ddRAD-seq based identification of genetic markers to facilitate the assessment of clonal versus sexual reproduction in *Dichrostachys cinerea*

Liam Lester Lumley

THESIS PRESENTED FOR THE DEGREE OF MASTERS OF SCIENCE IN THE FIELD OF MOLECULAR AND CELL BIOLOGY

SUPERVISOR: DR ROBERT INGLE

CO-SUPERVISOR: DR MICHAEL LENHARD

2025

Contents

Plagiarism Declaration

I know the meaning of plagiarism and declare that all of the work in the dissertation (or thesis), save for that which is properly acknowledged, is my own.

X Signed by candidate

Liam Lumley

Liam Lumley

Signed 31.01.2025

Acknowledgements

It is difficult to truly thank everyone who has been involved in this journey up until this point and there are many who are not on this list but are no less deserving. Many of the people who are mentioned here deserve far more from me than the few words written here.

First and foremost, I would like to thank my supervisor, Dr Robert Ingle, as he has been a guiding light throughout this entire project. As a result of my interactions with him, I have grown far beyond what I thought I was capable of when I first walked into his lab. Thank you for always pushing me to do better, encouraging me and keeping me focused Rob. You have had a stronger influence than you realise on the scientist I wish to become.

I would then like to thank Dr Michael Lenhard, Dr Mathias Scharmann, and Dr Christian Kappel for their invaluable guidance and impromptu teachings in evolutionary and bioinformatic concepts, and for making me feel welcomed during my time in their lab. I have still not seen someone write code as fast as Dr Kappel.

I would also like to thank Dr Laurence Kruger, for his invaluable guidance and assistance during the sampling phase of the project in KNP and his aid with the selfing experiments.

I would like to thank all the members of the Illing Lab, especially Dr Nicola Illing and Eugene Kabwe Ntuntu Kabamba for their advice on aspects of the project. Additionally, I would like to thank Dr Lara Donaldson and Tim Reid for allowing me to use the facilities necessary for the flow cytometry experiments.

I would also like to thank Manyoni Game Reserve for allowing collection of samples and Olwen Jakumeit for arranging access to the site. Additionally, I would like to thank the owners of the Tarentaalnes Guesthouse for allowing us to collect leaflets on their property. Finally, thank you to SANParks for allowing us to collect leaflets within the Kruger National Park.

On a more personal note, I would like to thank my friend Sabine Kirchner, together we have been through thick and thin, helping each other through the lowest and highest points of our academic and personal endeavours. To many more, my friend. Thank you to Caroline Roberston for the welcomed advice and the pleasant conversations. Thank you to Luc Pegram for all the conversations and lab antics, hopefully we shall publish together one day.

Thank you to the university of Cape Town (UCT) for financial support in to form of Merit and Need financial aid. Without the contribution of this aid, I would not have been able to support myself nor would I have been able to complete my studies.

Finally, thank you, Cindy and Fanie, for all that you have done for me. You really are where all of this started and where it will all go.

Abstract

Dichrostachys cinerea, commonly referred to as “sicklebush”, is a semideciduous nitrogen-fixing tree species belonging to the Fabaceae family and is native to parts of Africa, India and Northern Australia. In addition to propagation via seed dispersal, the species is capable of asexual reproduction via the process of root suckering. *D. cinerea* can aggressively colonize grasslands and quickly becomes problematic given its ability to persist through treatments such as fire, chemical treatment and mechanical removal. Due to its shrub-like physiology; propensity for root suckering; and the production of large amounts of seed, removal efforts often fail, especially in areas where *D. cinerea* has established dense thickets. African savannahs in particular have experienced a rapid rise in woody plant encroachment, with *D. cinerea* often one of the major species held responsible, and it is a species of concern to local ecologists. While previous studies have investigated the various factors which contribute to woody plant encroachment within Savanna ecosystems, it is still unclear as to what role asexual reproduction plays in this phenomenon and current methods to study this are laborious and time consuming. This study sought to develop a genetic based workflow to allow the detection of clonal individuals in *D. cinerea*. To facilitate this study, 225 *D. cinerea* samples were collected across the species range of *D. cinerea* in South Africa. DNA extracted from these samples allowed for the identification 39 299 SNP markers generated from ddRAD-seq libraries. These SNPs were then used to elucidate the phylogenetic relationships between individuals in order to ascertain whether clonality could be detected within the species, via the use of a genetic-distance-based threshold value which would discriminate between clonal and non-clonal individuals. Findings from this study revealed that across its species range, *D. cinerea* consisted of a mixed cytotype population, exhibiting a clear phylogenetic divide between diploid and tetraploid individuals. Identity-by-state analysis of the Kruger National Park diploid population confirmed that naturally occurring clonal individuals could be detected via ddRAD-seq generated SNP markers. This workflow will facilitate the study of clonality in *D. cinerea* in future studies which seek to answer ecologically relevant questions.

Introduction

Dichrostachys cinerea

Dichrostachys cinerea, commonly referred to as “sicklebush”, is a semideciduous nitrogen-fixing tree species belonging to the Fabaceae family and can grow up to 10 meters in height upon reaching maturity [1, 2]. The sicklebush produces fragrant inflorescences (from September to February in Southern Africa) and which are bicoloured pendulous spikes that are pink on the sterile upper half and yellow on the lower half of the inflorescence (Figure 1A). The leaves are compound and pinnate and can vary considerably in size (Figure 1B) [3]. Branches harbour strong alternating thorns that can grow up to 8 cm in length while the tree bark is usually rough on older trees, with bark colour varying between yellow to grey/brown (Figure 1C). The resulting seed pods are usually twisted or spiralled in morphology and are considered nutritionally valuable to animals due to their high nitrogen content (Figure 1D). In addition to propagation via seed dispersal, the species is capable of asexual reproduction via the process of root suckering, lending to its persistent nature in the face of environmental disturbances such as fire [4]. The species is considered to be both fire and drought resistant, but is notably sensitive to frost and waterlogging [5]. *D. cinerea* is native to parts of Africa, India and Northern Australia but has also been introduced to other areas where it has become highly invasive, such as Cuba. In fact, recent studies revealed that in Cuban province of Ciego de Avila, *D. cinerea* coverage increased from 61 977 ha to 91 533 ha from the years 1994 to 2022, despite government-backed efforts to minimize its spread [6].

Figure 1. Morphological traits of D. cinerea. In Southern Africa, flowering generally occurs between September to February (A). Compound leaves have been observed to display a large variety in leaflet size (B). Tree bark of this species is usually rough in mature trees and is commonly used as firewood (C). Seed pods are consumed by natural wildlife, however, may provide enhanced nutritional value when supplemented in livestock feed (D). Image sources: A and D from [7], B and C from [5]

Literature relating to *D. cinerea* has revealed several interesting characteristics of the tree and it's varied uses. Previous work on *D. cinerea* suggested that the roots of this species may harbour a natural source of antimicrobial compounds. Tannins extracted from the root tissue exhibited inhibitory properties against common yet problematic bacteria, such as *Staphylococcus aureus*, which are known to be the causative agent in a variety of human infections [8]. Further investigation into the medical applications of this species revealed that existing phenolic compounds extracted from the plant could harbour antioxidant as well as anti-inflammatory activity for use in humans [9]. Additionally, due to its abundance and easily harvestable fruit, several studies have sought to assess the effects of supplementing animal feed with seed pods obtained from locally occurring trees, in order to reduce the feeding expenses faced by small-scale local farmers as well as improving livestock performance during dry seasons [10, 11]. One study even suggested that the supplementation of *D. cinerea* seed pods into existing molasses-urea block recipes could serve as an additional and inexpensive source of nitrogen and fibre [12]. Research has also been conducted in order to assess the potential of the species as a source of solid biofuel, where overall findings suggested that *D. cinerea* displays similar emissions to that of other woody biomass biofuels available [13]. The

bark of *D. cinerea* has also been studied to assess its viability as a source of natural fibre to substitute current synthetic fibres, however, further research is still required prior to implementation [14].

During their taxonomic assessment of *D. cinerea* across its range, Brenan & Brummitt suggested that four subspecies were present in South Africa, namely subsp. *nyassana*, subsp. *africana*, subsp. *argillicola*, subsp. *forbesii* [15], with a total of eight taxonomic variants based on several morphological traits such as leaflet width, leaf size, peduncle arrangement and fruit-pod width. After the description of these subspecies, Ross (1974) performed an independent analysis of the variation of these morphological traits across the species' range in South Africa. This analysis did not support the existence of eight *D. cinerea* taxa in South Africa; instead the morphological traits used to delimit them were found to be continuous, suggesting that the sub-species described by Brenan & Brummitt merely represented taxonomic assignment to the extremes of continuous variation [3]. However, Ross proposed to maintain the distinction between subsp. *nyassana* and subsp. *africana* on the basis of whether the leaflet width exceeded 2 mm or not, with the larger leaflet sizes being assigned to subsp. *nyassana* [3], despite the continuous variation in this trait.

D. cinerea has proved highly invasive in areas where it has been introduced, notably in Cuba. Due to its shrub-like physiology, propensity for root suckering [4], the production of large amounts of seed (Figure 1D) and the presence of large thorns on branches, removal efforts quickly become strenuous, especially in areas where *D. cinerea* has established dense thickets (Figure 2A). *D. cinerea* tends to aggressively colonize land and quickly becomes problematic given its ability to persist through treatments such as fire, chemical treatment and complete mechanical removal. In Cuba, estimates in 1999 suggested *D. cinerea* grew on over 20% of the arable land (12% of the national territory) [13]. An assessment of the efficacy of these three treatments on the species revealed that none of the treatments could fully eradicate established individuals and would require constant and adaptable treatment, however, treatment via fire did allow for a balance between shrub removal and minimizing its impact on the surrounding flora [16]. Similarly, frequent fire treatment coupled with light grazing harbours the potential to halt and even reverse shrub encroachment with fire frequency seemingly playing a larger role, however, neither component on their own are as effective [17].

Figure 2. D. cinerea can establish impenetrable thickets requiring large resource allocation for removal (A). Image source: [7]. Savannahs are often characterized by large areas of grassland vegetation interspersed by woody vegetation (B). Image source: [18].

D. cinerea can also be invasive within its natural range. African savannahs in particular have experienced a rapid rise in woody plant encroachment, with *D. cinerea* often one of the major species held responsible, resulting in it quickly becoming a species of concern to local ecologists [2, 4, 16, 17]. Savanna ecosystems are characterised by a distinct vegetation structure; a continuous grass layer with scattered trees and shrubs (Figure 2B, [19]). The open canopy allows sunlight to reach the ground, facilitating the growth of a diverse array of grass species which supports a range of herbivores, and their predators. Savanna habitats around the world are threatened by bush encroachment, which refers to the gradual invasion of woody vegetation into grass-dominated areas, leading to a reduction in open spaces and altering the ecological dynamics of the savanna [20]. Investigation into the effects of *D. cinerea* savannah encroachment on existing herbaceous vegetation in East-Africa revealed that while sparsely encroached areas may provide a short-term increase in species richness and diversity, further increase in encroachment density resulted in a negative shift in species composition and richness [2]. Bush encroachment has been associated with various factors, including climate change, land-use practices and the suppression of natural fire regimes [21].

D. cinerea is just one of the wood plant species that contribute to woody plant encroachment in the savannas. Its adaptive characteristics, such as the ability to fix nitrogen and to resprout after top-killing by fire make it a key player in the transformation of African savanna ecosystems [22]. However, other species such as *Tarchonanthus camphoratus*, *Acacia tortilis* (*Vachellia tortilis*), and *Acacia mellifera* (*Senegalia mellifera*) have also been observed to contribute to the encroachment of woody plant species in the Northern Cape in South Africa [23]. Overall, six legume species have been broadly considered to be major contributors to woody plant encroachment throughout South Africa, namely, *Vachellia hebeclada*, *V. karroo*, *V. nilotica*, *D. cinerea* and the aforementioned *A.tortilis* and *A. mellifera*, however these species have different contributions in different parts of South Africa, with *D. cinerea* being considered problematic across the entire range of South Africa [24].

D. cinerea has proved very difficult to control [21]. Studies show that methods such as high intensity burning in the Kruger NP have no discernible effect after ten years [25]. Additionally, three strategies namely, cutting coupled with the application of glyphosate herbicide, cutting coupled with controlled burning of the remaining stem, and complete manual excavation of shoots and roots were previously assessed by Utaile et al. (2023) to assess the efficacy of the combination of these traits. The study revealed that neither of the treatment methods appeared to be significantly more effective than the other, however authors suggested that fire treatment may prove to be the most promising treatment method moving forward [16].

Factors that contribute to the causes of *D. cinerea* encroachment have been assessed previously. The study found a negative correlation between initial shrub coverage and the change in shrub coverage, with areas with low initial shrub coverage being susceptible to recruitment, while areas of high initial shrub coverage being susceptible to increased mortality. The authors concluded from this that the growth rate of a shrub population was dependent on the density of shrub. The study also found that an increase in fire frequency appeared to be associated with a decline in shrub coverage, reporting a decline in shrub coverage was observed in areas where a fire frequency was greater than one every three years. Interestingly the study revealed that fire frequency and grazing pressure appeared to be co-linear, and that this fire-grazing correlation was an important factor in controlling the spread of shrub encroachment, with authors suggesting light grazing coupled with frequent fires may prove an effective control strategy in minimizing woody plant encroachment [17].

