphs ('white sand' and 'kardoesie') stretching from Komaggas through Vanrhynsdorp with one population near Citrusdal area; (iii) S3, orange and peach morphs ('orange', 'peach north', 'peach loeries','white north') from the Northern Cape region, stretching from Koebus to Loeriesfontein; (iv) S4, three morphs ('mulberry', 'darling candles' and 'kardoesie') from Darling, Langebaan and south of the Olifants Valley; (v) S5, morphologically diverse morphs ('white hairy', 'red hairy', 'white ringed south', 'peach ringed south', 'peach south', 'peach south klawer','tricolor') from the Nieuwoudtville and Clanwilliam area. There was a degree of admixture between the southernmost populations of S3 and northernmost populations of S5, and a large degree of admixture between south-distributed populations of S5 and the S4 cluster. Under the K=4 assignment (Figure 4B) the first three clusters are identical (P1-P3) but S4 and S5 are combined to comprise P4.

PCoA of all individuals showed three highly distinct groupings, corresponding to S1, S2 and S3+S4+S5 from the K=5 assignment (Figure 6A). Most individuals from the S3 (i.e. P3) cluster

could be differentiated from the S4 and S5 (i.e. P4) clusters, which provides support for the K=4 assignment. Two populations corresponded to the 'orange' a priori morph (Population code 27, 30; Appendix A) and the one population of 'peach south klawer' showed potentially significant genetic distance from individuals of the same morph, which corresponded to the degree of admixture between S3 and S5 seen in the sMNF result. A priori morphs generally grouped together, with the only exception the aforementioned 'peach south klawer' population (Population 16, Appendix A). When subsetting out clusters S1 and S2, there was separation between the S3, S4 and S5 cluster which provides some support for the K=5 assignment (Figure 6B). However, the degree of separation between S4 and S5 was similar to internal groupings within S5. In addition, genetic distance between clusters S3, S4 and S5 showed a strong correlation with geographical distance.

Pairwise *F*st^{WC} (Weir and Cockerham's estimator) primarily provided support for the three highly distinct groupings in the PCoA (Tables 1&2), as I am considering *F*st^{WC} of between 0.1 and 0.2 a large degree of between-group genetic variance. For the K=4 assignment the P3-P4 comparison fell below this range (0.060) which supports three clusters. However, for the K=5 assignment the S3-S4 comparison was within this range (0.105) which suggests that failure of pairwise *F*st^{WC} to distinguish between P3/S3 and P4 (S4 and S5) was due to the within-group variance in P4 as well as admixture between southern populations of S3 and S5. This is reflected in that pairwise *F*st^{WC} of the P1/P2 to P3 comparison was similarly large as in the P1/P2 to P4 comparison. This provides some support for both the K=4 and K=5 assignment.

Preliminary exploration of variation in discrete morphological characters between a priori morphs grouped S1 with S2 together based on ray fruits with low sculpturing, white unpigmented rays and elongate disk appendages (Figure 7). S3 morphs showed strongly tubercled ray fruits and pigmented (orange to white-peach) dominant ray colouring with 'teardrop' appendages. S5 in particular was highly variable in ray colour patterning. However, almost all morphs of S5 had 'teardrop' disk appendages and dimorphic ray fruits (smooth and tubercled). Morphs of S4 could not be grouped with any cluster based on discrete traits, with dimorphic fruits similar to S5 but a twisted, elongate disk appendage structure with flattened

tips in some morphs ('kardoesie' and 'darling candles') and an expanded 'teardrop' appendage structure in one morph ('mulberry') as well as diverse ray colour patterning.

Linear discriminant analysis for the K=5 assignment clearly differentiated three clusters based on continuous morphological measurements: S1, S3 and S5 (Figure 8B). The S2 cluster showed significant overlap with S1 and marginal overlap with S3, whereas S4 overlapped with all other clusters except S1. As seen in Table 3, these differences were primarily driven by ray fruit length on the first linear discriminant axis (LD1, linear coefficient (LC) =1.74), and the combination of inner disk appendage length and capitulum curvature on the second linear discriminant axis (LD2, LC=-1.37 and 0.70 respectively). The higher absolute values of the linear coefficients and hence influence on discrimination was supported by differences in ranges of values in univariate comparisons. In terms of ray fruit length, S1 had the longest fruits ranging from 6-8 mm, S2 from 4-6mm, S5 from 4-4.5 mm and S3 from 2-2.5 mm (Figure 9D). Appendage length and capitulum curvature on LD2 most clearly separated S5 from S3 and S2. S5 had generally lower appendage lengths ranging from 0-0.05 mm in contrast to S3 and S2 which ranged from 0.05-0.6 and 0.6-1.2 respectively (Figure 9G). S5 also had higher capitulum curvature, ranging from 1-1.2 as opposed to 0.9-1.05 in S2 and S3 (Figure 9E). S4 showed intermediate values for all three of the most influential characters in univariate results, except lower ray fruit lengths than S1, which explains the level of overlap in the LDA. Most patterns were repeated for the LDA of the K=4 assignment (Figure 8A); but here there was significantly more overlap between P4 (S4,S5) and P2 (S2), a reduced influence of appendage length in driving separation of groups on LD2 (LC = 0.60 vs -1.37, Table 3) and an increased influence of capitulum curvature (LC = 0.99 vs 0.70, Table 3).

The LDA of morphology thus produced the following groups: (i) S1 and S2, long ray fruit lengths and elongated disk appendages; (ii) S3, short ray fruits and intermediate length appendages; (iii) S5, intermediate ray fruit lengths, shorter disk appendages with high capitulum curvature. S4 could not be delimited in any of these morphological groupings and did not group well with individuals of the S5 cluster. Additional differences which were thought to be potentially informative such as ray number, ray length, ray width, appendage diameter and inner disk tube length showed large overlap between most clusters on boxplots with a few consistent patterns

(Figure 9). The range of ray numbers differed dramatically within the complex (10-21 rays), with S3 and S5 having a particularly large variance (11-21 rays). Ray length and width variance was relatively large in S1 (17.5-42.5 mm and 5-11mm respectively, Figure 9B & C) in comparison to other clusters. Lastly, appendage diameter was near 0 in S1 and S2 (Figure 9F), with larger values in S3 and S5.

The LDA of environmental data for the K=5 assignment clearly differentiated S1, S3 and S5 with some overlap between S2 and S3 as well as S4 with S5 and S1 (Figure 10B). These differences were primarily driven by MAP, PWQ and elevation on LD1 (LC=2.99, -1.10 and -1.01 respectively, Table 4), and soil clay levels on LD2 (LC= -0.99). Supporting this, these variables had minimal overlap in univariate comparisons, with precipitation, elevation and clay levels grouping clusters differently (Figure 11). MAP grouped S1 with S4 (300-550 mm), S2 with S3 (100-200 mm) and had S5 distinct (250 mm). PWQ showed identical trends to MAP. Elevation grouped S1, S2 with S4 (0-250 m above sea level) and S3 with S5 as high elevation clusters (250-750 m). S5 cluster localities had higher clay levels (8-15%) in contrast to more sandy soils in other clusters. Thus, under the K=5 assignment, clusters exhibited ecological niche separation as follows: (i) S1 and S4, high precipitation, low elevation, sandy soils (ii) S2, low precipitation, low elevation, sandy soils; (iii) S3, high elevation, low precipitation, sandy soils and (iv) S5, high elevation, intermediate precipitation, clay soils. LDA of the K=4 assignment (Figure 10A) showed larger overlap between the P4 (S4, S5) cluster and P1 (S1). Influential variables were similar to the K=5 assignment, but pH was more influential on LD2 (LC=-0.93) than clay (LC=0.64). Ecological niche partitioning was thus not present in the K=4 assignment.