While it is unclear what contribution root suckering provides in areas of *D. cinerea* encroachment, one previous research project has investigated the effects of habitat disturbance on the frequency of clonality. A previous small-scale study of *D. cinerea* sought to elucidate the importance of clonal spread in areas of varying degrees of fire disturbance and herbivory in the Hluhluwe-iMfolozi game reserve in Kwa-Zulu Natal [4]. The researchers excavated the soil around ±30 juvenile plants at 11 sites (370 plants in total) in the reserve and looked for evidence of either a taproot (indicating that the plant was derived from a seedling – sexual reproduction) or connection to a horizontal root (root sucker – asexual reproduction). There appeared to be no correlation between fire disturbance or herbivory and the extent of clonal reproduction across the 11 sites. Instead the authors reported that root suckering was consistent across sites, with a mean proportion of 0.55 root suckers across the sites [4]. Besides being a laborious way to assess the extent of clonal reproduction (it took 6 people 21 days to excavate the soil around the 370 plants) it was not possible to assign up to 30% of the juvenile plants as being derived from seed or root sucker at some sites due to damage to the roots or abnormal physiology [4]. An additional complication is that the connecting root between the parent plant and root sucker breaks down over time, which can lead to incorrect assignation as non-clonal individuals. An alternative approach to this method lies in the development of a genetic marker-based assay to allow determination of genetic relatedness between individuals in a population, which is the aim of this thesis.

Next generation sequencing

Of all the sequencing technologies available, Illumina has become the gold standard for the high throughput sequencing of short reads [26], and has become a staple in the design

process of sequencing projects which include research in the fields of metagenomics, transcriptomics and whole genome DNA sequencing (WGS). The cost of Illumina sequencing has also decreased in recent years, allowing for the inclusion of sequencing in experimental design in smaller labs, while also allowing for the generation of large amounts of both single-end and paired-end read data [27]. With WGS data researchers are often able to use a combination of both second generation sequencing technologies and classical sanger sequencing to assemble plant genomes, however, due to the limitations bought forth by the shorter read lengths and complex plant genomes, WGS often captures a large amount of repetitive genome elements which then limits the quality of genome assembly [28]. If a reference genome for an organism is unavailable, the Genomescope software provides a reference free k-mer based analysis of the raw sequence data in order to provide basic genome characteristics such as estimated genome size, heterozygosity and k-mer coverage [29] such as in *Chrysanthemum makinoi*, in which the genome size and heterozygosity was assessed based on k-mers (k = 31) extracted from paired-end HiSeq Illumina read data, allowing author suggest a genome size estimate of 3.19 Gb [30]. Another use for sequencing data is for the identification of genetic/molecular markers.

A number of genetic markers have been established over the last decades [31]. Restriction fragment length polymorphisms (RFLP) were some of the first genetic markers to be developed for use in genetic studies. This method relies on the high specificity of restriction enzymes (RE) for their recognition site present on the target DNA. Digestion of a target DNA fragment with a RE would allow for the observation of a robust fragment length profile on an electrophoretic gel, and any polymorphisms present in the RE recognition site would result in an altered DNA fragment length profile. This technique could be used to investigate for the presence of DNA polymorphisms that could be utilized to distinguish between individuals and species [32]. For example, the analysis of mosquito bloodmeals using PCR-RFLP to identify polymorphisms in the cytochrome B sequence across vertebrate species allowed for the identification of the origin of the vertebrate-derived blood-meals [33]. Microsatellites, also referred to as short tandem repeats (SSRs) are tandem repeated motifs, which can reach a length of between one and six nucleotide bases and are distributed both prokaryotic and eukaryotic genomes [34]. SSRs are first identified *in silico* using software designed to assess the target genome for the presence and position of SSRs across the genome [35]. This information can then be used for various downstream functions such as the construction of genetic linkage maps, quantitative trait locus mapping and for the assessment of population structure [36].

More relevant to the current study are methods used in the identification and utilization of large amounts of single nucleotide polymorphisms (SNPs) as genetic markers for population studies. SNP markers are the result of a single base change that has occurred within the DNA sequence at a given position within the genome [37]. The use of SNP markers for population studies has shown a sharp rise in recent years, in part, due to the advancement of high throughput sequencing technologies within the last two decades, which has led to a large reduction in the cost of sequencing whole genomes for an organism of interest [26]. However, the recent success of SNPs as a reliable marker choice for population studies can also be attributed to the emergence of high-throughput marker discovery techniques. High

throughput marker discovery was initially pioneered by three original techniques, namely reduced representation libraires (RRL), genotyping by sequencing (GBS) and restriction-site-associated DNA sequencing (RAD-seq) [38]. While each of the techniques were different, they all shared the initial step of digesting the whole genome with a restriction enzyme and the final step of sequencing a highly fragmented representation of the original genome. This allowed for the identification of hundreds to thousands of shared SNPs across several samples in a population of interest from only a sub-set of the original genome.

In particular, the RAD-seq technique comprised of the digestion of high molecular weight (HMW) DNA with a single restriction enzyme, which was followed by the ligation of adapters, random shearing and size selection of these fragments which were then sent for sequencing (Figure 3) [39]. Since its induction, marker identification by RAD-seq has become a popular technique in plant studies, for example, to assess the spatial genetic structure of the asexually reproducing species *Cardamine leucantha*, RAD-seq was used to identify SNP markers in order to determine which individuals were ramets of the same genet across the sampling space. The study found a high degree of clonal diversity, with 61 genets being identified across the 20 x 20 m study plot. Findings suggested that both seed dispersal and clonal ramet production played a role in determining the genetic structure of the population in question [40]. Given its efficiency in marker identification, further advancements to the RAD-seq design then led to the development of several other RAD-seq based techniques, such as 2b-RAD [41], ezRAD [42] and double-digest RAD-seq (ddRAD-seq) [43].

The ddRAD-seq protocol exhibits two main differences from the original RAD-seq protocol. Firstly, instead of using a single restriction enzyme to digest the HMW DNA, an additional enzyme is incorporated into the initial digestion step, thereby excluding the need for random shearing of DNA fragments. Secondly, ddRAD-seq protocol sees the inclusion of a precise size selection step, which allows the user greater control over the fragment length present in the final library, allowing read counts across regions to be similar between individuals [43]. The bioinformatic analysis of ddRAD-seq data usually consists of de-multiplexing and trimming of barcodes from the resulting reads, a task that is effectively performed by the STACK process_radtags program [44]. If a reference genome is available, reads can be aligned to the reference genome where SNP calling can then be performed, however, if no reference genome is available then a pseudo reference genome can be assembled *de novo* [45]. SNPs are represented in the standard VCF format, a file format that can be manipulated via programs such as VCFtools [46] and BCFtools [47] to perform quality control on the resulting SNP set. Several tools exist to process data in studies that utilise RAD-seq technology, such as STACKS [44], UNEAK [48] and dDocent [49]. dDocent in particular is a user-friendly pipeline that seeks to automate the SNP calling process, making it very accessible to researchers who are new to the field of bioinformatics.

ddRAD-seq is commonly used to assess the relatedness between individuals within a population, including economically important crop species. For example, a study on the Campanian and Apulian grapevine varieties within southern Italy sought to elucidate the population structure and genetic diversity of a collection of clones, propagated from mother tissue, originating from these varieties. SNP marker datasets derived from both GBS and

ddRAD-seq techniques were used to identify seven separate clonal sub-populations, each separating strongly on the basis of geographical population, suggesting a strong history of clonal propagation within these areas. Interestingly, it was discovered that one of the sub-populations separated into two groups, showcasing the sensitivity of ddRAD-seq in the assessment of relatedness, even when the technique was used on highly similar individuals [50].

As ddRAD-seq harbours a strong potential to be used in the assessment of clonality between individuals, several studies have made use of its ability to distinguish between highly genetically similar individuals in wild populations. For example, a study by Amor, Johnson & James (2020) sought to inform the conservation of the endangered species *Bossiaea vombata*, a species considered to be functionally sterile due to its lack of viable seeds and high frequency of asexual reproduction. The authors sought to determine whether ddRAD-seq technology was suitable for detecting clonality in this non-model species. By assessing the pairwise genetic distance among individuals, a threshold value of genetic distance was established in order to infer clonality, which acted to identify possible clones within the population by suggesting the minimum amount of genetic similarity that would be required in order to be considered clonal. This threshold value was based on the expected diversity levels that reflect sequencing error and somatic mutations between clones. This threshold value was then used to identify the presence of five clonal organisms across four sampling sites, with three of the sites being monoclonal and the fourth site consisting of two clonal organisms [51].

Fairy circles (FC) are circular patches of grassland vegetation in the Namib desert characterized by the absence of vegetation within the circle centre. While several hypotheses have been proposed to explain how FC form, none have yet been confirmed. In order to assess whether these patches arose due to clonal propagation, Kappel et al. (2020) used ddRAD-seq derived SNP markers from *Stipagrostis ciliata* and *S. uniplumis* to assess the extent of clonality present in FC observed in the Namibian desert. They reported that for both species the samples from within one circle generally did not cluster together tightly and were instead interspersed with samples from other circles, indicating that the FC are not derived from clonal propagation of a single genet [52]. Interestingly, they found that both species exhibited intraspecific ploidy variation as several individuals from each species displayed over 1000 loci with more than two distinct haplotypes, indicating that they are likely tetraploids.

A study of the seagrass *Posidonia australis* used SNPs derived via ddRAD-seq to assess population generic diversity and structure across ten populations within Shark Bay, Australia. The study found that nine of the ten populations exhibited a high degree of observed heterozygosity and that the same nine high heterozygosity populations clustered separately from the low heterozygosity population. Authors used technical replicate samples to determine the rate of SNP calling error, which was used as a cut-off value for the degree of genetic similarity individuals would need to exhibit in order to be considered clonal. Interestingly, karyotyping experiments on fresh *P. australis* tissue revealed that all nine high heterozygosity populations were tetraploid, while the diploid population exhibited a lower

degree of heterozygosity in comparison, revealing an instance of intraspecific ploidy variation [53].

There are several different types of ploidies that are observed in living organisms. Haploid organisms, such as some algae, have only a single set of chromosomes and exhibit haploid dominant life cycles. Most other organisms, including most plant species, exhibit diploid dominant life cycles i.e. with two sets of chromosomes. However, it is well known that plants alternate between the haploid and diploid phase during their reproductive cycle [54]. It is also well know that plant genomes commonly undergo polyploidization events [55]. Polyploid organisms have multiple sets of chromosomes and can arise in a species through a variety of pathways, including whole genome duplication [55]. Polyploidy can be further characterized into either autopolyploidy or allopolyploidy. Briefly, autopolyploidy is polyploidy developed from the inheritance of an additional set of chromosomes from a member of the same taxonomic species, while allopolyploidy is ploidy developed from the transmission of chromosome sets from another species i.e. hybridization [56]. Some species also exhibit intraspecific ploidy variation, where one species could consist of a mixture of different ploidies within the population, a phenomenon which has been observed in previous studies [53, 57].

The effects of intraspecific ploidy differences in a species have been shown to contribute to morphological differences between individuals of different cytotypes, such as an increase in leaf size in individuals with a higher ploidy [58-60]. For example, a study on *Buddleja macrostachya* sought to assess whether morphological characteristics could be used to assay individual ploidy level. Investigating the measurements of various morphological characteristics of individuals of known ploidy revealed a significant difference in the leaf size, flower length and fruit length between hexaploid and dodecaploid plants [59].

The ploidy of an organism can be inferred via methods such as karyotyping to assess chromosome number [53], or flow cytometry by comparing the relative genome sizes of members of the same species [61]. The ploidy of an organism can also be inferred *in silico* using NGS sequencing data and the relevant bioinformatic tools such as PloidyNGS, a model free visualization tool which automates the assessment of ploidy using short read data [62]. SNP data can also be used to estimate the ploidy of an organism, based on the distribution of read counts at biallelic SNPs [63]. In theory, within a diploid genome, the mean read counts (allele depth) at heterozygous SNP positions have a single mode at 0.5 where half of the reads should account for each allele, while in a tetraploid individual the mean read counts can display modes at 0.25, 0.5 and 0.75 (Figure 3).

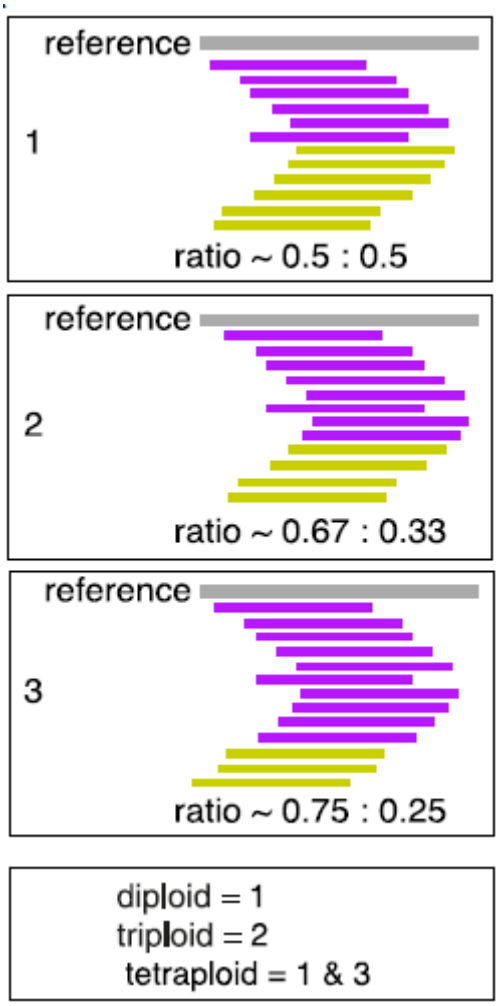

Figure 3. Expected read count distributions in organisms of variable ploidy. In diploid individuals a read count distribution of 50% of reads are expected at a heterozygous site. Source: [63].