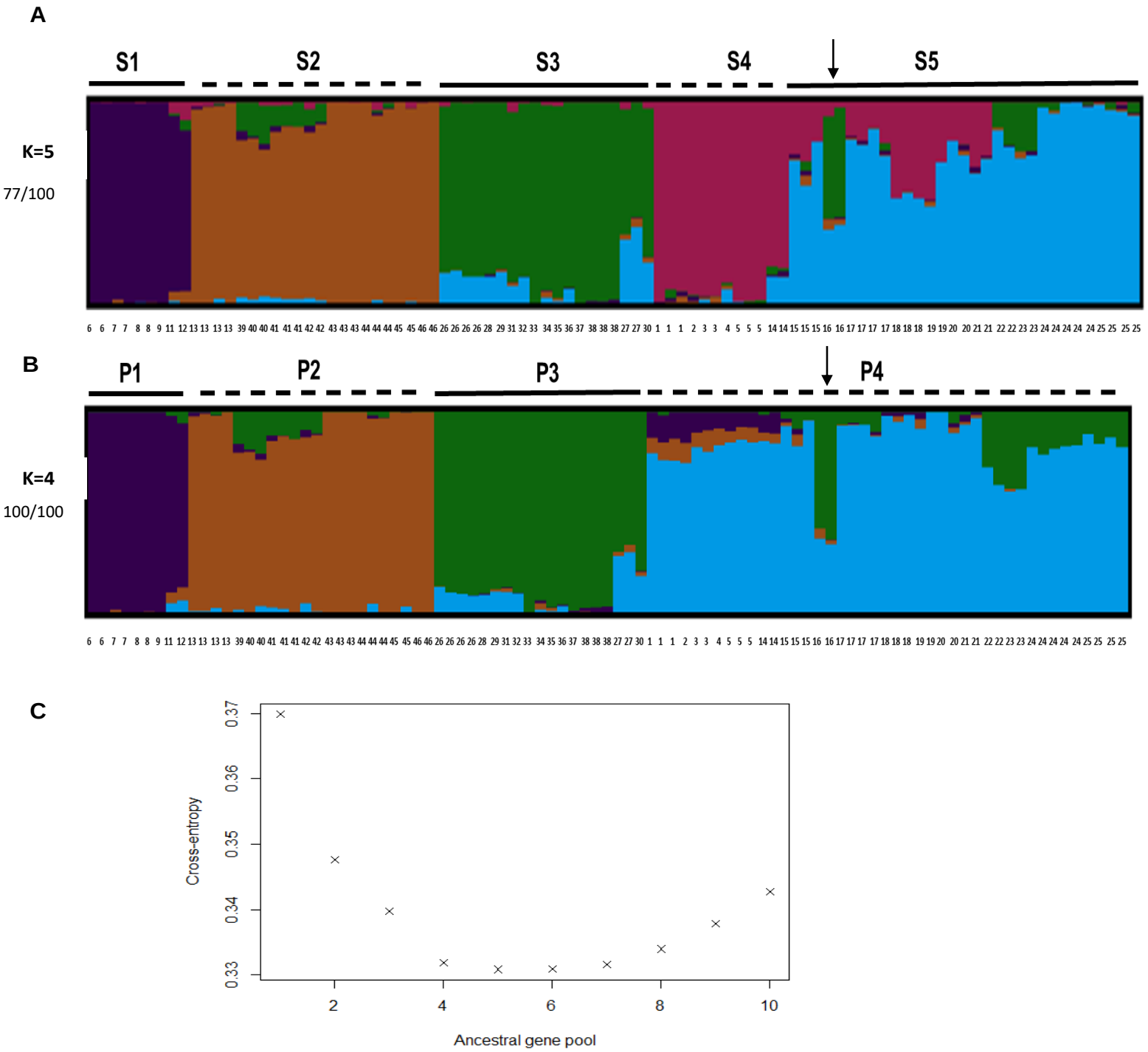


Figure 4: Plots of sNMF genomic assignment for the *Dimorpotheca pluvialis-sinuata* complex for K (number of ancestral pools) with two lowest entropies: (A) K = 5 (B) K = 4. Each individual is represented by a single vertical bar, different ancestral gene pools are indicated by different colours, numbers below indicate the population code of each individual (Appendix A). Group assignments above bars indicate genetic cluster hypothesis used for subsequent analysis; with arrows outlier population 16 which was assigned to cluster S3/P3. Only major mode ancestry coefficient summaries are shown. Ratios to the left of boxes indicate the percentage runs out of a total of 100 runs supporting major modes, (C) The cross-entropy criterion plotted for each value of K between K = 2 and K = 10.

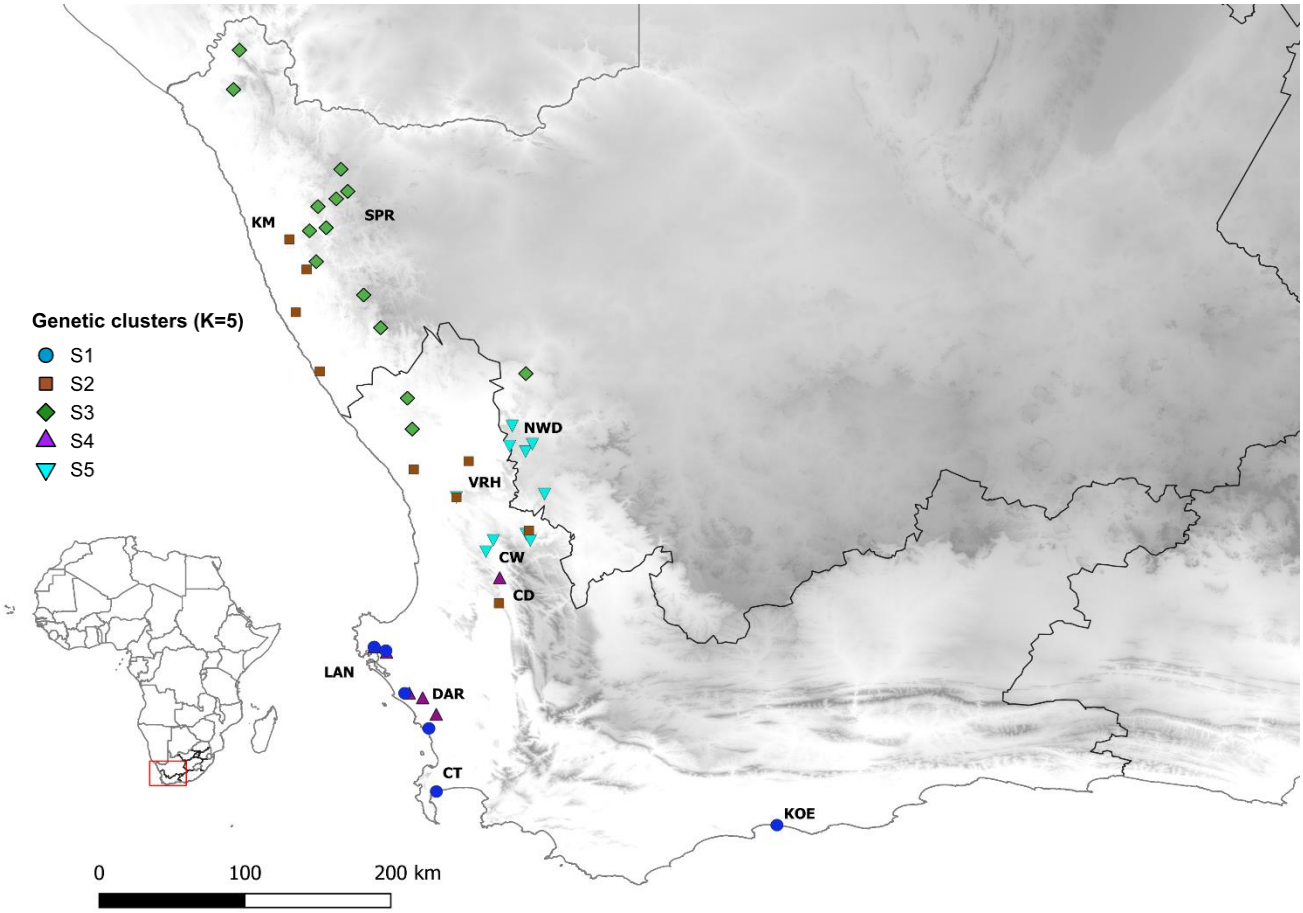


Figure 5: Distribution of populations within each genetic cluster, symbols and colours differentiating clusters. Elevation overlayed, darker areas showing elevation of over 300 m above sea level. Letters showing approximate locations of important towns: CT – Cape Town, LAN – Langebaan, DAR – Darling, CD – Citrusdal, CW - Clanwilliam, VRH – Vanrhynsdorp, NWD – Nieuwoudtville, KM – Komaggas, SPR – Springbok.

Table 1: Pairwise ***Fst***^{WC} between genetic clusters for K=4 hypothesis; clusters as populations

	Genetic Clusters (K=4)		
	P1	**P2**	**P3**
Fst^{WC}			
P2	0.154		
P3	0.173	0.130	
P4	0.178	0.111	0.060

Table 2: Pairwise ***F***st^{WC} between genetic clusters for K=5 hypothesis; clusters as populations

	Genetic Clusters (K=5)			
	S1	**S2**	**S3**	**S4**
Fst^{WC}				
S2	0.154			
S3	0.173	0.130		
S4	0.178	0.136	0.105	
S5	0.162	0.118	0.060	0.062

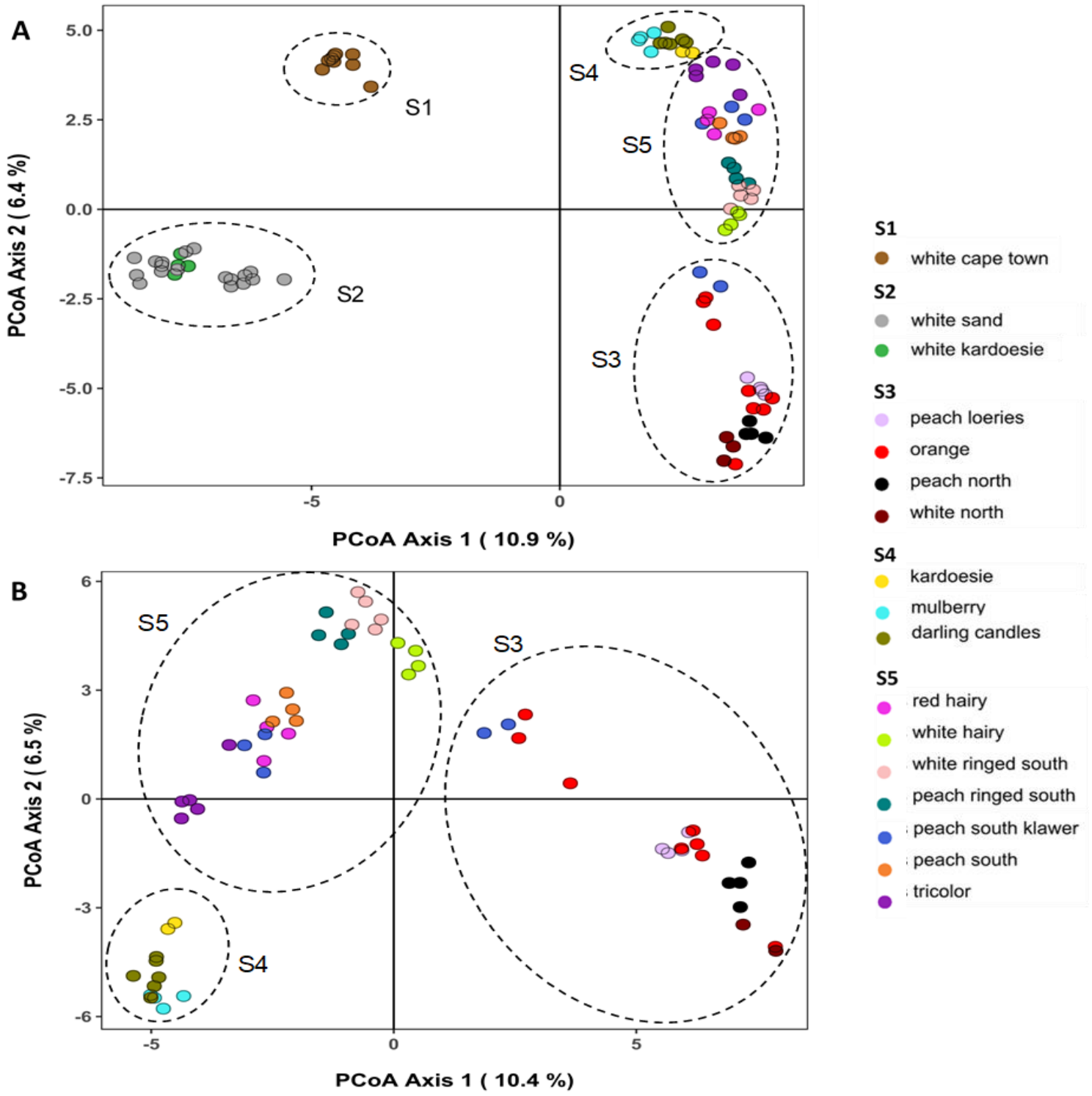


Figure 6: Principle Coordinate Analysis (PCoA) of individuals based on SNP dataset with 8506 SNPS; Axes show the first two principal components and the percentage variance explained. Ellipses arbitrarily show sMNF genetic clusters (K=5). Colours indicate *a priori* morphs assigned to the complex before sampling. (A) Analysis of all individuals (B) Only cluster S3, S4 and S5

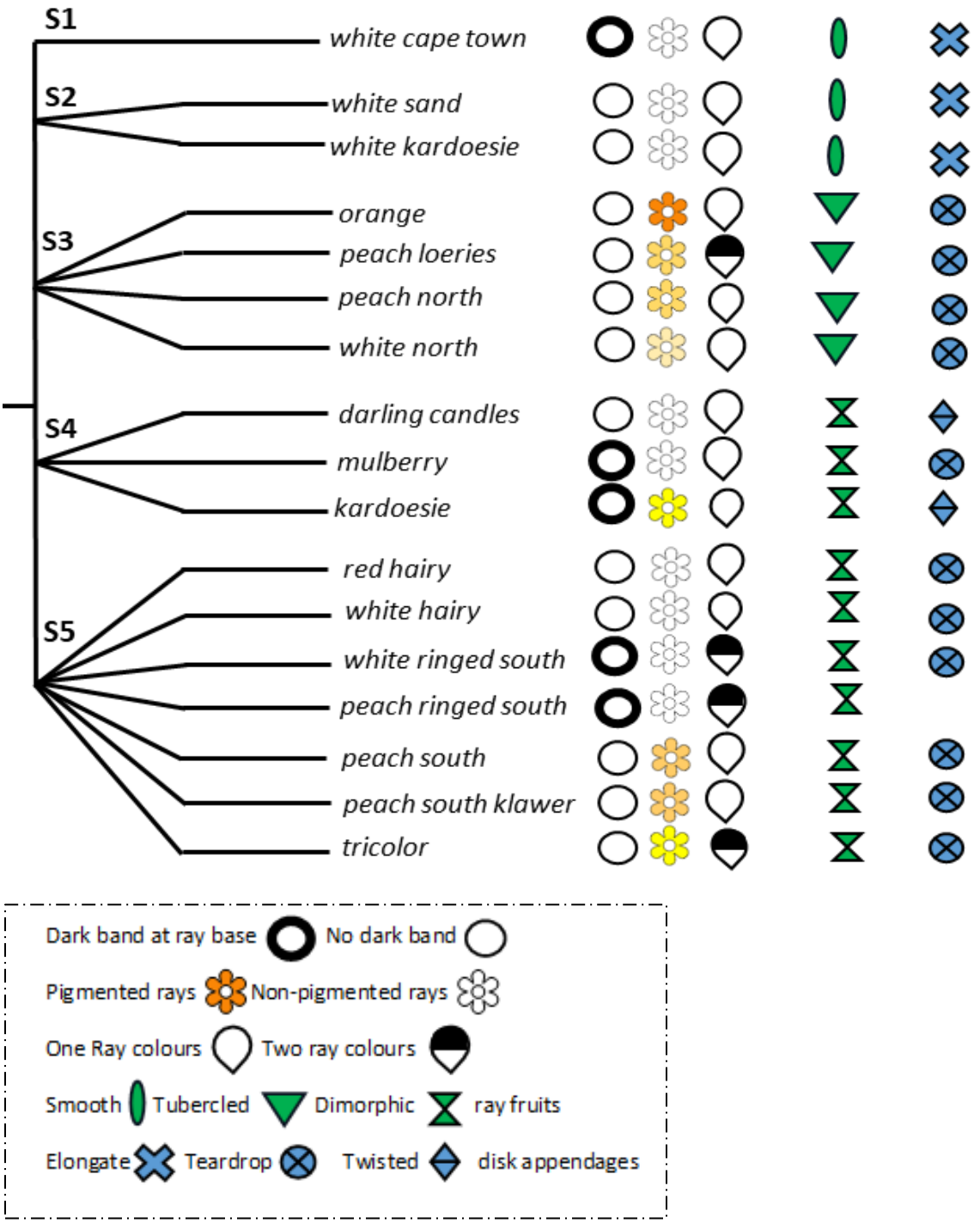


Figure 7: Differences in discrete morphological characters between a priori morphs, grouped by sMNF genetic clusters (K=5). Difference in ray colour patterns, ray fruit structuring and inner disk appendage length structures shown. No symbol indicates absence of this character.

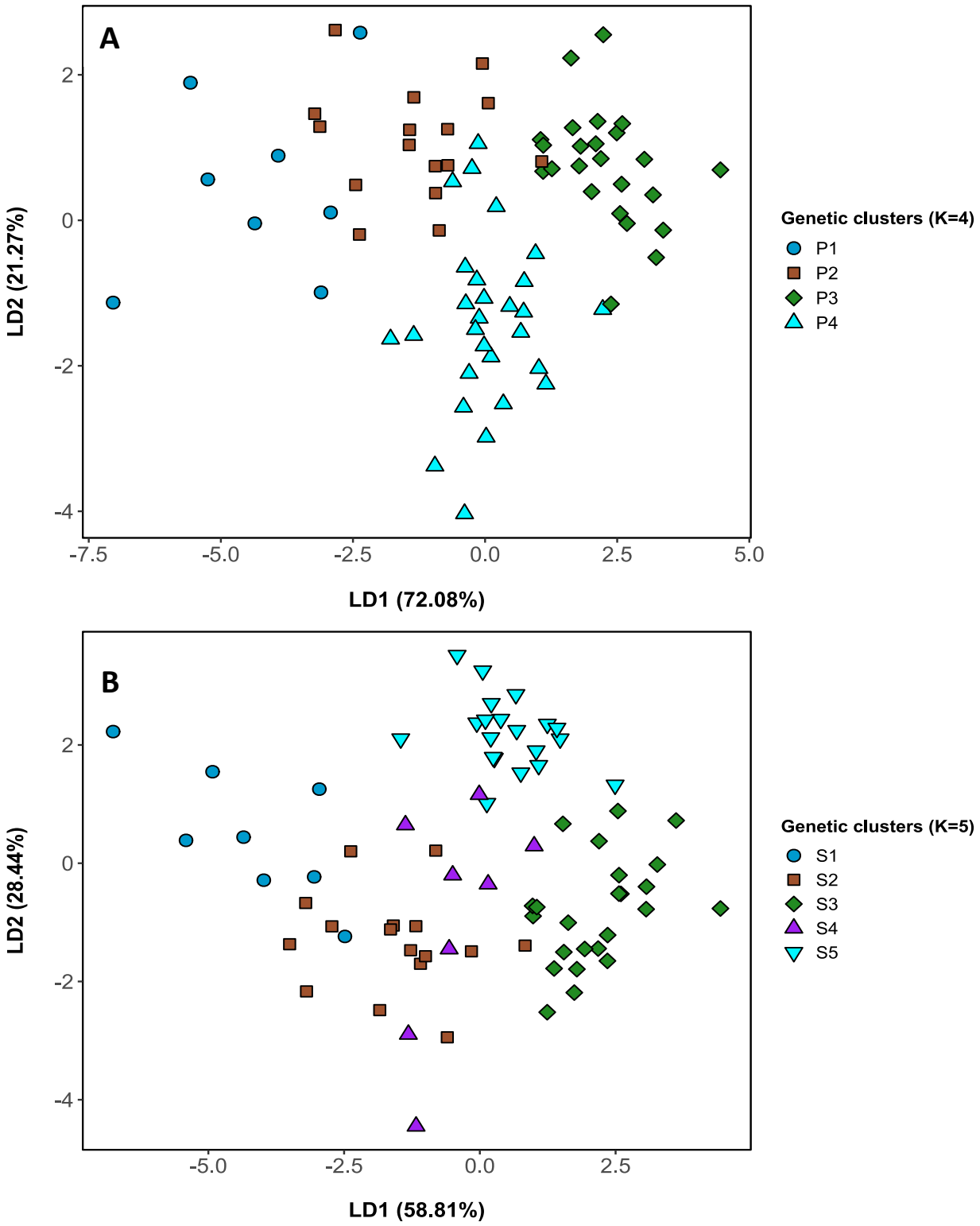


Figure 8: Linear Discriminant Analysis (LDA) of morphological characters for two individuals from each of the 45 populations using sNMF genetic clusters as grouping factor; colors and cluster names consistent with Figure 1; (A) K=4 hypothesis (B) K=5 hypothesis