Flow cytometry

Flow cytometry (FCM) is a laser-based technology that allows for the rapid measurement of single cell properties via a combination of several components that collectively aim to organize cells into a single file, detect cells as particles via laser-based interrogation, and generate several light-scatter based data for downstream analysis [64]. The process of FCM analysis of a cell population of interest can be broken down to three fundamental components, namely fluidics, optics, and electronics. Fluidics in flow cytometry refers to the use of an appropriate buffer for the cells of interest which is used to both sort cells into single file and guide cells throughout the machine at a constant rate [65]. Optics in FCM refer to the use of a highly sensitive light source with a wide range of excitation wavelengths to assess the light scatter or fluorescence resulting from a cell to determine the cell properties. [66]. As cells are guided through to the interrogation point, they encounter the laser which results in the emission of light scatter which are then filtered via specialized filters and detected by photodetectors (Figure 5). Within modern flow cytometers, photodetectors often take on the form of photomultiplier tubes (PMT) which are a complex arrangement of detectors to detect photos

on interest [65]. Light scatter can be categorized into two components: forward scatter which is generally indicative of the relative size of a cell and side scatter which can be used to infer the complexity/granularity of a cell [67]. Additionally, if cells have been fluorescently labelled, the resulting emission wavelength can be detected and quantified [68]. Finally, the resulting photons are converted into an electrical signal which can be stored as data and analysed.

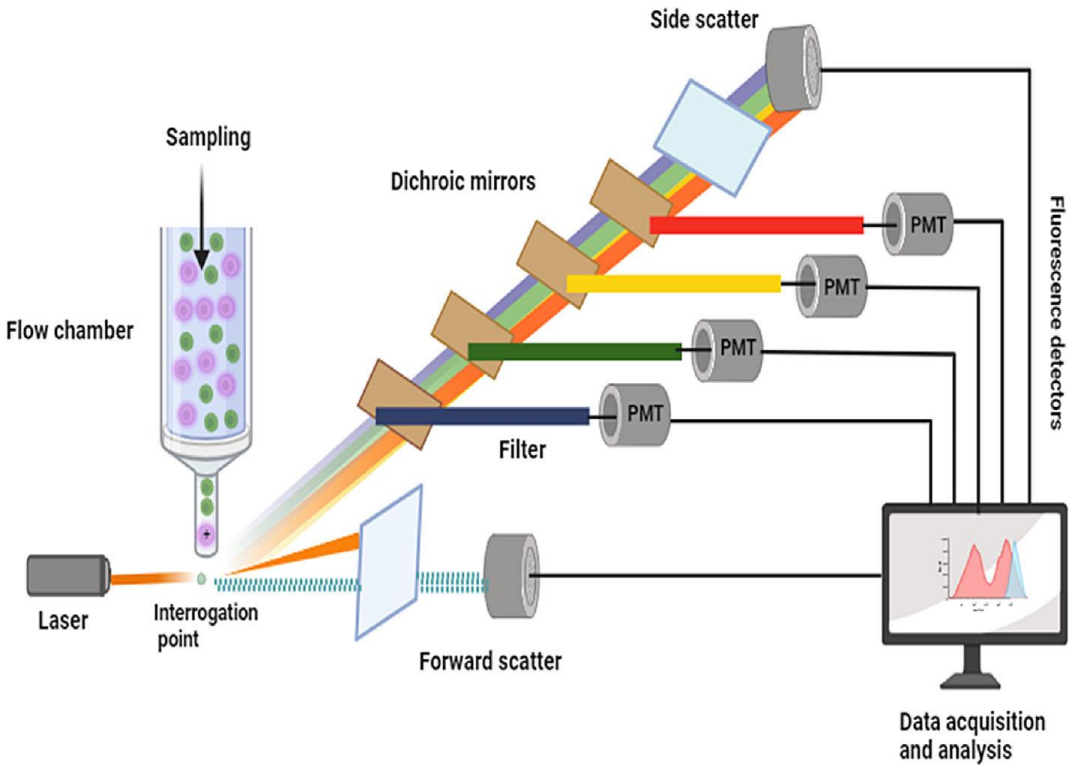

Figure 4. The process of flow cytometry begins with the loading of samples, which are funnelled into a single file of cells which can be individually interrogated via a laser. The emission of non-specific light and fluorescence is then directed and detected by highly sensitive equipment, which allow for the conversion of photons into electrical signals. Data analysis is then performed via specialized software. Image source: Jyoti, Chandel & Singh (2023) [67].

FCM has been creatively adopted to allow for several applications within the field of plant science. FCM has risen in popularity in the plant science community as a cost-effective alternative, and potentially high throughput manner, in which to assess the characteristics of a plant genome and investigate plant genetics. In the case of gene expression studies, it has been shown to be a viable alternative to traditional enzyme assays, with Hagenbeek and Rock (2001) reporting that the use of FCM to assess gene expression produced similar results to traditional enzyme assays in *O. sativa L.*, however, FCM proved to be less sensitive when detecting weak promoters [69]. FCM has also been used in plants to aid in the enrichment of tissue specific cells to allow for downstream RNA analysis [70].

FCM has also been used to karyotype several plant species in a practice coined 'flow karyotyping', whereby an estimate of the chromosome count can be provided in a fraction of

the time normally required for traditional karyotyping [71]. This process involves the induction of cell cycle synchronization in order to arrest cells in the metaphase state whereafter chromosomes are then released from the cells and stained for identification in downstream FCM analysis [71, 72]. While this practise may not be as reliable as traditional microscopic based karyotyping [72], the method has been successfully implemented in commercially significant plants, such as *Secale sereale L.* [73].

While several other applications for FCM in plant studies have arisen, arguably the most popular implementation is its use in the estimation of ploidy and relative genome size/DNA content. The history of DNA content estimation in plants began several decades ago when the first successful attempt was reported [74]. The estimation of nuclear DNA content is an amalgamation of several concepts, however at its core, it is the estimation of the amount of DNA present in an unknown sample of interest relative to a reference of known DNA content. Generally, nuclei are first extracted from plant cells via mechanical disruption and suspended in a specialized buffer that helps to both facilitate the release of nuclei from the damaged cells and keep the nuclei intact until analysis via FCM [75].

The use of FCM in genome size and ploidy estimations has been employed for many plant species. In *Ranunculus parnassifolius*, an angiosperm distributed throughout the central and southwestern parts of Europe, 112 individuals were assessed for 2C DNA size via FCM in order to elucidate the cytotype distribution of this species in the northwestern part of Spain. The use of the high throughput FCM method to assess genome size allowed researchers to not only determine the 2C DNA value of all the samples, ranging between 7.08 pg to 21.04 pg, but to also discover three levels of ploidy within the species, reporting a mixed ploidy population consisting of a mixture of diploids, tetraploids and a small portion of pentaploids in one of the eight populations sampled [61]. Furthermore, the study also reported that the quality of PI intensity peaks degraded depending on the tissue type used, with lower CVs, sharper peaks and less debris resulting from fresher tissue in comparison to frozen and preserved tissue types though the reduced quality did not seem to affect the estimation of DNA ploidy levels [61]. This supported previous work from Suda & Travnicek (2006), which assessed the performance of FCM in DNA ploidy estimation across 60 air-dried species, which reported that genome size and DNA ploidy estimated could be successfully obtained from tissue preserved for up to 3 years [76].

Although the use of FCM to investigate genome size and species DNA ploidy is a robust technique that allows for a high throughput manner in which to assess these traits, there are some limitations that can arise with the technique [77]. Firstly, plants are known to produce a wide range of secondary metabolites which may impair the binding of fluorescent dyes to the DNA of the cell, or are themselves auto-fluorescent, which can become particularly problematic in non-model species granted the lack of knowledge of their biochemistry [78]. Secondly, many cellular structures may undergo non-specific binding of the fluorescent dyes used which may lead to difficulties in data interpretation.

Project aims

D. cinerea, while native to the African continent, poses a severe threat to natural savanna biomes due to its aggressive encroachment within these areas. *D. cinerea* is particularly problematic due to its rapid colonization of these areas coupled with its ability to reproduce asexually via root suckering, allowing an increased advantage over naturally occurring species by not depending only on seed dispersal for reproduction. This makes *D. cinerea* an interest to ecologists concerned with the conservation and diversity within the savanna biome. Additionally, the role of plant clonality in savanna ecology has not been well studied but it is thought that it may play an important role both in initial encroachment, and in reinvasion following clearing [21]. The major barrier to studying clonality in non-model plant species such as *D. cinerea* is the absence of the necessary genetic resources to allow the application of population genetic techniques. The aim of this project was to develop a ddRAD-seq pipeline for high throughput marker discovery which would allow future studies of clonal reproduction in *D. cinerea* on a larger scale than has previously been possible. This will allow the impact of environmental factors, such as drought which has been shown to impact *D. cinerea* stem density [22] and/or human interventions on the mode of reproduction in this species to be determined.

Methods and materials

Sample collection

A total of 218 *Dichrostachys cinerea* trees were sampled from four areas across this species range within South Africa. In Kwa-Zulu Natal (KZN) 49 samples were collected from the Manyoni Game Reserve (M1 - M49), with an additional 7 samples collected from the town of St Lucia (S1 – S7). A further 25 samples (T1 – T25) were collected from the Tarentaalnes Gastehuis in the North-West Province (NW). Finally, 137 individuals (K1 - K137) were collected from the southern Kruger National Park (KNP, permit number: SS1353), seven of which were sampled in duplicate to serve as technical replicates in order to estimate the error rate in ddRAD-seq, thus the total number of leaflet samples collected for sequencing was 225. In Manyoni, KNP and Tarentaalnes a transect was performed where samples were obtained from all individuals along a linear distance of 30 – 50 meters. The collected leaf tissue from each individual was stored in silica gel. The latitudinal and longitudinal coordinates were recorded for each individual via INaturalist to provide a location estimate for each individual (Figure 5), however due to their close proximity to one another, samples which belonged to the same transect were assigned the identical location coordinates.

Leaflet dimension measurements

Using a Tork Craft™ digital vernier caliper, leaflet dimension data (leaflet length and width) was collected from the dry leaf material. To ensure accuracy and consistency, the leaflet length and width was determined from five or more leaflets until a CV value < 15 % was obtained. Once the ploidy of individuals within the population had been established, a Mann-Whitney test was used to assess whether leaflet dimensions were significantly different between diploid and tetraploid individuals.

Autonomous selfing experiments

To determine whether *D. cinerea* is capable of self-pollination we bagged 15 developing inflorescences prior to anthesis using pollinator-exclusion bags made of bridal veil (mesh diameter <1 mm) and observed whether any fruit pods had formed four weeks later.

DNA extraction

Total genomic DNA was extracted using an established CTAB based protocol targeted toward recalcitrant plant species [79], however, some modifications were made to optimise the protocol for high-throughput DNA extraction of dry leaf *D. cinerea* material (see results). In the final protocol used, around 50 mg of dry leaf tissue was ground up into a fine powder and added to 1 mL of CTAB extraction buffer (100 mM Tris-HCl pH 8, 25 mM EDTA pH 8, 1.5 M NaCl, 2% (w/v) cetyltrimethylammonium bromide (CTAB), 1% polyvinylpyrrolidone (PVP) and 0.3% (v/v) β-mercaptoethanol) and incubated at 65°C for 50 min, with mixing by inversion of the samples every 10 min. The solution was then centrifuged at 6000 × g for five min and the supernatant was transferred to a new microfuge tube. One volume of chloroform:isoamyl alcohol (24:1) was added and the solution was inverted for five min. Thereafter the solution was centrifuged at 6000 × g for five min, and the upper aqueous solution was transferred to a

new microfuge tube and treated with 10 µg of RNase A for 15 min at 37°C. Thereafter one volume of chloroform: isoamyl alcohol (24:1) was added and the solution was inverted for five min. The sample was centrifuged at 6000 × g and the upper aqueous phase was transferred to a new microfuge tube and mixed with 0.5 volume of 6 M NaCl, 0.1 volume of 3 M potassium acetate (50% (v/v) 5 M potassium acetate, 11.5% (v/v) glacial acetic acid) and 0.66 volume of the resulting final volume of ice-cold isopropanol. Samples were then incubated at -20°C for 30 min to precipitate DNA and then centrifuged at 8800 × g for 15 min. The resulting pellet washed twice with 1 mL of 70% (v/v) ethanol and resuspended in 100 µL TE (10 mM Tris-HCl pH 8, 1 mM EDTA) buffer overnight at 4°C prior to nanodrop measurement.

DNA quality assessment

DNA concentration and purity was evaluated using a Nanodrop. DNA integrity was assessed by electrophoresis on a 0.7% (w/v) TAE agarose gel infused with ethidium bromide (0.05 µL mL^{-1}). As restriction enzyme digestion is a critical step in ddRAD-seq, between 0.5 µg and 1 µg of each sample was digested with 10 U HindIII (Thermo Fisher Scientific) for 1 hour at 37°C and run alongside an undigested aliquot of the corresponding DNA for each sample to determine whether the extracted DNA could be completely digested by restriction enzymes.

ddRAD-seq Library generation and sequencing

Double digest and barcode ligation

Prior to sequencing library preparation, more precise DNA concentration values were determined via a Qubit fluorescence assay. The samples were then divided into six libraries with a maximum of 48 samples per library. Using Qubit readings, samples were individually aliquoted into a 96-well plate to ensure a sample-well contained around 100 ng of DNA, while also not exceeding 20 µL in volume. A double-restriction enzyme digest was then performed using EcoRI-HF (New England biolabs, NEB) and TaqI (NEB) by adding 2.5 µL CutSmart® Buffer (NEB), 2 µL water, 0.4 µL of each enzyme. The samples were then incubated in a thermocycler at 37°C for 30 min, then 65°C for 30 min, and finally at 80°C for 20 min to inactivate the restriction enzymes. Thereafter ligation was performed to ligate the P1-barcoded adapter and P2-biotin adapter (available at https://figshare.com/s/8f08cd720c1ab0af4fb5) to the resulting fragments by the addition of 2 µL of 3 µM P2-Biotin adapter, 2 µL of 0.3 µM P1 adapter, 3 µL 10 mM rATP, 0.8 µL 10x T4 ligase buffer, 1 µL T4 Ligase (NEB) to each DNA fragment mix. The mixture was then incubated at 23°C for 20 min and then 65°C for 10 min. This gave each sample in the library a unique barcode sequence to allow sample-to-sequence association. All DNA samples of a library were then pooled for subsequent size selection.