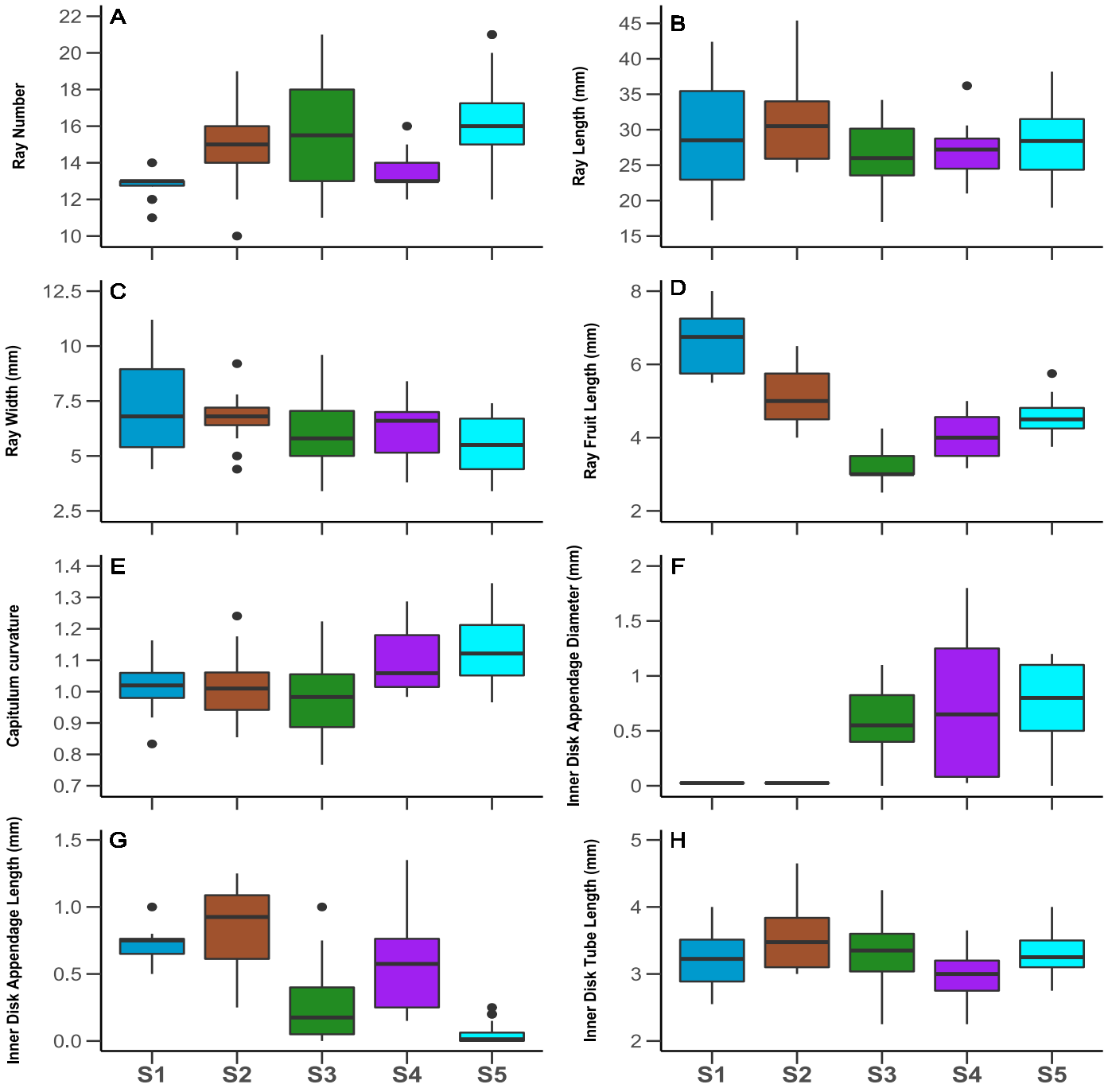


Figure 9: Boxplots of difference in key morphological traits between K=5 sNMF clusters, cluster names and colors consistent with Figure 1A. Traits include: (A) Ray number (B) Ray length (C) Ray width (D) Ray fruit length (E) Capitulum curvature (F) Inner disk appendage diameter (G) Inner disk appendage length (H) Inner disk tube length

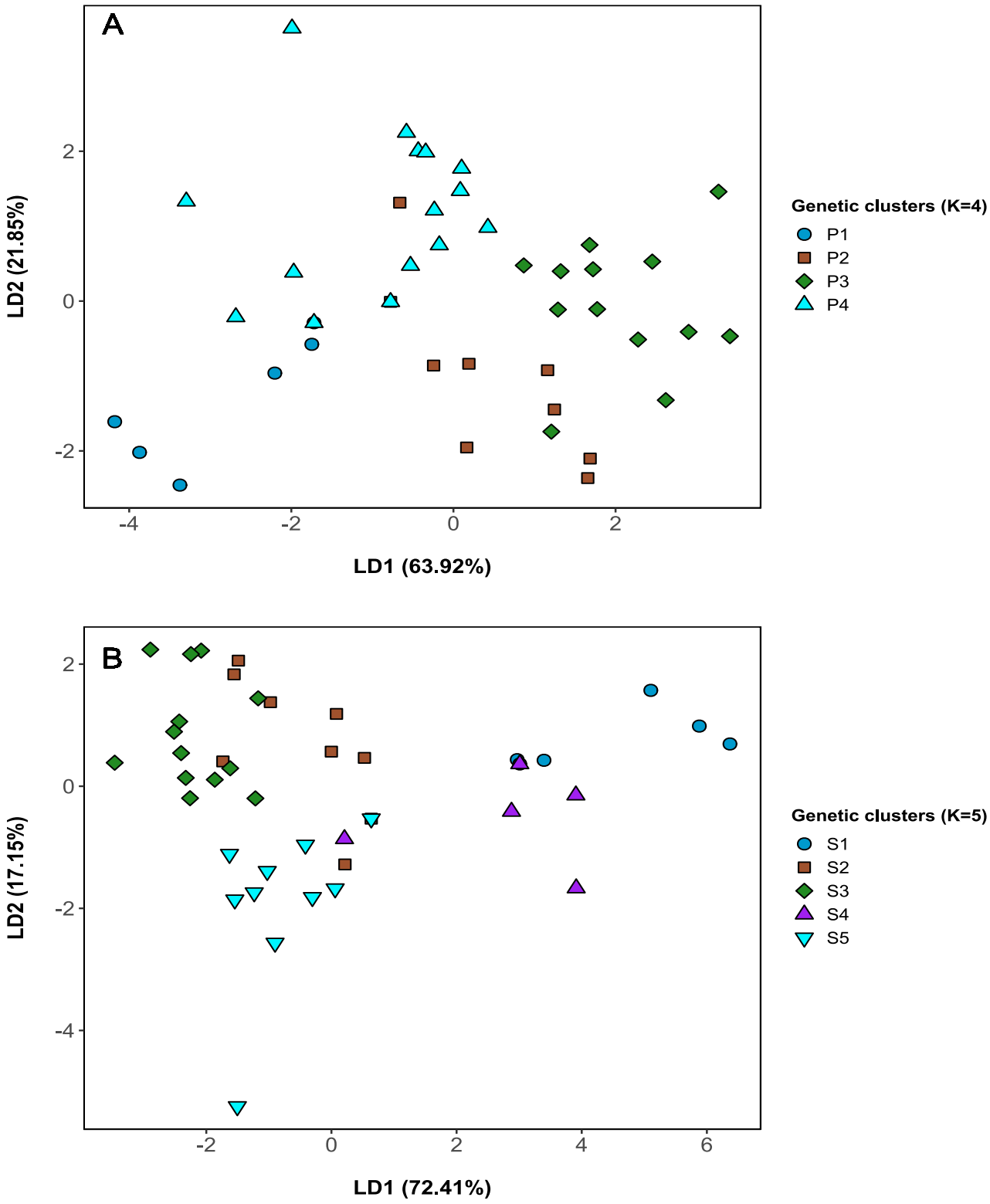


Figure 10: Linear Discriminant Analysis (LDA) of environmental variables at population locations using sNMF genetic clusters as grouping factor; colors and cluster names consistent with Figure 1 (A) K=4 hypothesis (B) K=5 hypothesis

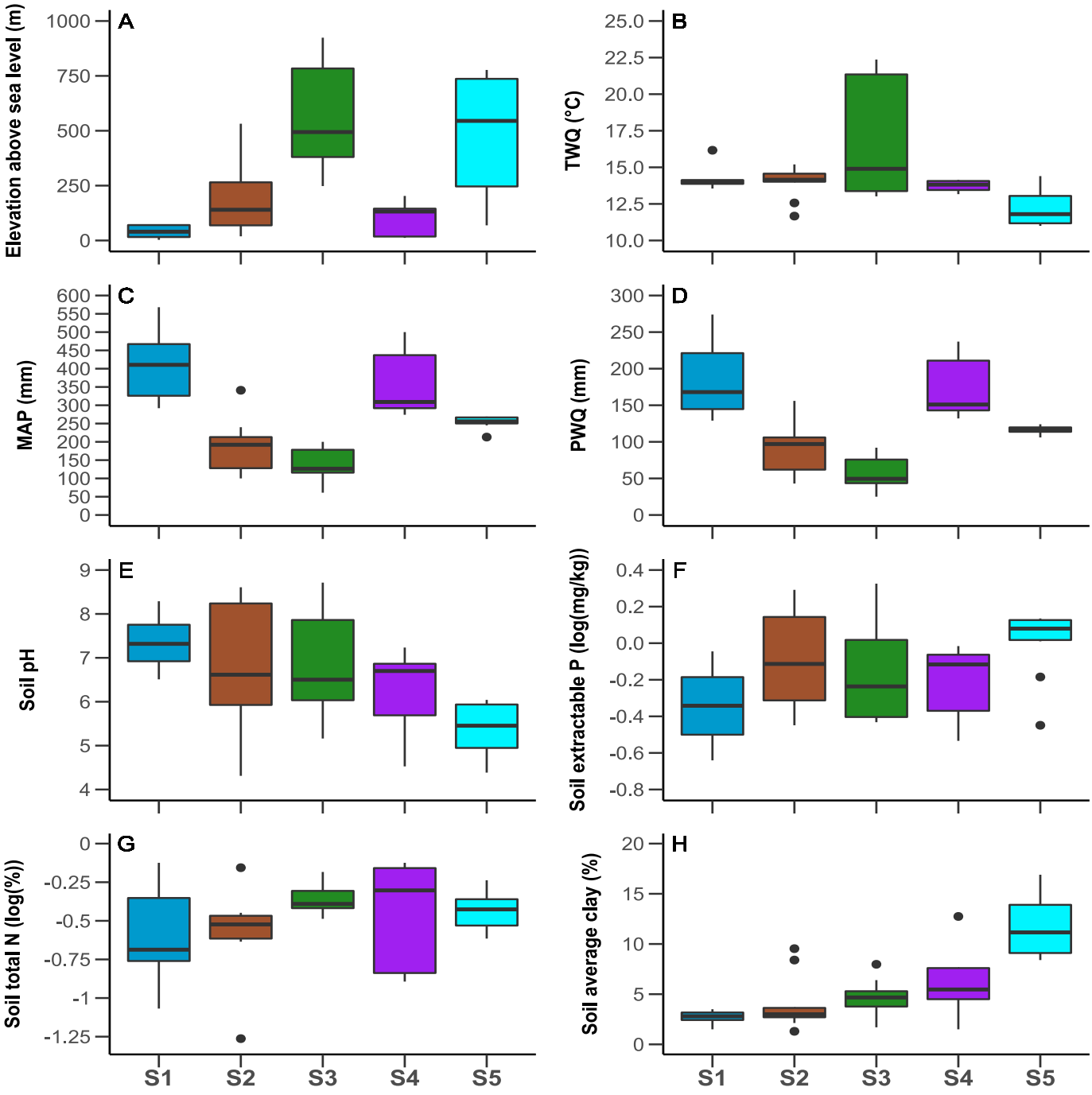


Figure 11: Boxplots of difference in key environmental traits between K=5 sNMF clusters, cluster names and colors consistent with Figure 1A. Variables include: (A) Elevation above sea level (B) Temperature of Wettest Quarter (C) Mean Annual Precipitation (D) Precipitation of Wettest Quarter (E) Soil pH (F) Soil extractable phosphorous (G) Soil total nitrogen (G) Soil average clay

Table 3: LDA linear coefficients of continuous variables for both morphological and environmental models where K=4 and K=5 sMNF genetic clusters, LD1 and LD2 showing first and second linear discriminant axis respectively.

	Variables	LDA model linear coefficients			
		K=4		K=5	
		LD1	LD2	LD1	LD2
Morphological	Bract Length	0.212	-0.376	0.267	0.343
	Ray Number	0.171	0.120	0.164	-0.111
	Ray Length	-0.022	-0.208	0.041	0.304
	Ray Width	-0.329	0.346	-0.408	-0.390
	Inner Disk Appendage Diameter	0.127	-0.332	0.135	0.177
	Inner Disk Appendage Length	-0.154	0.596	-0.449	-1.370
	Inner Disk Tube Length	0.282	0.850	0.246	-0.522
	Inner Disk Lobe Length	0.465	0.291	0.455	-0.212
	Outer Disk Tube Length	-0.139	-0.364	-0.083	0.386
	Outer Disk Lobe Length	-0.393	-0.767	-0.410	0.269
	Capitulum curvature	-0.531	-0.990	-0.462	0.700
	Disk Fruit Length	-0.102	-0.157	0.022	0.566
	Ray Fruit Length	-1.829	-0.320	-1.745	0.596
Environmental	Elevation	0.770	0.219	-1.013	0.099
	Mean Annual Precipitation (MAP)	-2.249	-0.448	2.995	0.224
	Precipitation of Wettest Quarter (PWQ)	0.959	0.791	-1.104	-0.478
	Temperature of Wettest Quarter (TWQ)	0.390	0.387	-0.035	0.204
	pH	-0.190	-0.937	0.350	0.569
	Soil extractable P	0.369	0.053	-0.427	0.046
	Soil total N	-0.084	0.653	0.177	-0.303
	Soil average clay	-0.539	0.646	-0.032	-0.994

DISCUSSION

Genetic data reveals the potential for recognizing four to five genetically distinct lineages within the *Dimorphotheca pluvialis-sinuata* species complex, the existence of which constitutes support for putative evolutionary species as described by Wiley (1978) and De Queiroz (2007). However, the best species-delimitation scheme (4 vs. 5 species) is unclear when considering the marginal cross-entropy difference between the two best-supported sMNF hypotheses (K=4 and K=5) for genetic cluster assignment. Both of these hypotheses recognize three identical genetic clusters, each of which is geographically isolated (Figure 5) as would be expected in allopatric speciation (Mayr, 1963): (i) S1/P1, the white form from the coastal Cape Town area which is likely to correspond with the type specimen of D. pluvialis, although no locality is specified by Linneaus (1753); (ii) S2/P2, a northern white-rayed sand form; and (iii) S3/P3, a group of forms with orange and peach rays, corresponding largely to the type specimen *D. sinuata* (which was collected at "Silwerfontein", close to the town of Springbok, by Drège in the 1830's). The discrepancy arises when considering the distinctness or uniformity of the S4 cluster (comprising southern forms with variable disk floret appendages and ray colours) and the S5 cluster (comprising several morphologically distinctive floral morphs from the Cederberg and Nieuwouldtville area). These are concatenated into one cluster, P4, under the K=4 hypothesis. While the K=5 hypothesis has the lowest cross-entropy, the degree of admixture between these two clusters corresponds to lower geographic distance between populations, which indicates the potential for significant gene flow in sympatry impacting delimitation of the clusters as evolutionary species (Petit & Excoffier, 2009). The stronger agreement by sMNF on the four-species assignment adds further complexity as it indicates a lack of certainty in ancestral gene pool assignment under the five-species hypothesis. Considering that the genetic clusters produced from sMNF were used as the underlying hypotheses for the analysis, it is necessary to evaluate support for both a five and four-species model outside strictly the sNMF results-primarily through PCoA and pairwise $F\text{st}^{WC}$ (Weir and Cockerham's estimator).

The PCoA of all individuals can at the onset be interpreted in favor of only three evolutionary species, but subsetting out S1/P1 and S2/P2 shows that the large degree of genetic variation

between these two species and the remaining species masks the differentiation between S3/P3 and P4 (S4 and S5). While an argument could be made that the genetic distance between S4 and S5 is equivalent to internal groupings within S5 and S3, hence four species would be more appropriate, if one applies this reasoning there would be scope to group individuals of S3 with S5 as well, which would result in large genetic distance between individuals within the same species. Thus, interpretation of the PCoA in this manner favours a broad three-species approach, which was not supported by sMNF and I feel does not accurately represent genetic variation within the complex. Using a five-species model, there is equivalent genetic distance between S3-S5 and S4-S5 in multivariate space which was mirrored by similar pairwise $F\text{st}^{WC}$ values for the S3-S5 and S4-S5 comparison. The five-species model also more accurately portrays the relationship between genetic distance and geographic distance. Notable as well is that while interpretation of pairwise $F\text{st}^{WC}$ is often context-dependent, a value of between 0.1 and 0.2 should be considered strong differentiation (Lemopoulos et al., 2019). That the majority of pairwise comparisons fall within this range further supports the sNMF results, particularly when considering that the distribution of individuals within each cluster exceeds the typical distribution of populations in $F\text{st}^{WC}$ calculations, likely resulting in greater within population variance (Leviyang & Hamilton, 2011, Willing et al., 2012). In conjunction these results suggest that combining S4 and S5 would result in lost genetic variation which could be informative, notwithstanding morphological and environmental considerations.

Morphological analysis failed to clearly distinguish between all five species, showing three major groups: S1 with S2, S3 and S5. Differences between these groups were in many cases subtle and based on only a few influential traits: ray fruit length and structuring, inner disk floret appendage differentiation and differentiation in height between the inner and outer capitulum (i.e. capitulum curvature). S1 and S2, the two species with the greatest genetic differentiation, were morphologically very similar which is a clear case of cryptic species (Crawford & Stuessy, 2016, Fourie et al., 2014). The strong genetic division of S1 and S2 from the remaining species is however reinforced by clear morphological differences; these forms have relatively smooth, long ray fruits and elongate inner disk floret appendages with no flattened tips. S3 in turn could be strongly characterized by small, strongly tubercled ray fruits.

The S5 species, while highly variable in ray colour and disk ornamentation, could be most effectively differentiated from S3 on the basis of having dimorphic (smooth and tubercled) intermediate length fruits, flattened 'teardrop' appendages which are directly attached to the lobe as opposed to being slightly elongated in S3, and inner disk florets which jut out (high capitulum curvature). However, these characters cannot be seen as definitive for all S5, as they are not uniform across all forms. For example, the influence of capitulum curvature is likely due to the distinctive inner capitulum presentation of the 'red hairy' and 'white hairy' morphs (Figure 1), and the 'peach-ringed south' morph exhibited no appendage structures. In addition, dimorphic fruits are present in S4 as well. S4 in general was the most poorly resolved species in terms of morphology, exhibiting high variance between morphs in all the key influential characters, meaning it could not be grouped with any one species. Thus, while continuous morphological analysis could discriminate between three major groups, only S1+S2 and S3 had uniform defining characteristics across all morphs. What is clear from analysis of morphology is that the current classification of the complex by Norlindh (1943) based on ray colour fails to define the natural taxa and that some of the most influential characters, primarily differences in ray fruits, were not considered by him.

The high degree of ecological niche separation between localities of the proposed species (barring S1 some overlap in S4) based on differing elevation, precipitation and soil clay provided greater support than morphology for a five-species approach. Species occupying different adaptive zones are typically thought to have become distinct through differential selection (Andersson, 1990), and as such ecological niche separation can be viewed as evidence for genetic species delimitation. To support this, recent studies in Asteraceae have shown a relationship between divergence of microsatellite loci and ecological divergence (Friar et al., 2006) as well as genome size and ecological pressures (Torrell & Vallès, 2001). Assessment of ecological niche space differentiation can thus provide a useful framework for supporting genetic divergence as well as resolving discrepancies in morphological variation. For example, the high difference in precipitation between S1 and S2 localities provides support for genetic divergence not seen in morphology. In addition, consistent differences in elevation between S2 and S3 localities provides support for why there is minimal gene flow between these

populations despite in many cases low geographic distance between populations (Chapman et al., 2013). Difference in elevation in particular have been shown to correspond to differing flowering and germination times in semi-arid winter annuals (Crimmins et al., 2010) providing support for speciation within the complex. In general, I think that despite morphological overlap between the five species, when viewed in conjunction with the distinct ecological niche separation, the integrated taxonomic approach applied here supports a five-species model for *Dimorphotheca pluvialis-sinuata.*

The study thus supports revisiting the existing classification of *Dimorphotheca pluvialis-sinuata* as comprising of two species. This current approach is extremely conservative in being based on traditional, morphology-based approach which is not geared towards the identification of independently evolving metapopulation lineages (Wiley, 1978). While formal taxonomic reclassification is beyond the scope of this study's analysis, I here propose the existence of five distinct evolutionary species within the *Dimorphotheca pluvialis-sinuata* complex; these correspond to the five-species model and reflect entities with varying degrees of morphological distinction and occupy fairly distinct ecological niches: (i) **S1,** a white rayed form with long, smooth ray fruits and elongate inner disk appendages, distributed in the low elevation, high precipitation Cape coastal sands region; (ii) **S2**, forms which were highly genetically distinct but morphologically indistinguishable from S1, primarily distributed in the coastal red sands of Namaqualand; (iii) **S3**, forms with uniform orange or peach rays and most strongly characterized by highly tubercled, small ray fruits, primarily distributed in the high elevation, low precipitation red sands of the Namaqualand Hardeveld escarpment; (iv) **S4**, forms which lacked morphological consistency, primarily distributed in intermediate-high precipitation, lowland Darling/Langebaan and crossing into the southern Olifants valley; (v) **S5**, highly morphologically variable forms with a wide variety of ray colours and disk floret ornamentation, with most forms characterized by flattened 'teardrop' inner disk appendages and a raised inner capitulum, distributed in the high elevation, clay soils of the Cederberg and Vanrhynsdorp area.