Size selection

Size selection was performed on each library to generate an average fragment size of 550 bp. For each library, 150 µL of the pooled solution was transferred to a microfuge tube where 90 µL of AMPure XP beads (0.6x of the total volume) was added and the mixture was left to incubate at RT for 10 min. The solution was then placed on a magnetic stand to pull-down the undesired sized fragments. The supernatant was then transferred to a new microfuge tube where 0.12x volume AMPure XP beads were added and incubated for 10 min at RT. The

solution was then placed on a magnetic stand and the supernatant was discarded. Keeping the tube on the magnetic stand, the pellet was washed twice with 70% (v/v) ethanol and left to air dry for up to 10 min. Once dry, 30 µL of water was added to resuspend the pellet. After two min of incubation, the solution was once again placed on a magnetic stand and the supernatant, containing the DNA library with the desired fragment length, was transferred to a new microfuge tube.

PCR amplification

Dynabeads M-270 Streptavidin beads (Invitrogen) were used to select for fragments with P2-biotin labelled adapters. Prior to this step, beads were first washed three times in 100 µL of the supplied Dynabeads Binding and Washing (BW) buffer then resuspended in 30 µL BW buffer. The entire 30 µL bead solution was then mixed with the size selected DNA and incubated at RT for 15 min, ensuring the solution was mixed every 5 min. Thereafter the solution was placed on a magnetic stand for 2 min. The bead pellet was washed thrice with 100 µl of 1xBW buffer and resuspended in 45 µL of water. PCR was then performed to add the DNA sequences necessary for Illumina flow cell compatibility. A unique primer pair was selected for each library. A mastermix of the components necessary for the PCR was generated containing; 45 µL of the size selected bead suspension, 3 µL of each primer (10 µM) of the primer pair, and 50 µL KAPA HiFi Hotstart ReadyMix. The mixture was then divided into four new microfuge tubes to minimize amplification bias (~ 25 µL per tube). After eight cycles of PCR (98°C for 20 sec, 65°C for 20 sec and 72°C for 30 sec) the PCR reactions were pooled into a new microfuge tube. The solution was placed on a magnetic stand to pellet the beads, and the supernatant was transferred to a new microfuge tube.

PCR cleanup and sequencing

As a final clean up procedure, 0.7x AMPure XP beads were added to the PCR mix and incubated for 10 min at RT. A magnetic stand was used to collect the beads, which were washed twice with 70% ethanol and left to air dry for 10 min. The resulting pellet was resuspended in 20 µl nuclease free water. After a two min incubation the solution was placed on a magnetic stand to separate the beads from the now complete library. The final library was stored at -20°C until sequencing. To assess library fragment distribution, a High Sensitivity D1000 assay was performed on each of the final libraries (Agilent 4150 TapeStation system). A Qubit fluorescence assay was also performed to ensure a library DNA concentration between 1-8 ng μL^{-1}. The resulting ddRAD-seq libraries were sequenced on the Illumina NovaSeq 6000 sequencing instrument for PE150 reads by Novogene (Cambridge, UK).

Whole genome sequencing and analysis

Total DNA extracted from sample K47 was sequenced via Illumina NovaSeq 6000 sequencing instrument for PE150 reads by Novogene (Cambridge, UK). FastQC was used to assess the sequencing quality of FASTQ files. The jellyfish software was used to count kmers (21mers), providing the -C parameter to count the canonical representation of strands. Jellyfish histo was then used the generate a histogram of kmer count to provide as input for GenomeScope 2.0. Basic genome characteristics were then estimated with GenomeScope 2.0, with the initial kmer coverage estimate provided as 10 via --kcov parameter [80]. MEGAHIT (v 1.2.9) was used

to generate a rudimentary assembly of the resulting paired end sequencing data, and a BUSCO (v 5.7.1) analysis was performed on the resulting assembly, having the fabales lineage dataset specified in the command line.

Bioinformatic analysis of ddRAD-seq data

Demultiplexing and quality assessment

For each ddRAD-seq library, paired-end sequencing resulted in separate FASTQ files for the forward and reverse reads. The data was demultiplexed via the process_radtags program originating from the Stacks software pipeline [44]. The --inline-null flag was used, as with ddRAD-seq data the barcode is present on a single-end read even though the data is paired end, along with the -q and -r flags to discard any reads with low quality scores and rescue barcodes/RAD-Tag cut sites respectively. Consequently, this grouped all reads containing the same barcode into a single file per individual which could be renamed with custom bash script to ensure each set of reads were correctly assigned to a unique sample identifier. FastQC (v0.12.1) was used to generate a report of the sequencing reads per sample [81] and compiled with multiQC to gain an overview of sequencing quality.

Genotyping and SNP quality control

Demultiplexed FASTQ files were then supplied as input to dDocent (v 2.9.4) , a software pipeline that provides a user-friendly alternative to using the Stacks [44] pipeline as it aims to autonomize the SNP calling process with minimal user input, requiring only the demultiplexed FASTQ files which have been renamed to suit dDocent specific file name requirements [49]. The dDocent pipeline was implemented to identify SNPs by aligning to a *de novo* reference genome constructed from the pooled ddRAD-seq reads form all samples. To achieve this, the software first filters sequencing reads based on user specified input on coverage thresholds to reduce the number of reads considered for assembly in order to save on computational resources and run-time. Reads were then clustered via CD-HIT [82] whereafter they are supplied to Rainbow for assembly of a *de novo* reference sequence [83]. dDocent then mapped sequencing reads to the resulting *de novo* assembly via BWA [84] and SNP calling was performed via Freebayes [85]. Default dDocent values were used for clustering and alignment.

The resulting VCF file was then filtered to reduce the amount of low quality and uninformative SNPs present in the data via VCFtools (v 0.1.16) [46]. To this end, any SNPs with a minor allele frequency of below 5% were excluded. The VCF file was also filtered to only keep variants that were bi-allelic SNPs, and with a minimum read depth of 5. SNPs with any missing data across all individuals were also removed, resulting in 39299 SNPs across the 225 samples. SNP filtering followed recommendations by Ellis et al. (2020) [45].

In silico ploidy estimation

To gain insight into the ploidy of *D. cinerea*, the allelic depth of all sites was extracted from the filtered VCF file using bcftools (version 1.3.1) [47] and imported into R (v4.4.1) where the minor allele frequencies were calculated and used to infer ploidy based on the distribution of

minor allele frequency around expected frequencies. To elaborate, the reasoning follows that in a diploid genome, the mean read counts at heterozygous positions have a single mode at 0.5 where half of the reads account for each allele, while in a tetraploid individual the mean read counts allow for modes at 0.25, 0.5 and 0.75 [52, 63, 86]. All minor allele frequencies values were filtered to reduce noise and minor allele frequency plot for each individual were generated using this filtered dataset. Multidimensional scaling was performed on all samples via the SNPRelate package in order to assess degree of genetic variability between samples of different cytotypes.

Clonality and population structure

Identity-by-state analysis was then performed on all individuals to assess relatedness between individuals via the SNPRelate package (v 1.38.0) [87]. The VCF file containing all individuals was first converted to a genomic data structure (GDS) file for use in SNPRelate. Identity-by-state analysis was then performed on the GDS file, whereafter hierarchical clustering was performed to generate a dendrogram of all individuals. From the VCF file containing all samples, all diploid individuals collected from the Kruger National Park area were isolated to generate a new diploid- only VCF file via the vcfR package (v1.15.0) [88]. Similarly, identity-by-state analysis was performed on diploid individuals and a dendrogram was used to visualize population structure and identify potential clonal individuals. The STACKs populations (v 2.65) program was then used to assess population level diversity statistics and to determine the level of inbreeding and heterozygosity within this population.

Flow Cytometry

Flow Cytometry experiments were conducted on *D. cinerea* tissue to verify *in silico* estimates of ploidy. Flow cytometry experiments were performed on both fresh and dry plant material. Fresh tissue included well known standards for plant-nuclei based flow cytometry [68]. Fresh tissue for *D. cinerea* was available from seedlings germinated from seeds obtained within the NW province, as well as a sapling originating from within KNP which was purchased at the Skukuza Indigenous Plant Nursery. Flow cytometry experiments were also performed on a subset of dry tissue samples already included in the ddRAD-seq sequencing libraries for which *in silico* ploidy estimates were available.

The final protocol for nuclei extraction and flow cytometry analysis of both fresh and preserved tissue used woody plant buffer (WPB) (200 mM Tris-HCl pH 8, 4 mM $MgCl_2$, 2 mM EDTA, 86 mM NaCl, 10 mM sodium metabisulfite, 1% (w/v) PVP-10, 1% (v/v) Triton X-100) as described by Loureiro et al. [89]. Eighty mg of leaf tissue was chopped up on ice for 1 min in 1 mL of WPB. The solution was then mixed via pipetting several times to ensure nuclei were in suspension. The nuclei-tissue mixture was then filtered through a 40 µm cell strainer (SPL Life Sciences) and treated with propidium iodide (50 µg mL^{-1}) and RNase (50 µg mL^{-1}), then left to incubate for five min on ice. Samples were analysed with a BD Fortessa (LSRII) flow cytometer. The parameters analysed were propidium iodide intensity (PI) to identify fluorescently stained nuclei, forward scatter (FS) for a rough estimation of particle size, and side scatter (SS) for a rough measure of particle optical complexity. Gain settings were as follows: 280 FSC, 230 SSC and 300 ECD. For each dataset gating was performed to ensure the minimization of debris and

false positive peaks. The data were first subjected to a time gate to ensure no blockages or air bubbles had occurred. Data were then loosely gated to exclude events with both high FC and SS. Highly localized populations of events with a similar PI were then selected as possible nuclei, whereafter these events were then gated to exclude doublets. The final PI histogram is then used to determine values for G1 peaks. All flow cytometry results were analysed and generated using FlowJo™ v10.10.0 software (BD Life Sciences).

To estimate genome size, the following equation was used as described by [68]:

$$2C\ content\ of\ standard\ (pg)\ x\ \frac{sample\ G1\ peak\ mean}{standard\ G1\ peak\ mean} \quad 1$$

Results

Sample collection

Sample collection occurred in three primary regions: Kruger National Park (KNP), the Rustenburg area (NW) and Manyoni Private Game Reserve (KZN_MA), with an additional seven samples obtained from the St. Lucia area (KZN_ST) (Figure 5). The sampling thus covered a large part of the species' range of *D. cinerea* within South Africa. At each site, opportunistic sampling was conducted over several kilometres to assess baseline genetic diversity at these sites. To assess for potential clonality between individuals in close proximity, samples were collected along a 30 – 50 m transect. In KNP, technical replicates were taken from seven individuals (K115, K116, K117, K119, K122, K123 and K127) to allow for estimation of ddRAD-seq error rate.

Leaf samples were obtained from 218 individuals, with seven of these trees being sampled twice as technical replicates, giving a total of 225 samples. The majority of the samples (n = 137, excluding the seven technical repeats) collected originated from KNP across an area of ~ 500 km^2. Within the KNP area the degree of encroachment varied from highly dense thickets of *D. cinerea* individuals to areas where individuals were sparsely spread throughout the collection area. Additionally, within the KNP samples were collected along two transects; one 50 m in length consisting of individuals K115 to K137 as well as a 20 m long transect consisting of individuals K98 to K112. Technical replicate samples were included in order to allow for the estimation of the error rate that occurred during the sequencing of ddRAD-seq libraires, as although the genetic sequence of technical replicates should be identical to one another, sequencing error will result in a small amount of apparent genetic divergence between the technical replicates. Estimation of this error rate therefore allows for the detection of clonal individuals within the population.

Sample collection within the NW region (T1 to T25) occurred in a single site within a severely encroached farmland within the Rustenburg area. Interestingly, a large thicket was present on the farmland with individual trees exhibiting close proximity to one another. Samples collected from the KZN region were collected across two sites. The first site was located on the grounds of Manyoni Private Game reserve (M1 to M49) which consisted of both opportunistic sampling of individuals (M1 to M25), as well as a linear transect of individuals (M26 to M49) and resulted in an estimated sampling area of 15 km^2. The second site was located within the town of St Lucia (S1 to S7) which only consisted of opportunistic sampling.

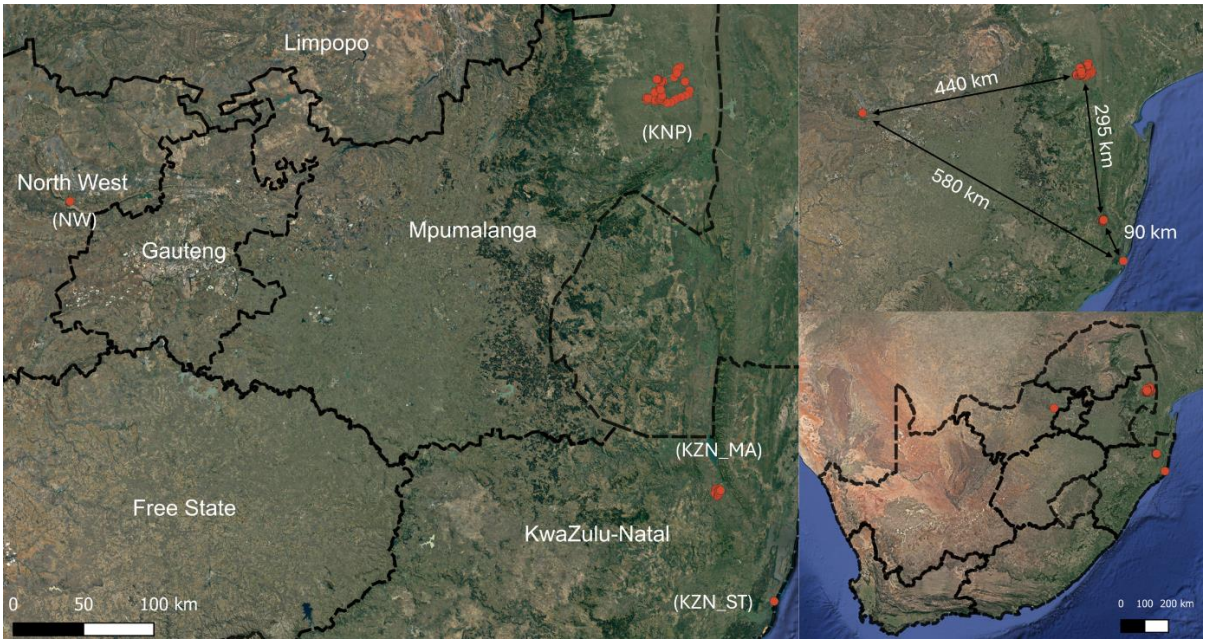

Figure 5. Sample map indicating the geographical positions for samples collected across the three major regions and the distances between the collection sites.

Optimization of DNA extraction protocol

An essential first step in ddRAD-seq is the isolation of high-quality genomic DNA. There are many published protocols for the extraction of genomic DNA (gDNA) from silica-dried plant material [90-92], but it is often necessary to modify these for any given plant species. During the optimization period, modifications were implemented to allow the development of a procedure that would isolate high quality and intact genomic DNA that could undergo complete digestion by restriction enzymes, as this is one of the key steps in the ddRAD-seq library generation protocol. The characteristics of high-quality DNA which is suitable for NGS are described in Healy et al. [79]. A high quality DNA sample would exhibit low levels of both carbohydrate/secondary metabolite contamination and protein contamination as indicated by an A260/230 ratio close to 2 and an A260/280 ratio between 1.8 – 2 respectively [79]. Additionally, the isolated DNA should exhibit a single high molecular weight (HMW) band with minimal evidence of shearing or degradation when analysed using gel electrophoresis, indicating that it is intact. The DNA extraction protocol also needed to yield several micrograms of high-quality DNA per sample which could then undergo endonuclease digestion. Results pertaining to the optimization steps taken to achieve these qualities are provided below.

Initial DNA extractions were performed on silica dried *D. cinerea* leaflet tissue according to the protocol of Healey et al. [79], which uses 95% (v/v) ethanol and 5 M NaCl during the DNA precipitation step. The DNA yield per extraction provided over 10 µg of DNA from only 20 mg dry tissue, an amount much larger than the minimum required DNA amount (100 ng) for the generation of ddRAD-seq sequencing libraries [43] (Table 1). However, the resulting spectrophotometric measurements for protein contamination (A260/280) were below the 1.8

– 2 range recommended. Similarly, A260/230 ratios were far below the threshold suggested (Table 1), which implied that modification of the DNA extraction procedure was necessary to counter carbohydrate/secondary metabolite contamination, a common problem when extracting nucleic acids from plant samples. To assess the integrity of the extracted DNA, and whether it can be cleaved by restriction enzymes, gel electrophoresis was performed on 1 µg of gDNA following incubation at 37°C for 60 min with or without HindIII (Figure 6). This indicated the presence of intact HMW gDNA which was completely digested by HindIII.

Table 1. Spectrophotometric analysis results for naïve DNA extraction following the protocol described in Healey et al. [79]. The DNA extraction was performed in duplicate (PL-1A/B).

Sample	A260/280	A260/230	Concentration (ng/µL)
Sample 1	1.67	0.95	118.8
Sample 2	1.63	0.94	130

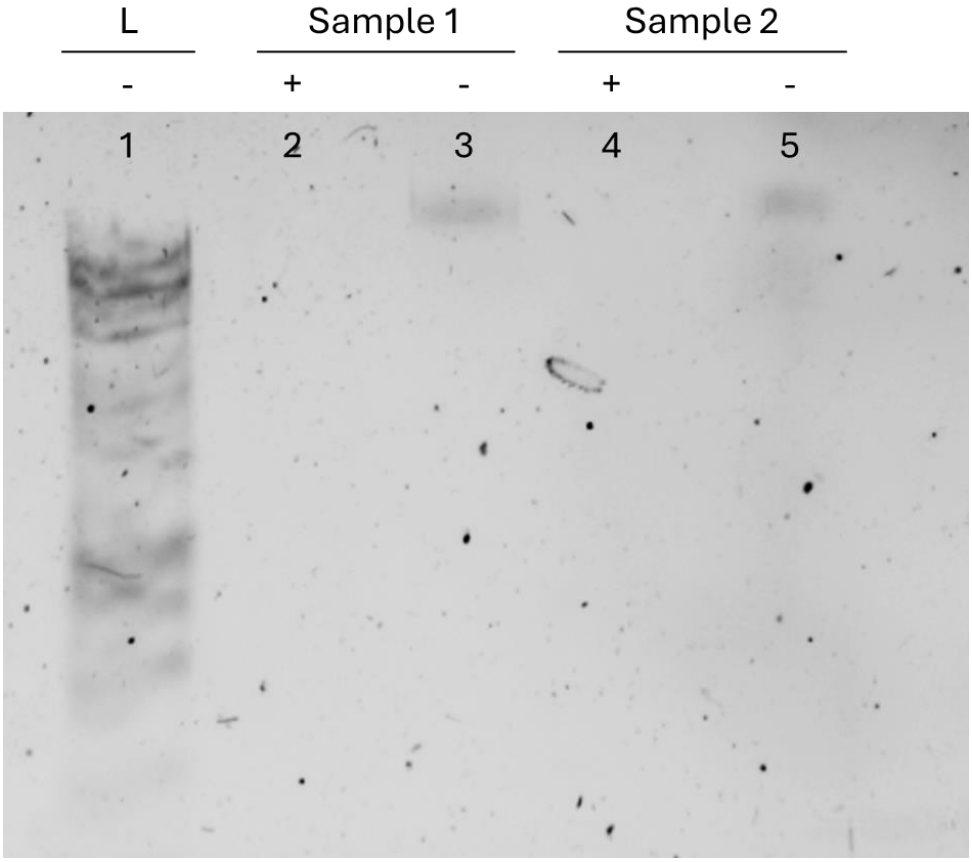

Figure 6. Electrophoretic separation of 1 µg DNA on an 0.7% agarose gel, indicating whether DNA had undergone enzymatic digestion via HindIII (+) or not (-) prior to loading. A 1 kbp ladder (L) was loaded in lane L.

As the A260/230 and A260/280 ratios were lower than recommended, the protocol was modified to replace the components used in the DNA precipitation step with those

recommended by Aboul-Maaty et al. [93] which uses a NaCl, potassium acetate and isopropanol precipitation. The modified precipitation step was performed on triplicate samples at either RT or -20°C, each for 30 minutes to assess if any differences in performance would result due to this change. Spectrophotometric measurements indicated a reduced DNA yield per sample, regardless of incubation temperature when compared to the original protocol (Table 2). Furthermore, there was no obvious improvement in 260/280 ratios using this modified protocol, and 260/230 ratios were highly variable. Nonetheless, gel electrophoresis revealed that the DNA isolated was intact, with a single HMW band observed, and could be completely digested by HindIII (Figure 7).

Table 2. Spectrophotometric measurements obtained from following the protocol by Healey et al. [79], however substituting the DNA precipitation step mixture for the mixture provided by Maaty [93]. Samples were either incubated at room temperature (-RT) or -20°C (-20) for 30 minutes.

Sample	A260/280	A260/230	Concentration (ng/µL)
Sample 1-20	1.35	-2.53	8.9
Sample 1-20	1.61	2.57	28.7
Sample 1-20	1.02	-1.57	5.2
Sample 1-RT	1.26	1.12	13.7
Sample 1-RT	1.35	1.41	12.3
Sample 1-RT	1.65	2.89	31.1

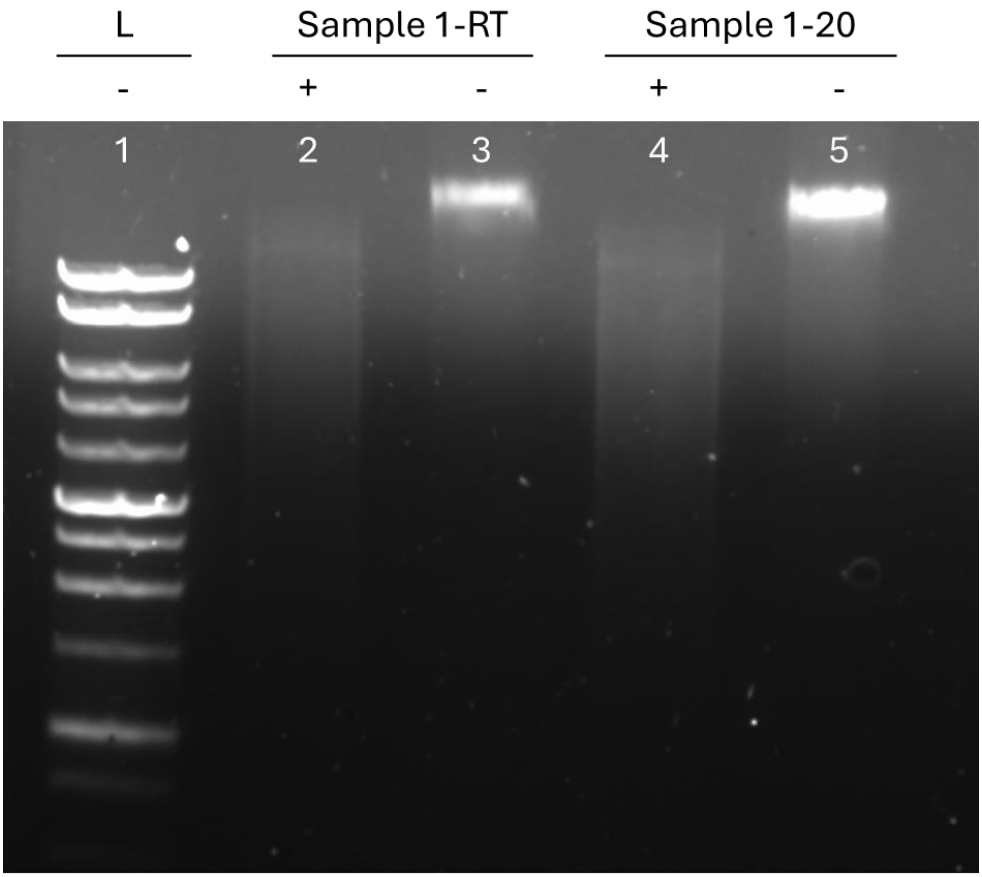

Figure 7. Electrophoretic separation of 0.5 µg of DNA resulting from the modified protocol. During the DNA precipitation step DNA was either incubated at RT (Sample 1-RT) or -20°C (Sample 1-20). Only extracts with the highest yield per experimental group were assessed, indicating whether DNA had undergone enzymatic digestion via HindIII (+) or not (-) prior to loading. A 1 kbp DNA ladder was loaded in lane L.

As this modified protocol resulted in reduced yield, with no real improvement in 260/230 or 260/280 ratios, further optimization was necessary. Numerous authors have suggested that the addition of polyvinylpyrrolidone (PVP) may reduce carbohydrate contamination in the final DNA extract [79, 94]. To assess whether the addition of PVP (mol wt. 10 000) would improve A260/230 ratios, 1 % (w/v) PVP was included in the extraction buffer. DNA precipitation was again performed as per Aboul-Maaty et al. [24]. Spectrophotometric measurements of duplicate extractions (Table 3) indicated a modest improvement in both A260/280 and A260/230 ratios in extractions containing 1 % (w/v) PVP when compared to extractions performed in the absence of PVP (Table 3). Furthermore, measurements suggested that DNA yield was several fold higher in extractions containing PVP when compared to extractions devoid of PVP (Table 3). Electrophoretic separation of both duplicate DNA extracts containing PVP revealed intact HMW bands which could undergo complete digestion by HindIII (Figure 8).

Table 3. Spectrophotometric measurements obtained from extractions following the modified protocol either supplemented with 1 % PVP or 0 % PVP.

Sample	% PVP (w/v)	A260/280	A260/230	Concentration (ng/µL)
Sample 1A	1	1.85	1.41	139.6
Sample 1B	1	1.92	1.57	250.6
Sample 2C	0	1.67	1.29	39.1
Sample 2D	0	1.92	1.39	15.1

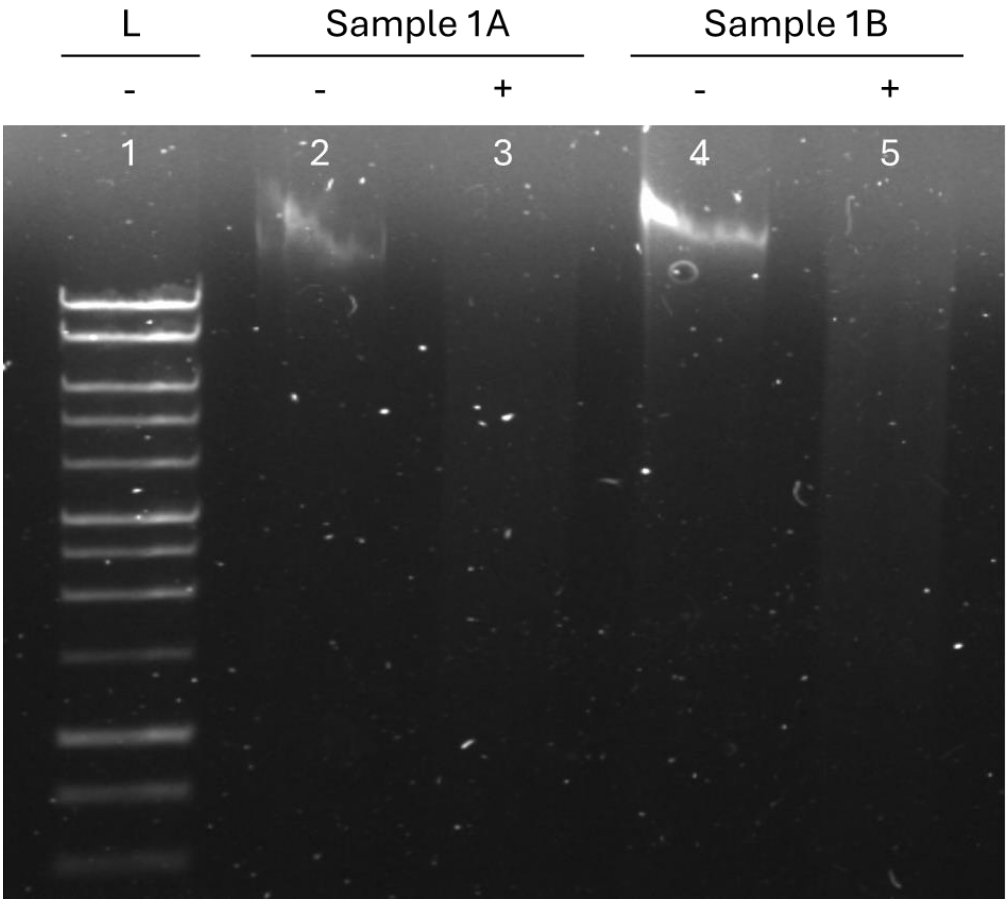

Figure 8. Electrophoretic separation of 1 µg of DNA extracted from PVP containing duplicates (PL-1A/B), also indicating whether DNA had undergone enzymatic digestion via HindIII (+) or not (-) prior to loading. A 1 kbp DNA ladder was loaded in lane L.