While these groupings serve as a baseline hypothesis for evolutionary species, I believe that it would be prudent to conduct Bayesian species delimitation on the complex as well, which could produce a greater number of species as seen in numerous studies which utilize the

method in a similar integrated framework to this study (Solís-Lemus et al., 2015, Yang & Rannala, 2010, Zhang et al., 2011). This is particularly relevant to the S5 gene pool, which exhibited low genetic variation between morphs in my analysis but could potentially be delimited into multiple species based on morphology alone. The degree of genetic distance and morphological variability between S1 and S5 also raises the possibility that the *D. pluvialis-sinuata* complex is not monophyletic, with some of its component species potentially being more closely related to perennial species of *Dimorphotheca*. The complex as a whole presents potentially interesting questions with regards to evolutionary mechanisms which could help resolve the link between genetic, morphological and ecological differentiation; such as the effect of pollinator distribution and the role of habitat in fruit structuring. For example, dimorphism of ray fruits as seen in S4 and S5 has been shown to be environmentally adaptive in other Asteraceae (de Clavijo, 1995), and differentiation in flower color could affect pollinator selection within plant communities (Chittka & Menzel, 1992, Cooley et al., 2008, de Jager et al., 2011) and subsequently speciation.

The study thus paves the way for future work in *Dimorphotheca pluvialis-sinuata* complex and *Dimorphotheca* as a whole. While I am confident that the five species I have proposed here are a good representation of evolutionary species within the complex, I do believe there is scope for a more detailed and comprehensive analysis of morphology: particularly with regards to ray and disk fruit characters, continuous ray colour characters as well as vegetative characters. In addition to Bayesian species delimitation, I would also like to add on to genetic data with more populations encompassing a greater geographic distribution. A combination of a stronger morphological and genetic framework for species delimitation will allow me to formally reclassify the taxonomy of the species complex. In general, I believe that an integrated approach as implemented here shows how genetic data can be highly informative in resolving variation, particularly in the context of a complex of semi-cryptic species, and in the context of diversification in South African flora as a whole.

Acknowledgements

I am thankful to my supervisors Assoc. Prof Tony Verboom, Dr Nicola Bergh and Prof. Allan Ellis for their immense help and support in completion of this project. Special thanks as well to Ms Zaynab Shaik and Mr Seth Musker for their tireless efforts in assisting me with SNP data analysis. Lastly, thanks to all students of Allan who assisted in field collection.

APPENDIX A

Population codes of 45 sampled localities in South Africa with associated co-ordinates, locality description and a priori morph allocation

Population	Morph	Locality	Latitude	Longitude (deg)
1	ds mulberry	Saldanha (road to Cape Town)	-33.014	18.102
2	ds mulberry	Saldanha	-32.972	18.009
3	ds darling candles	Darling		
4	ds darling candles	Darling	-33.370	18.374
5	ds darling candles	Mamre	-33.500	18.476
6	dp white cape town	Saldanha (Weskus Padstaal)	-33.334	18.239
7	dp white cape town	Saldanha (road to Cape Town)	-33.001	18.097
8	dp white cape town	Saldanha (steel plant)	-32.972	18.009
9	dp white cape town	Melkbos	-33.610	18.420
11	dp white cape town	Muizenberg	-34.104	18.477
12	dp white cape town	Koensrust	-34.367	21.038
13	dp white kardoesie	Piekinerskloof Pass, Kardoesie	-32.631	18.948
14	ds kardoesie	Kardoesie, Citrusdaal	-32.430	18.953
15	ds peach south klawer	Klawer south (N7 roadside)	-31.801	18.627
16	ds peach south klawer	Nuwerus (Road to Lutzville)	-31.265	18.295
17	ds peach south	Bottertop, Cederberg	-31.775	19.290
18	ds tricolor	Clanwilliam (dam)	-32.226	18.849
19	ds tricolor	Clanwilliam	-32.138	18.905
20	ds red hairy	Biedouw valley top	-32.089	19.153
21	ds red hairy	Biedouw valley bottom	-32.139	19.183
22	ds white hairy	Nieuwoudtville	-31.442	19.148
23	ds white hairy	Nieuwoudtville	-31.384	19.199
24	ds white ringed south	Grasberg, Nieuwoudtville	-31.399	19.028
25	ds peach ringed south	Vanrhynsdorp pass	-31.241	19.047
26	ds peach loeries	Loeriesfontein	-30.831	19.150
27	ds orange	Bitterfontein	-31.024	18.259
28	ds orange	Studer's Pass top	-30.472	18.057
29	ds orange	Komaggas	-29.712	17.523
30	ds orange	Koingnaas	-29.954	17.573
31	ds orange	Nigramoep air field	-29.523	17.586
32	ds orange	Kamieskroon	-30.215	17.930
33	ds orange	Naries Retreat, Springbok	-29.688	17.648
34	ds peach north	Springbok to Steenkopf road	-29.403	17.810
35	ds peach north	Bulletrap to Nigramoep road	-29.462	17.723
36	ds peach north	Steinkopf	-29.230	17.759
37	ds peach north	Khubus	-28.605	16.950
38	ds white north	Sendelingsdrift	-28.297	16.995
39	dp white sand	Between Komaggas and	-29.779	17.371
40	dp white sand	Meselpad to Koingnaas road	-30.016	17.500
41	dp white sand	Hondeklip to Gariesbaai road	-30.351	17.419
42	dp white sand	Groenriviersmond	-30.813	17.600
43	dp white sand	Lutzville	-31.582	18.307
44	dp white sand	Klawer south	-31.801	18.627
45	dp white sand	Vanrhynsdorp	-31.517	18.719
46	dp white sand	Doringbos, Cedeberg	-32.060	19.173

APPENDIX B

Floral character descriptions of *a priori* morphs from the *Dimorphotheca pluvialis-sinuata* complex used for field-based sampling

Morph name	Floral character description
white cape town	White rays, small black ray base fading to pink. Gaps between rays. Disk appendages elongated and black. Involucre with long white hairs.
white sand	White rays, small black base fading to blue. Rays long with no gaps. Disk appendages elongated and black. Some tubercled nuts. Involucre with fine hairs
white kardoesie	White rays, rounded with small blue base. Similar to dp white sand but disk appendages yellow.
mulberry	White rays, large purple base, small flat inflorescence. Disk appendages have teardrop appearance, very expanded and protruding prominently above the plane of flower.
darling candles	Similar to ds mulberry; but have twisted elongated disk appendages
tricolor	Rays tricolored; orange tips, yellow center, purple black base. Disk appendages with flattened teardrop appearance
peach south klawer	Rays salmon-orange with darker red-purple ring at base. Disk appendages with flattened teardrop appearance.
peach south	Rays peach brown with whiter lower part, base brown. Short ray tube. Disk appendages with flattened teardrop appearance. Involucral hairs long and black
kardoesie	Peach upper rays with wide red-purple base. Disk appendages black, twisted and elongate.
red hairy	Rays white with peach tips, base brownish. Back of rays purple, short ray tube. Central disks with black teardrop appendages which are black but small in diameter. Central disk corolla tube with short brush-like hairs, protruding above plane of the flower. Involucre with long black hairs.
white hairy	Rays white with golden base. Central disk with teardrop appendages, protruding prominently with well-developed ring of hairs.
white ringed south	Rays upper white with peach tips, large black base with blue-green sheen; outer yellowish with blackish streaks. Undifferentiated disk appendages
peach ringed south	Rays darker peach tip, pale peach base, purple basal part. Undifferentiated disk appendages.
orange	Rays glossy orange with small black basal part; backside orange. Disk appendages lack, either teardrop or twisted. Blacks hairless.
peach north	Similar to ds orange, but peach rays with peach/white pollen
white north	Similar to ds orange, but white-ish peach rays with white pollen
peach loeries	Rays upper peach with white centre and small black base, outer dull peach with green streaks. Inner disks with teardrop appendages, purple upper half. Pollen white