To assess whether this optimized protocol was functional across a range of field-collected samples to be included in the sequencing libraries the optimized protocol was performed on seven samples collected from KNP. Spectrophotometric measurements revealed that A260/280 ratios were between, or close to, the recommended 1.8 to 2 range (Table 4) and A260/230 ratios remained consistent between samples, ranging from 1.41 to 1.82. Furthermore, this method consistently provided several micrograms of DNA per extraction (Table 4). Electrophoretic separation of DNA revealed all samples yielded intact high molecular

weight bands which could be completely digested by HindIII (Figure 9). This protocol was thus used to isolate DNA from all samples collected in the field.

Table 4. Spectrophotometric measurements obtained from extractions following the final protocol on seven silica-dried samples collected from KNP.

Sample	A260/280	A260/230	Concentration (ng/µL)
K10	1.81	1.45	99.9
K11	1.87	1.62	209.1
K13	1.79	1.41	68.4
K14	1.74	1.48	216.9
K37	1.91	1.82	138.4
K90	1.87	1.71	184.2
K91	1.89	1.75	191.5

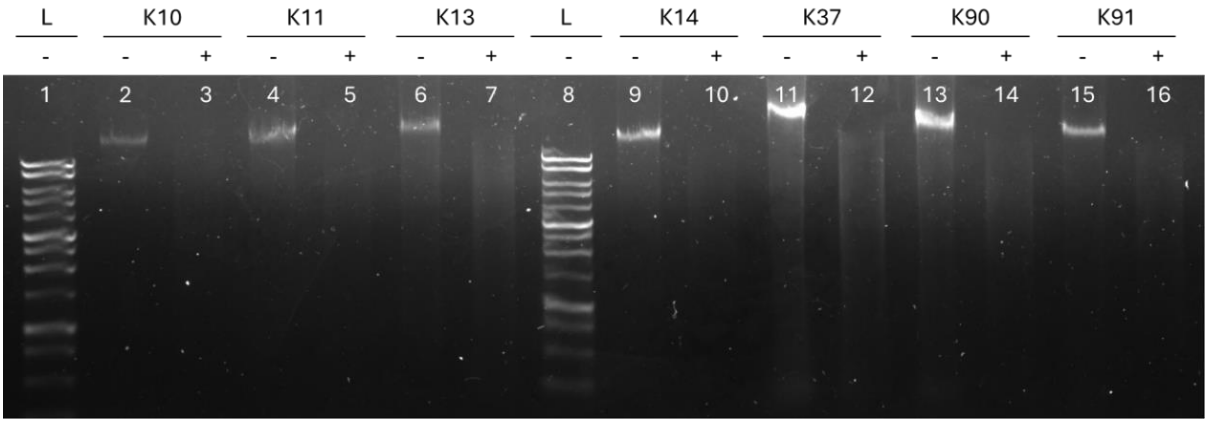

Figure 9. Electrophoretic separation of 0.5 - 1 µg of DNA extracted from KNP samples, indicating whether DNA had undergone enzymatic digestion via HindIII (+) or not (-) prior to loading. A 1 kbp DNA ladder was loaded in lane L.

Estimation of *D. cinerea* genome size using GenomeScope 2

Prior to the study, a review of the available literature at the time revealed that no prior sequence information for *D. cinerea* was available, apart from a complete chloroplast genome and some gene sequences used as markers for prior studies [95, 96]. In an effort to obtain an estimate of the genome size of *D. cinerea,* genomic DNA was extracted from sample K47 and sent for low coverage whole genome short read Illumina sequencing. Sequencing data was then assessed via FastQC analysis to investigate sequencing quality, whereafter the data was then used to estimate preliminary genome characteristics using genomescope 2, a kmer-based approach to infer genome characteristics [29, 80].

Whole genome sequencing of sample K47 resulted in 27.5 GB of 150 bp paired-end Illumina short read data. The sequencing quality of the resulting data was then assessed via FastQC analysis, which revealed high average sequence quality scores (>30) for both forward and

reverse reads (Figure 10A). FastQC analysis also revealed that virtually all base calls could be assigned with confidence across all reads (Figure 10B). FastQC analysis also revealed that the read-set contained a cumulative percentage count of around 1% of adapter sequences, suggesting that the adapter sequence was present in only a small portion of the read-set and could be disregarded, alleviating the need for adapter trimming (Figure 10C). Furthermore, no substantial differences within the GC content could be identified between the forward and reverse reads (Figure 10D). FastQC analysis thus suggested that the sequencing data was of high quality.

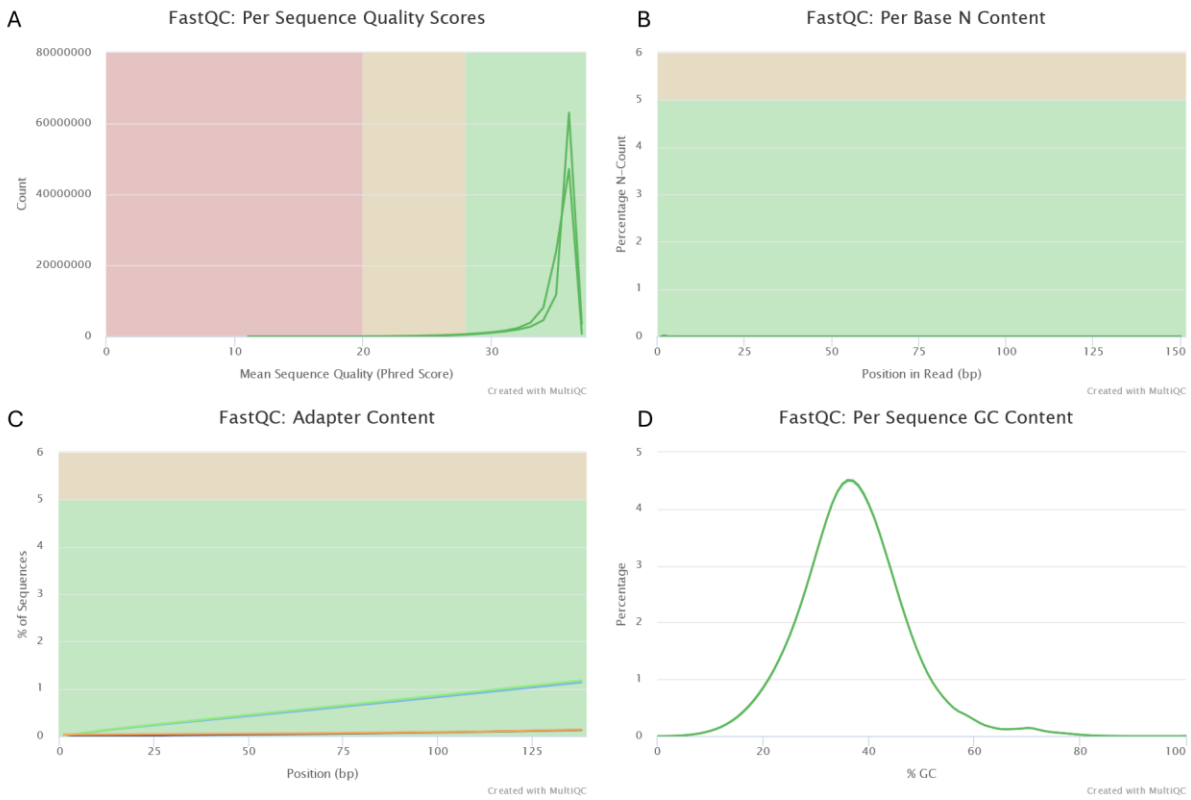

Figure 10. FastQC assessment of the quality of sequencing reads available for sample K47 - per sequence quality scores (A), per base N content (B), adapter content (C) and per sequence GC content (D).

GenomeScope is a software package that is able to provide overall genome and read characteristics without the need for a reference genome, as it infers these characteristics via statistical analysis of the read-set kmer profile [29]. Visualization of the kmer profile revealed only a single kmer coverage peak around 10x coverage (12.5x), thought to correspond to the homozygous peak (Figure 11). The absence of a clear heterozygous peak was most likely due to the coverage of the WGS data being lower than the recommended coverage required for adequate GenomeScope 2 modelling, causing the heterozygous peak to merge with the error peak, preventing the model from confidently identifying the heterozygous peak withing the kmer spectra [29]. Sample heterozygosity was estimated to be 4.21%, however, this estimate

was most likely inflated due to the low coverage of sequencing which resulted in the merging of erroneous base pair calls (sequencing error) and kmers with different alleles (heterozygous sites) into a single peak, therefore, heterozygosity measurements resulting from GenomeScope analysis were not regarded as a true representation of the heterozygosity of the species genome (Figure 11). GenomeScope analysis estimated the genome size of sample K47 to be ~0.9Gbp.

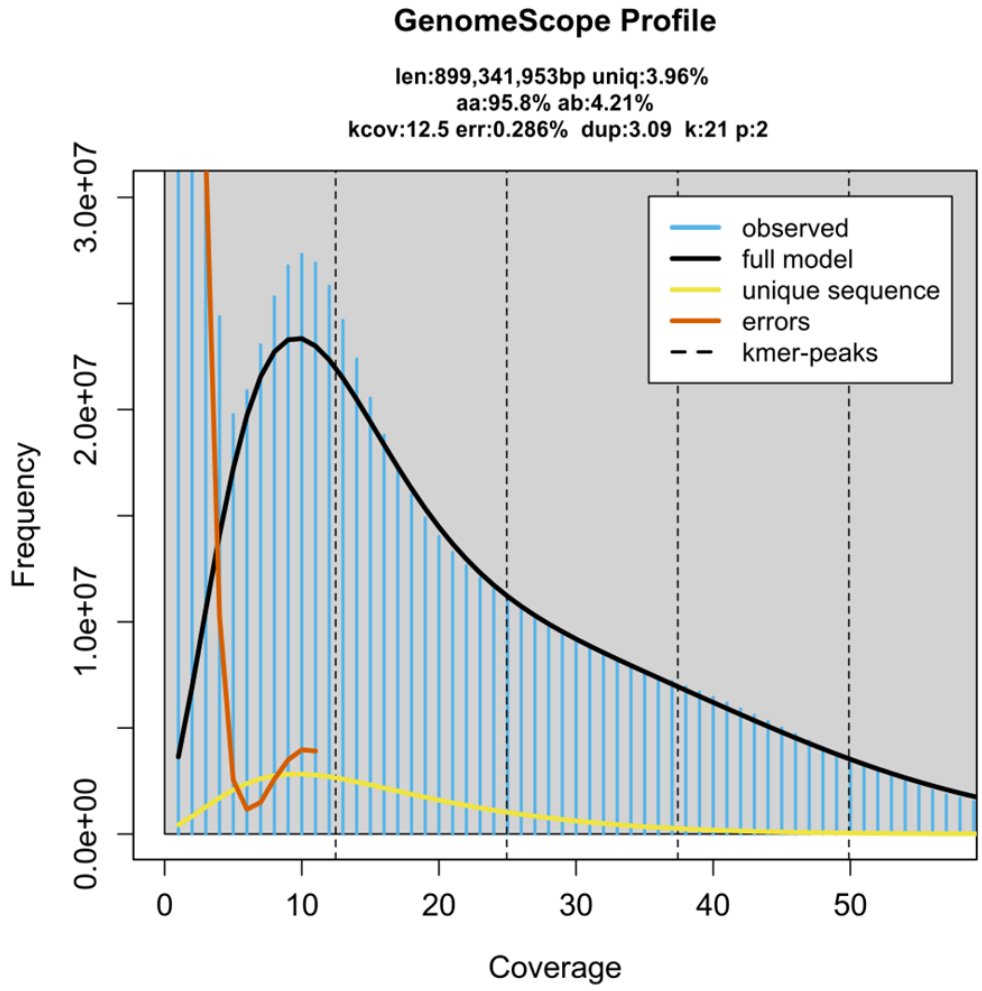

Figure 11. GenomeScope 2 analysis of raw sequencing data obtained from sample K47 short read sequencing.

Prior to this study, limited genetic data was available for *D. cinerea*. This led to the question of whether the sequencing data that was made available by the WGS sequencing of sample K47 would be sufficient to generate a suitable reference genome for SNP calling. This however was unlikely due to the depth of the sequencing, around 10x according to GenomeScope analysis, being below the minimum requirements for the generation of a high-quality genome assembly [97, 98]. Regardless, the raw paired-end sequencing data was provided to the MEGAHIT software in order to perform a basic assembly of the data to assess the quality of the assembly. Analysis of the scaffolds N50 statistic revealed a value of 311 kbp (Appendix A

Table A1). Thereafter, a BUSCO analysis was performed on the assembly, revealing the BUSCO statistic Complete: 73.6% [Single-copy: 67.2%, Duplicated: 6.4%], Fragmented: 7.4%, Missing: 19.0%, n:5366 (Appendix A Figure A1). Granted the sequencing depth was low, no further analysis was performed using the WGS data.

ddRAD-seq library generation and quality control

After successfully extracting high-quality gDNA from all samples of interest, the next step involved generating six ddRAD-seq libraries for sequencing and SNP identification. Double restriction enzyme digest was performed on 100 ng of gDNA from each sample with a combination of EcoRI-HF and TaqI in order to fragment the DNA in preparation for adapter ligation and size selection. Following digestion, samples within a library were uniquely barcoded by the ligation of P1 barcode adapters which allowed for the separation of total library sequencing data into sample-specific read sets. Simultaneously, P2-biotin adapters were also ligated to DNA fragments to allow for downstream fragment isolation. All samples of a library were then pooled, and size selection was performed to isolate fragments for PCR amplification. Following library generation, a quality assessment was conducted to evaluate fragment size and concentration across all libraries.

Tapestation analysis of each ddRAD-seq sequencing library revealed a fragment size distribution of around 500 bp with a mean fragment length of 566±21 bp for all libraries (Table 5). Comparison of library fragment length revealed no additional fragment peaks that correlate with the presence of primer dimers or adapter dimers and also indicated that all libraries showed a similar fragment distribution around the average fragment size (Figure 12). Library DNA concentrations obtained via qubit fluorescence assay were also within the acceptable range of 2-4 nM as recommended by the official Illumina website https://knowledge.illumina.com/library-preparation/general/library-preparation-general-reference_material-list/000001252 (Table 5).

Table 5. Agilent Tapestation results for each sequencing library generated. Library DNA concentration was determined via Qubit fluorescence assay and converted to nM according to the official Illuminal website recommendations.

Library	Average size (bp)	Concentration (ng/uL)	Concentration (nM)
1	592	2.26	5.78
2	578	2.6	6.82
3	538	2.56	7.21
4	550	2.86	7.88
5	549	2.84	7.84
6	589	3.86	9.93

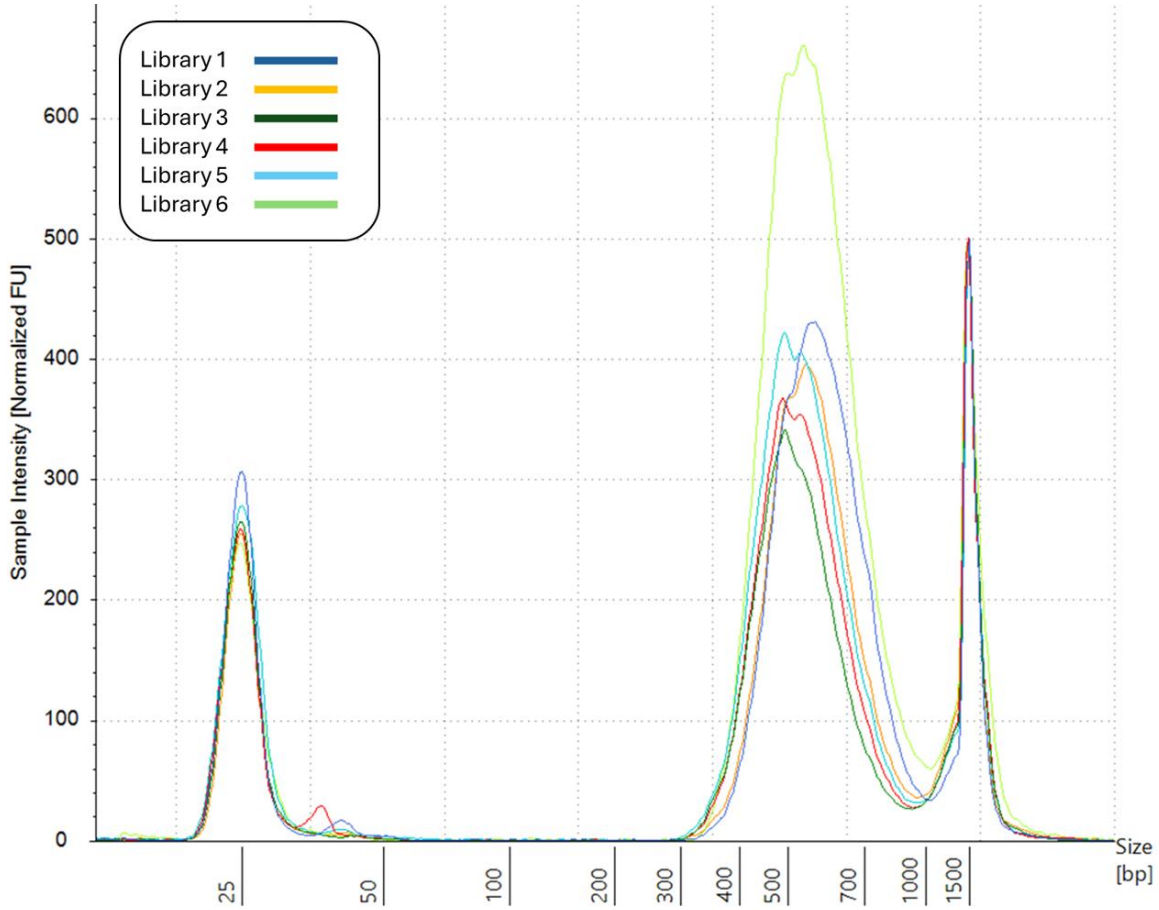

Figure 12. Comparison of all six ddRAD-seq sequencing library fragment distributions. Fragment size (bp) indicated on the x-axis while normalized fluorescence units (FU) shown on the y-axis.

FastQC analysis of the paired end read files for all samples showed no significant sequencing errors in either the forward or reverse reads, with the sequencing reads displaying a high average Phred quality score (±36) (Figure 13A). Furthermore, the majority of base calls were made with strong confidence, as the percentage of positions where 'N' was called was below 1% (Figure 13B). Samples also exhibited a low degree of cumulative adapter contamination across reads (Figure 13C). The GC content measured across the length of each sequence in a file revealed a ± 2 % difference in the GC distribution for a proportion of the samples (Figure 13D). During the process of generating a ddRAD-seq sequencing library, DNA fragments undergo both PCR amplification and size selection, which consequently generates a high degree of duplication of each fragment to be sequenced [43], as seen in Figure 14.

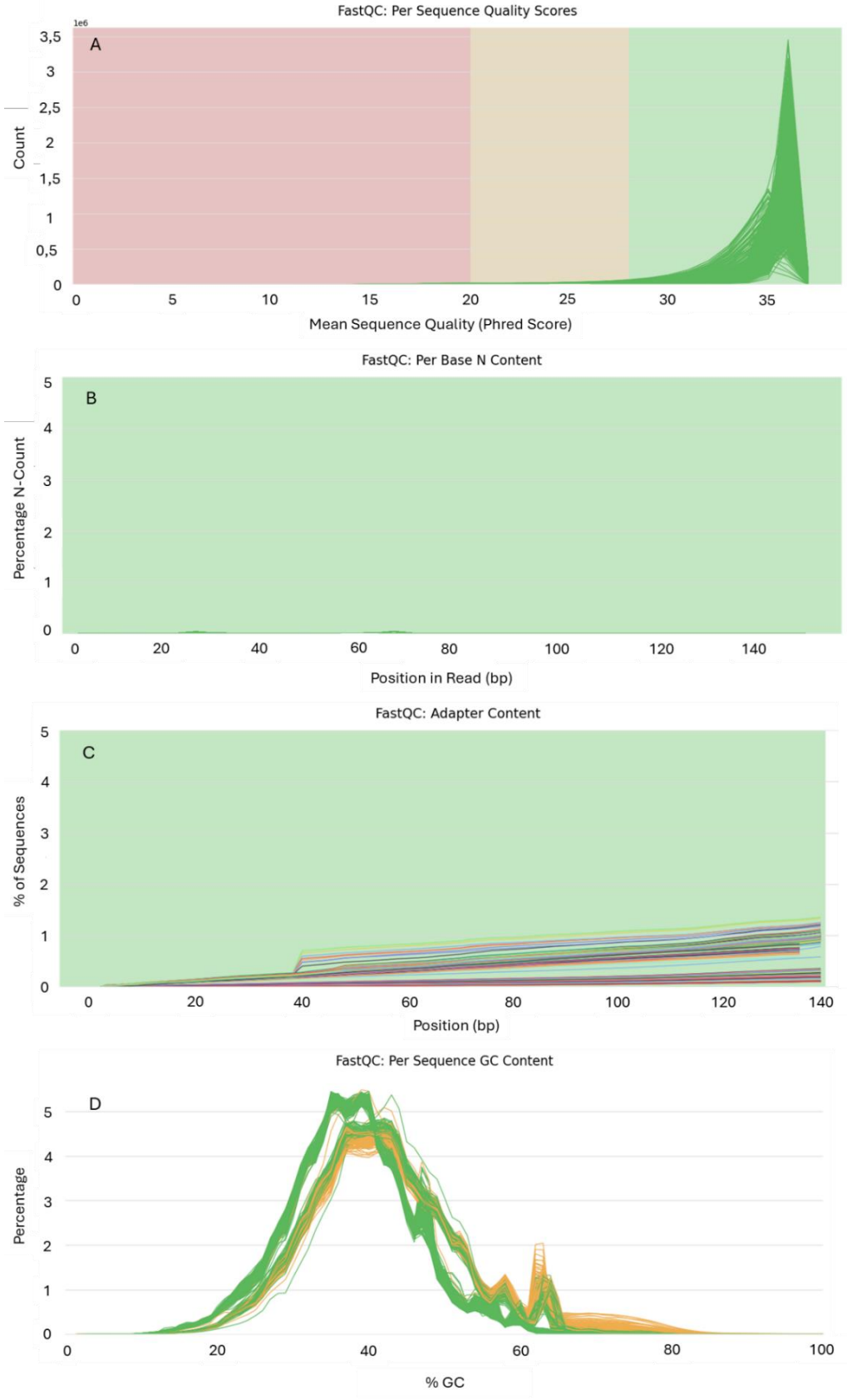

Figure 13. FastQC analysis of each sample compiled via MultiQC. Both forward and reverse reads were analysed with fastQC. Per sequence quality scores determined for the reads of each individual (A), per base N content (B), cumulative adapter content across all reads (C) and GC content for all reads belonging to a sample (D).

Figure 14. The percentage of duplicate reads across all samples sequenced. Due to the nature of ddRAD-seq library preparation steps a high proportion of duplicate reads are present in all samples.

Population genetic analysis of *D. cinerea*

As the FastQC analysis indicated that the sequencing data was of high quality, the next step performed was to identify SNPs. Sequencing data was first demultiplexed via the process_radtags program from the STACKS software, in order to correctly assign reads to their unique sample within a sequencing library [44]. The dDocent pipeline was then used to achieve two aims. Firstly, dDocent would generate a *de novo* reference genome assembly based on the provided ddRAD-seq reads. Although WGS data was made available by earlier sequencing experiments (Figure 10), a high-quality reference genome assembly could not be generated due to the low sequencing coverage [98]. Second, dDocent would automate the SNP calling process and provide a raw VCF file from which the final SNP set would be derived.

Variant calling based on alignment to the ddRAD-seq based *de novo* genome, resulted in 1 143 007 biallelic SNP sites identified across all 225 samples. The removal of indels, and both low quality and uninformative SNPs resulted in 39 299 SNP sites remaining for downstream analysis. The filtered SNP set was then used to perform identity-by-state (IBS) analysis to assess the genetic relatedness among samples in the study, based on the proportion of shared SNPs. Following IBS analysis, hierarchical clustering was performed on the resulting distance matrix in order to generate a dendrogram to assess relatedness between individuals. This analysis revealed that the samples from the NW, KZN_MA, and KZN_ST populations clustered together on geographical location (Figure 15). However, this was not the case for the samples from KNP. Seventeen of the KNP individuals were found within the KZN_MA clade, while 28 clustered with the KZN_ST individuals. Thus, the genetic distance between individuals does not appear to be solely driven by the geographical distance between them.