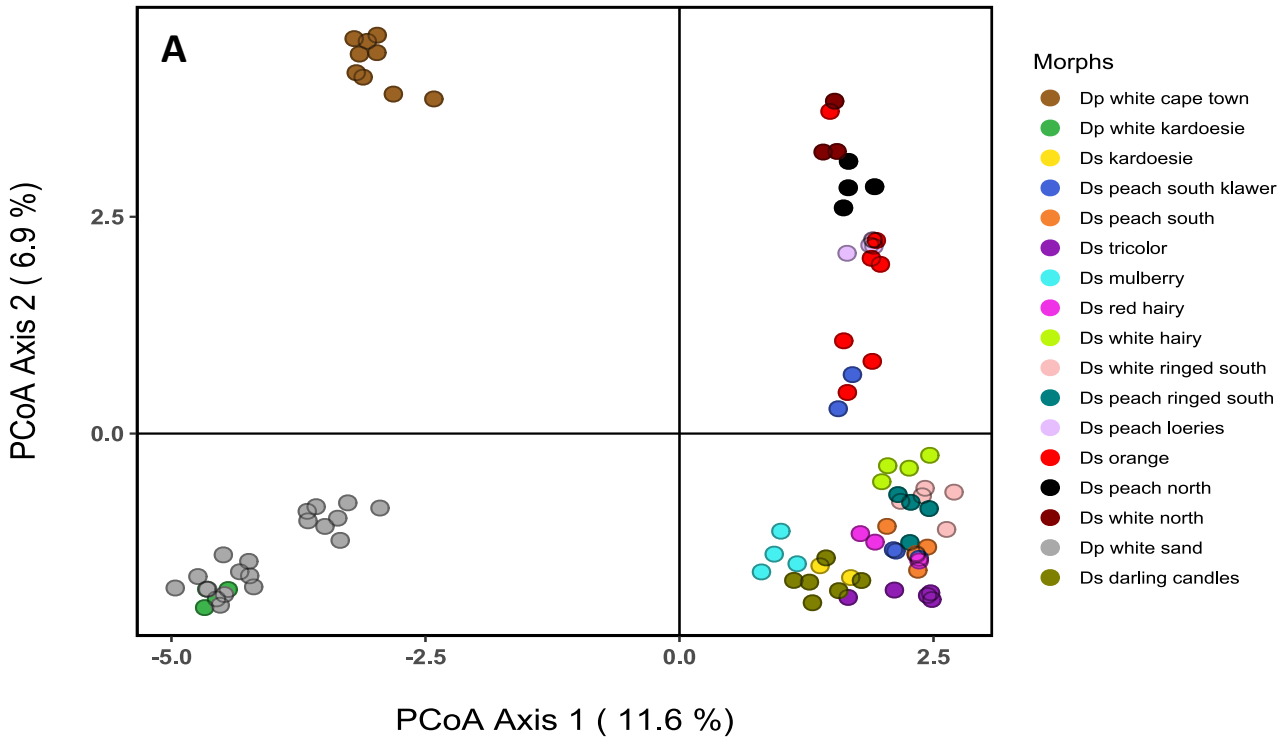
APPENDIX C

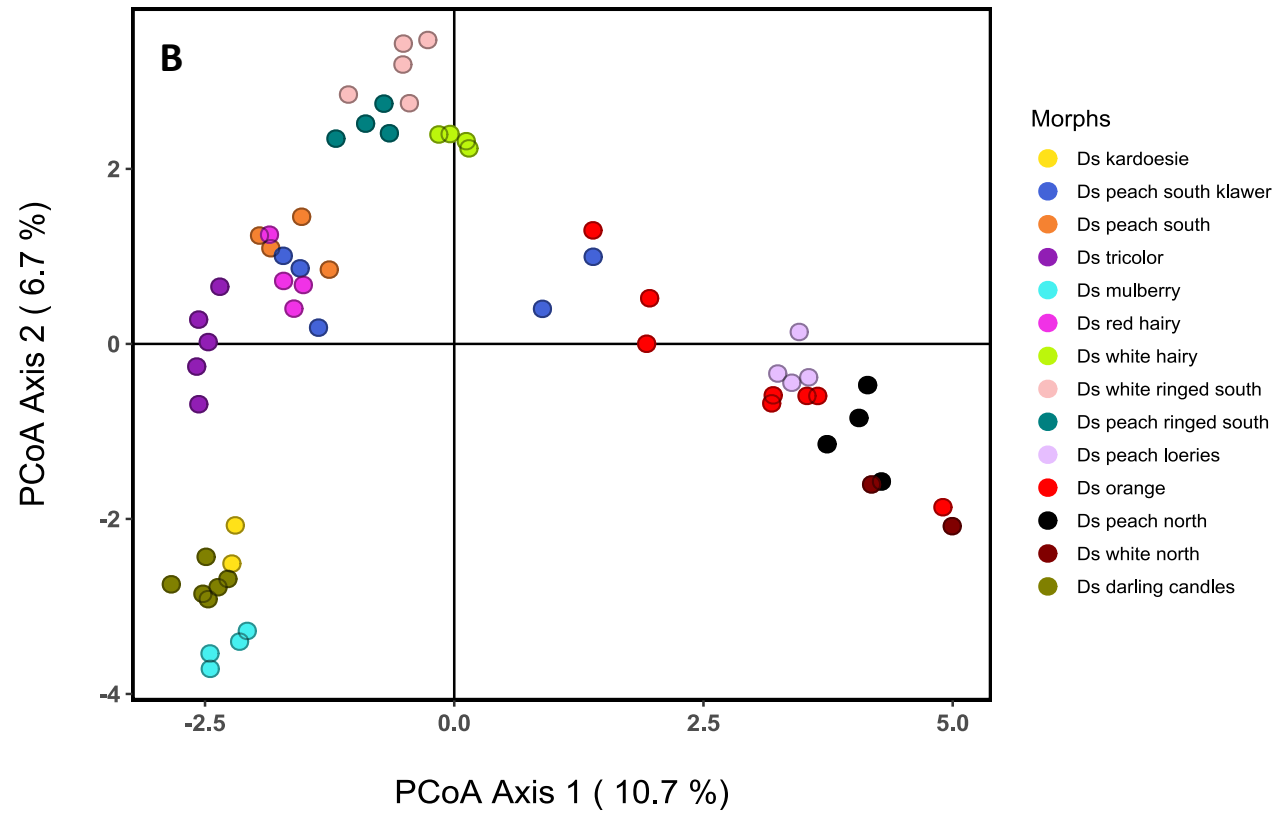
Description of flower characters used in analysis of morphology, characters dropped from analysis due to collinearity or being uninformative highlighted in red

Flower character	Description
Bract Length	Length of capitulum involucral bract
Ray Number	Number of rays
Ray Length	Length ray florets (corolla tube and ligule)
Ray Width	Width of ray ligule at the widest point
Inner Disk Appendage Diameter	Diameter of inner disk appendages as viewed from the top
Inner Disk Appendage Length	Length of inner disk appendages from point of attachment to lobe upwards
Inner Disk Tube Length	Length of inner disk corolla tube
Inner Disk Lobe Length	Length of inner disk corolla tube
Outer Disk Tube Length	Length of outer disk corolla tube
Outer Disk Lobe Length	Length of inner disk corolla lobe
Capitulum curvature	Ratio of total inner: outer disk floret lengths (ovary + corolla tube+ lobe) excluding appendage length
Disk Fruit Length	Length of disk fruits from lowest till highest point
Ray Fruit Length	Length of ray fruits from lowest till highest point
Bract number	Number of capitulum involucral bracts
Disk number	Number of disk florets

Appendix E

Principle Coordinate Analysis (PCoA) of SNP dataset filtered with strict parameters. Axis showing first two principal components and associated percentage variance explained. Colors indicating *a priori* morphological morphs assigned to the complex before sampling: (A) Analysis of All individuals (B) Only S3, S4 and S5 clusters.





APPENDIX F

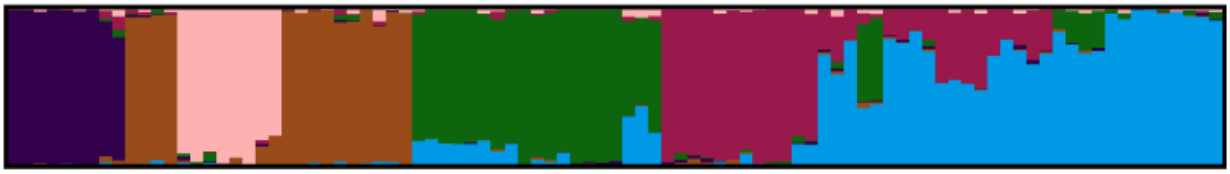
Minor mode of sMNF output for K=5 genetic cluster assignment as well as major and minor modes for K=6 assignment. Numbers to the right of bars indicate number of runs which supported the solution. Population codes of bars shown below.

K=5 MinorCluster1

23/100

6 6 7 7 8 8 9 11 12 13 13 13 13 39 40 40 41 41 41 42 42 43 43 43 44 44 44 45 45 46 46 26 26 26 26 28 29 31 32 33 34 35 36 37 38 38 38 1 1 1 1 2 3 3 4 5 5 5 14 14 15 15 15 16 16 17 17 17 17 18 18 18 19 19 20 20 21 21 22 22 23 23 24 24 24 24 24 25 25 25 25 27 27 30

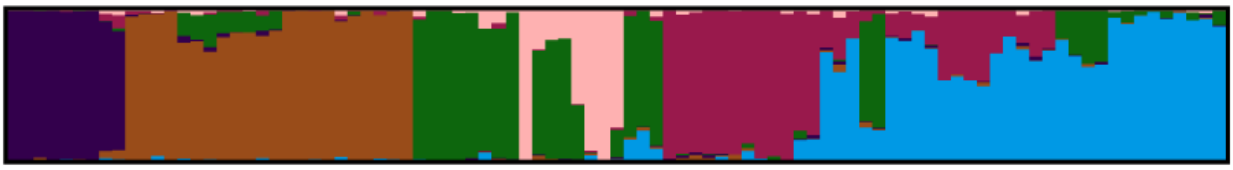
K=6



59/100

6 6 7 7 8 8 9 11 12 13 13 13 13 39 40 40 41 41 41 42 42 43 43 43 44 44 44 45 45 46 46 26 26 26 26 28 29 31 32 33 34 35 36 37 38 38 38 1 1 1 1 2 3 3 4 5 5 5 14 14 15 15 15 16 16 17 17 17 17 18 18 18 19 19 20 20 21 21 22 22 23 23 24 24 24 24 24 25 25 25 25 27 27 30

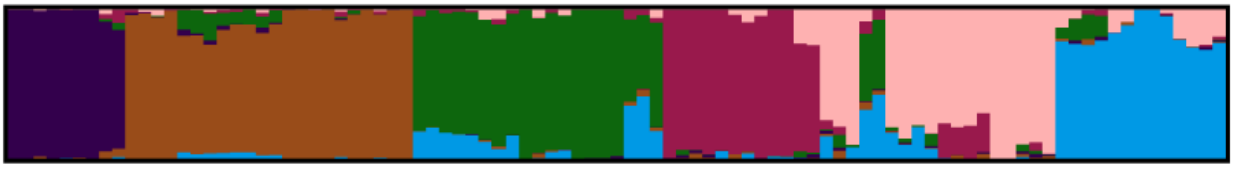
K=6 MinorCluster1



21/100

6 6 7 7 8 8 9 11 12 13 13 13 13 39 40 40 41 41 41 42 42 43 43 43 44 44 44 45 45 46 46 26 26 26 26 28 29 31 32 33 34 35 36 37 38 38 38 1 1 1 1 2 3 3 4 5 5 5 14 14 15 15 15 16 16 17 17 17 17 18 18 18 19 19 20 20 21 21 22 22 23 23 24 24 24 24 24 25 25 25 25 27 27 30

K=6 MinorCluster2



20/100

6 6 7 7 8 8 9 11 12 13 13 13 13 39 40 40 41 41 41 42 42 43 43 43 44 44 44 45 45 46 46 26 26 26 26 28 29 31 32 33 34 35 36 37 38 38 38 1 1 1 1 2 3 3 4 5 5 5 14 14 15 15 15 16 16 17 17 17 17 18 18 18 19 19 20 20 21 21 22 22 23 23 24 24 24 24 24 25 25 25 25 27 27 30

REFERENCES

Andersson, L. 1990. The driving force: Species concepts and ecology. *Taxon.* 39(3):375-382. DOI:10.2307/1223084.

Bolger, A.M., Lohse, M. & Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 30(15):2114-2120.

Chapman, M.A., Hiscock, S.J. & Filatov, D.A. 2013. Genomic divergence during speciation driven by adaptation to altitude. *Molecular Biology And Evolution.* 30(12):2553-2567.

Chesters, D., Wang, Y., Yu, F., Bai, M., Zhang, T.-X., Hu, H.-Y., Zhu, C.-D., Li, C.-D. et al. 2012. The Integrative Taxonomic Approach Reveals Host Specific Species in an Encyrtid Parasitoid Species Complex. *PLOS ONE.* 7(5):e37655. DOI:10.1371/journal.pone.0037655.

Chittka, L. & Menzel, R. 1992. The evolutionary adaptation of flower colours and the insect pollinators' colour vision. *Journal of Comparative Physiology A.* 171(2):171-181. DOI:10.1007/BF00188925.

Cooley, A.M., Carvallo, G. & Willis, J.H. 2008. Is Floral Diversification Associated with Pollinator Divergence? Flower Shape, Flower Colour and Pollinator Preference in Chilean Mimulus. *Annals of botany.* 101(5):641-650. DOI:10.1093/aob/mcn014.

Cowling, R.M., Esler, K.J. & Rundel, P.W. 1999. Namaqualand, South Africa – an overview of a unique winter-rainfall desert ecosystem. *Plant Ecology.* 142(1):3-21. DOI:10.1023/a:1009831308074.

Cramer, M.D., Wootton, L.M., van Mazijk, R. & Anthony Verboom, G. 2019. New regionally modelled soil layers improve prediction of vegetation type relative to that based on global soil models. *Diversity and Distributions.* 0(0). DOI:10.1111/ddi.12973.

Crawford, D.J. & Stuessy, T.F. 2016. Cryptic variation, molecular data, and the challenge of conserving plant diversity in oceanic archipelagos: the critical role of plant systematics. *Korean Journal of Plant Taxonomy.* 46(2):129-148.

Crimmins, T., Crimmins, M. & Bertelsen, C. 2010. Complex responses to climate drivers in onset of spring flowering across a semi-arid elevation gradient. *Journal of Ecology.* 98:1042-1051. DOI:10.1111/j.1365-2745.2010.01696.x.

de Clavijo, E.R. 1995. The Ecological Significance of Fruit Heteromorphism in the Amphicarpic Species Catananche lutea (Asteraceae). *International Journal of Plant Sciences.* 156(6):824-833. DOI:10.1086/297306.

de Jager, M.L., Dreyer, L.L. & Ellis, A.G. 2011. Do pollinators influence the assembly of flower colours within plant communities? *Oecologia.* 166(2):543-553. DOI:10.1007/s00442-010-1879-7.

de Queiroz, K. 2005. Ernst Mayr and the modern concept of species. *Proceedings of the National Academy of Sciences.* 102(suppl 1):6600-6607. DOI:10.1073/pnas.0502030102.

De Queiroz, K. 2007. Species concepts and species delimitation. *Systematic Biology.* 56(6):879-886. DOI:10.1080/10635150701701083.

Degnan, J.H. & Rosenberg, N.A. 2006. Discordance of Species Trees with Their Most Likely Gene Trees. *PLOS Genetics.* 2(5):e68. DOI:10.1371/journal.pgen.0020068.

Doyle, J.J. & Doyle, J.L. 1990. Isolation of plant DNA from fresh tissue. *Focus.* 12(13):39-40.

Edwards, D.L. & Knowles, L.L. 2014. Species detection and individual assignment in species delimitation: can integrative data increase efficacy? *Proceedings of the Royal Society B: Biological Sciences.* 281(1777):20132765. DOI:doi:10.1098/rspb.2013.2765.

Fick, S.E. & Hijmans, R.J. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International journal of climatology.* 37(12):4302-4315.

Fourie, A., Wingfield, M.J., Wingfield, B.D. & Barnes, I. 2014. Molecular markers delimit cryptic species in Ceratocystis sensu stricto. *Mycological Progress.* 14(1):1020. DOI:10.1007/s11557-014-1020-0.
Friar, E.A., Prince, L.M., Roalson, E.H., McGlaughlin, M.E., Cruse-Sanders, J.M., Groot, S.J.D. & Porter, J.M. 2006. Ecological speciation in the East Maui-endemic Dubautia (Asteraceae) species. *evolution.* 60(9):1777-1792. DOI:10.1111/j.0014-3820.2006.tb00522.x.
Frichot, E. & François, O. 2015. LEA: an R package for landscape and ecological association studies. *Methods in Ecology and Evolution.* 6(8):925-929.
Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G. & François, O. 2014. Fast and efficient estimation of individual ancestry coefficients. *Genetics.* 196(4):973-983.
Gremer, J.R. & Venable, D.L. 2014. Bet hedging in desert winter annual plants: optimal germination strategies in a variable environment. *Ecology Letters.* 17(3):380-387. DOI:10.1111/ele.12241.
Gruber, B., Unmack, P.J., Berry, O.F. & Georges, A. 2018. dartR: An r package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Molecular ecology resources.* 18(3):691-699.
Jensen, R.J. 2009. Phenetics: revolution, reform or natural consequence? *Taxon.* 58(1):50-60. DOI:10.1002/tax.581008.
Jombart, T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics.* 24(11):1403-1405.
Kopelman, N.M., Mayzel, J., Jakobsson, M., Rosenberg, N.A. & Mayrose, I. 2015. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Molecular ecology resources.* 15(5):1179-1191.
Leaché, A.D. & Oaks, J.R. 2017. The Utility of Single Nucleotide Polymorphism (SNP) Data in Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics.* 48(1):69-84. DOI:10.1146/annurev-ecolsys-110316-022645.
Leaché, A.D., Fujita, M.K., Minin, V.N. & Bouckaert, R.R. 2014. Species Delimitation using Genome-Wide SNP Data. *Systematic Biology.* 63(4):534-542. DOI:10.1093/sysbio/syu018.
Lemopoulos, A., Prokkola, J.M., Uusi-Heikkilä, S., Vasemägi, A., Huusko, A., Hyvärinen, P., Koljonen, M.-L., Koskiniemi, J. et al. 2019. Comparing RADseq and microsatellites for estimating genetic diversity and relatedness — Implications for brown trout conservation. *Ecology and Evolution.* 9(4):2106-2120. DOI:10.1002/ece3.4905.
Leviyang, S. & Hamilton, M.B. 2011. Properties of Weir and Cockerham's Fst estimators and associated bootstrap confidence intervals. *Theoretical Population Biology.* 79(1):39-52. DOI:https://doi.org/10.1016/j.tpb.2010.11.001.
Li, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics.* 27(21):2987-2993.
Li, H. & Durbin, R. 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics.* 26(5):589-595.
Nordenstam, B. 2007. Tribe Calenduleae. In *Flowering Plants. Eudicots: Asterales.* J.W. Kadereit and C. Jeffrey, Eds. Leipzig, Germany: Springer. 241-245.
Norlindh, T. 1943. Monograph of the genera *Dimorphotheca*, *Castalis, 0steospermum, Gibbaria* and *Chrysanthemoides*. In *Studies in the Calenduleae I.* Lund, Sweden.
Pembleton, L.W., Cogan, N.O. & Forster, J.W. 2013. St AMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. *Molecular ecology resources.* 13(5):946-952.
Perea, C., De La Hoz, J.F., Cruz, D.F., Lobaton, J.D., Izquierdo, P., Quintero, J.C., Raatz, B. & Duitama, J. 2016. Bioinformatic analysis of genotype by sequencing (GBS) data with NGSEP. *BMC genomics.* 17(5):498.

Pritchard, J.K., Stephens, M. & Donnelly, P. 2000. Inference of Population Structure Using Multilocus Genotype Data. *Genetics.* 155(2):945-959. Available: https://www.genetics.org/content/genetics/155/2/945.full.pdf.
Puritz, J.B., Hollenbeck, C.M. & Gold, J.R. 2014. dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ.* 2:e431.
R Development Core Team. 2018. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing
Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A., Firth, D. & Ripley, M.B. 2013. Package 'mass'. *Cran R.*
Sneath, P.H.A. 1976. Phenetic Taxonomy at the Species Level and above. *Taxon.* 25(4):437-450. DOI:10.2307/1220526.
Solís-Lemus, C., Knowles, L.L. & Ané, C. 2015. Bayesian species delimitation combining multiple genes and traits in a unified framework. *evolution.* 69(2):492-507. DOI:10.1111/evo.12582.
Spinks, P.Q., Thomson, R.C. & Shaffer, H.B. 2014. The advantages of going large: genome-wide SNPs clarify the complex population history and systematics of the threatened western pond turtle. *Molecular Ecology.* 23(9):2228-2241. DOI:10.1111/mec.12736.
Torrell, M. & Vallès, J. 2001. Genome size in 21 Artemisia L. species (Asteraceae, Anthemideae): Systematic, evolutionary, and ecological implications. *Genome.* 44(2):231-238. DOI:10.1139/g01-004.
Van Zyl, J.J. 2001. The Shuttle Radar Topography Mission (SRTM): a breakthrough in remote sensing of topography. *Acta Astronautica.* 48(5-12):559-565.
Weir, B.S. & Cockerham, C.C. 1984. Estimating F-statistics for the analysis of population structure. *evolution.* 38(6):1358-1370.
Wiley, E.O. 1978. The Evolutionary Species Concept Reconsidered. *Systematic Biology.* 27(1):17-26. DOI:10.2307/2412809.
Willing, E.-M., Dreyer, C. & van Oosterhout, C. 2012. Estimates of Genetic Differentiation Measured by FST Do Not Necessarily Require Large Sample Sizes When Using Many SNP Markers. *PLOS ONE.* 7(8):e42649. DOI:10.1371/journal.pone.0042649.
Yang, Z. & Rannala, B. 2010. Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences.* 107(20):9264-9269.
Zachos, F.E. 2015. Taxonomic inflation, the Phylogenetic Species Concept and lineages in the Tree of Life – a cautionary comment on species splitting. *Journal of Zoological Systematics and Evolutionary Research.* 53(2):180-184. DOI:10.1111/jzs.12088.
Zhang, C., Zhang, D.-X., Zhu, T. & Yang, Z. 2011. Evaluation of a Bayesian Coalescent Method of Species Delimitation. *Systematic Biology.* 60(6):747-761. DOI:10.1093/sysbio/syr071.

Name: Thaabiet Parker

Student Number: PRKTHA005

Course: BIO4000W

Declaration

I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.

I have used the ……Harvard UCT…… convention for citation and referencing. Each contribution to, and quotation in, this essay/report/project/……project…… from the work(s) of other people has been attributed, and has been cited and referenced.

This essay/report/project/……project…… is my own work.

I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.

Signature _____T.P_____

Date _____18/10/2019_____