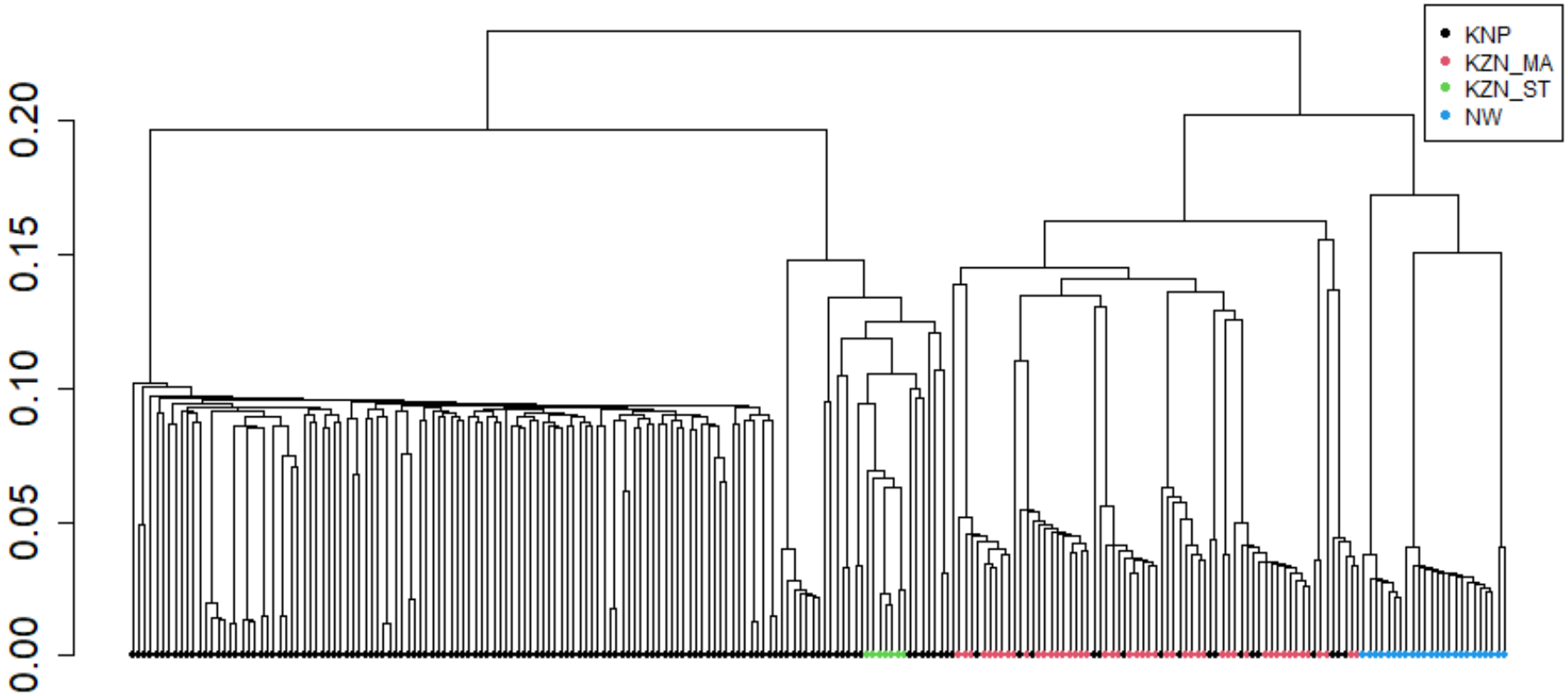

Figure 15. Dendrogram showing the genetic relatedness between individuals, based on the IBS score of an individual (y-axis). IBS analysis was performed on the total sample space, whereafter hierarchical clustering was performed to generate the dendrogram. Geographical location of the individual sampled is indicated by colour.

One possible explanation for this is that our sampling might have included individuals from the two currently recognised subspecies of *D. cinerea*, ssp. *africana* and ssp. *nyassana*. The exact status of these subspecies in South Africa is unclear as both their distributions and morphological characters overlap in this region. Leaflet width, leaf size and peduncle arrangement are all continuous characters in *D. cinerea* in South Africa [3], and there is currently no genetic evidence supporting the separation of these taxa. Nonetheless, as leaflet width (>2 mm in fresh samples) is the character most commonly used to assign a specimen to ssp. *nyassana* [3], we measured this parameter (and leaflet length) in our silica dried samples. Measurements were taken from a minimum of five leaflets until the CV of the mean was <15% to obtain more accurate values. Both leaflet width and length exhibited a bimodal distribution (Figure 16) rather than a normal distribution, suggesting the presence of two distinct groups within our samples

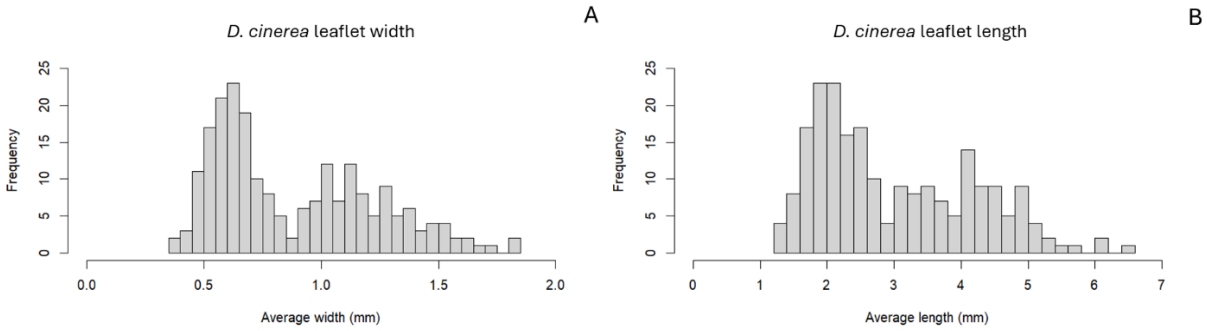

Figure 16. The frequency distribution of average leaflet width (mm) (A) and leaflet length (mm) (B) measurements of all individuals used in ddRAD-seq analysis.

In silico ploidy estimation

Several studies have shown that an increase in ploidy within a member of a species could be associated with an increase in leaf dimensions in comparison to a member of the same species with a lower ploidy level [58-60]. It was therefore hypothesized that the distribution of leaflet size (Figure 16) and the failure of the KNP samples to cluster together (Figure 15) might be attributed to variation in ploidy within *D. cinerea*.

To assess whether there was variation in ploidy among our sampled individuals, read depth data for both the reference and alternative alleles for each sample were extracted from the quality controlled VCF file and imported into R. Read count data was then filtered to reduce excessive noise, whereafter the minor allele frequency (MAF) distribution of each individual was generated in order to assign an *in silico* ploidy estimate for each sample. A multidimensional scaling analysis was then performed to elucidate the effect ploidy may have on the genetic distance between individuals.

Prior to assessing the allele frequency, filtering was performed on the allele depth for each sample to reduce background noise within the read depth data. To this end, sample quantiles

were selected as filter thresholds which would both retain informative data to allow for accurate calculation of MAF values while also minimizing the amount of noise for each sample. For the reference allele read depth, coverage values which fell outside of the 40 – 95 % quantiles were removed (Figure 17A and Figure 17B). Similarly, for the alternate allele read depth, coverage values which fell outside of the 3.5 – 85 % quantiles were removed (Figure 17C and Figure 17D).

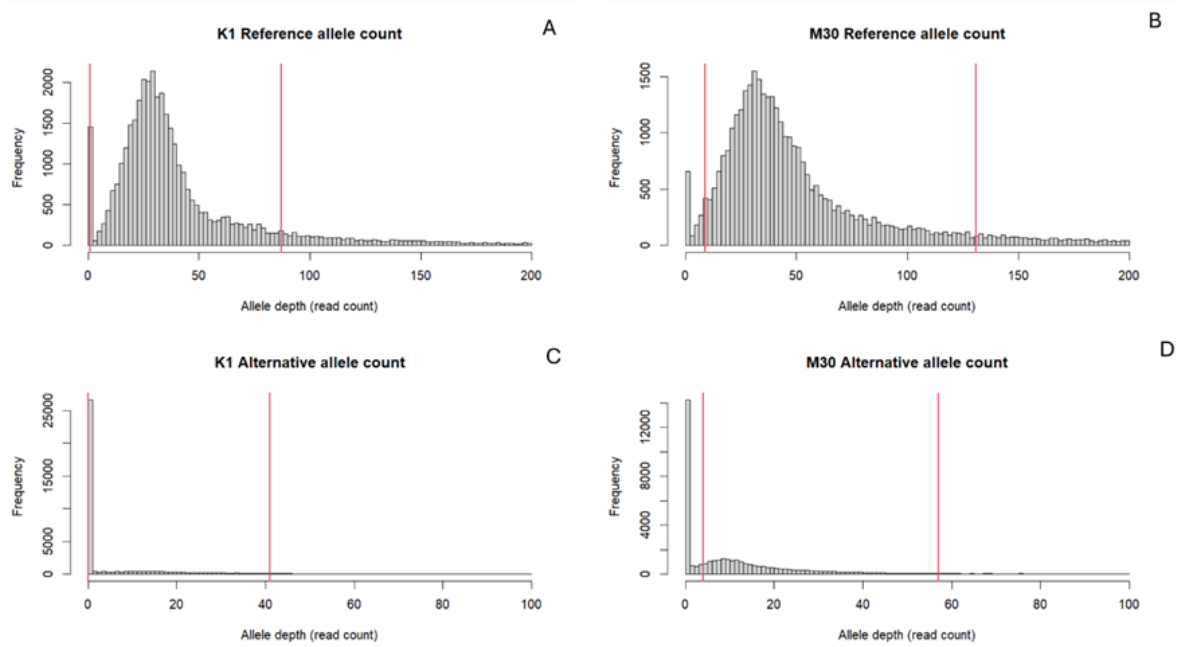

Figure 17. Allele depth filter thresholds were determined to minimize data loss while eliminating the majority of outlier data points. Frequency distributions of read count (x-axis) shown here are for a representative set of samples and are purely for the purposes of visualization.

Prior to filtering, the reference allele depth for all samples had a range of median values between 13 and 91 reads and a range of mean values between 27.93 to 194.17 reads (Figure 18A). After filtering, the reference allele depth for all samples had a range of median values between 14 and 73 reads, with a range of mean values reduced to between 22.03 to 75.22 reads (Figure 18B). Similarly, prior to filtering, the alternate allele depth for all samples had a range of median values between 0 to 16 reads, with a range of mean values between 4.78 to 36.73 reads for all samples (Figure 18C). After filtering, the range of median values for the alternate allele depth remained between 0 and 16, however, the range of mean values for all samples was reduced to between 4.48 and 17.41 reads (Figure 18D). Overall, the filtering process removed a large amount of the existing outliers within the data thereby reducing the background noise.

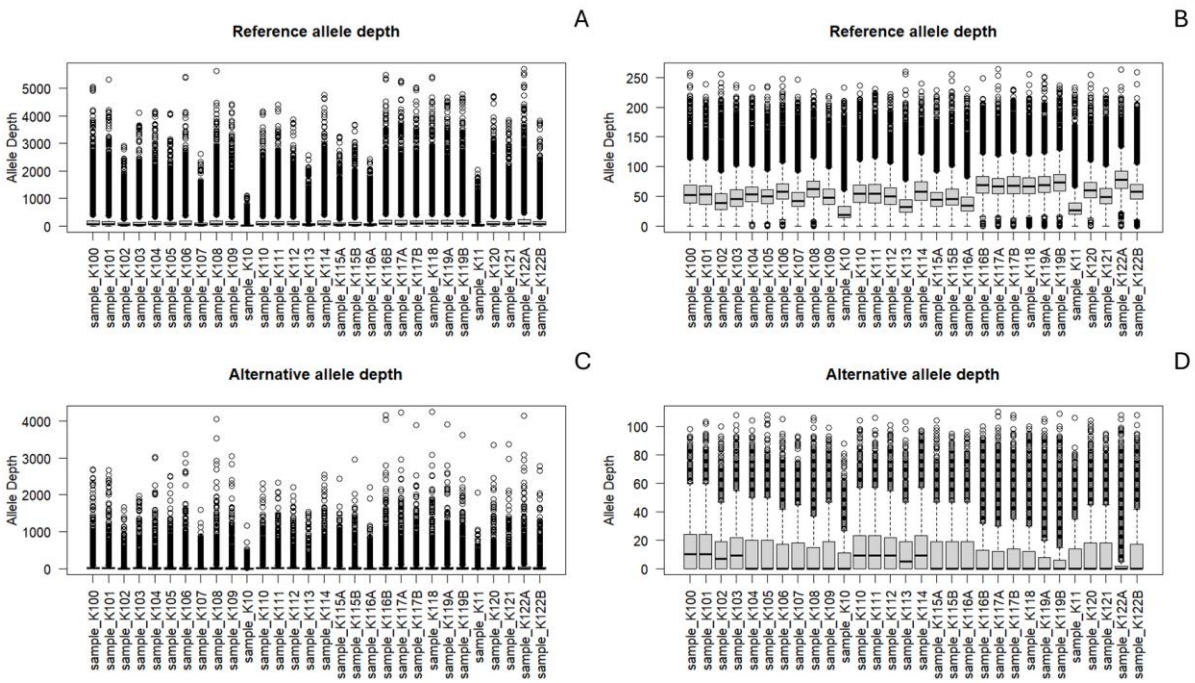

Figure 18. Box plots of allele depth for a representative set of samples. Outlier values of alternative and reference allelic depth were identified (A and C respectively) and the majority were removed (B and D). Allelic depth values of only 30 representative samples are shown, however, filtering was performed on all samples. The letter "A" and "B" after a sample number e.g. K117A and K117B, indicate that these are technical replicates from the same individual.

In silico ploidy estimations were then made for each individual sample. The method was based on whether the distribution of the minor allele frequencies (MAF) for each individual followed an expected distribution around modes known to be associated with a particular level of ploidy [63]. The reasoning follows that within a diploid genome, the mean read counts (allele depth) at heterozygous positions have a single mode at 0.5 where half of the reads should come from each allele, while in a tetraploid individual the mean read counts can display modes at 0.25, 0.5 and 0.75.

This analysis revealed that while 106 samples did follow the distribution of allele frequencies around the 0.5 mode expected for a diploid individual, the remaining 119 samples exhibited a distribution of allele frequencies around the 0.25 mode, coinciding with a distribution expected from a tetraploid individual. *In silico* ploidy estimates revealed that individuals originating from the KNP consisted of a mixture of both diploid and tetraploid individuals as indicated by the distribution of MAF around the 0.5 mode (Figure 19A and Figure 19B) and 0.25 mode (Figure 19C and Figure 19D) respectively. All individuals collected from the NW region appeared to be tetraploid based on the MAF distribution around the 0.25 mode (Figure 19E and Figure 19F). All individuals sampled from Manyoni game reserve in KZN (M1 to M49) were also found to be tetraploid (Figure 19G). Similarly, all samples collected from St Lucia in KZN (S1 to S7) exhibited a distribution of allele frequencies around 0.25 mode (Figure 19H) which suggested that these individuals were also tetraploid.

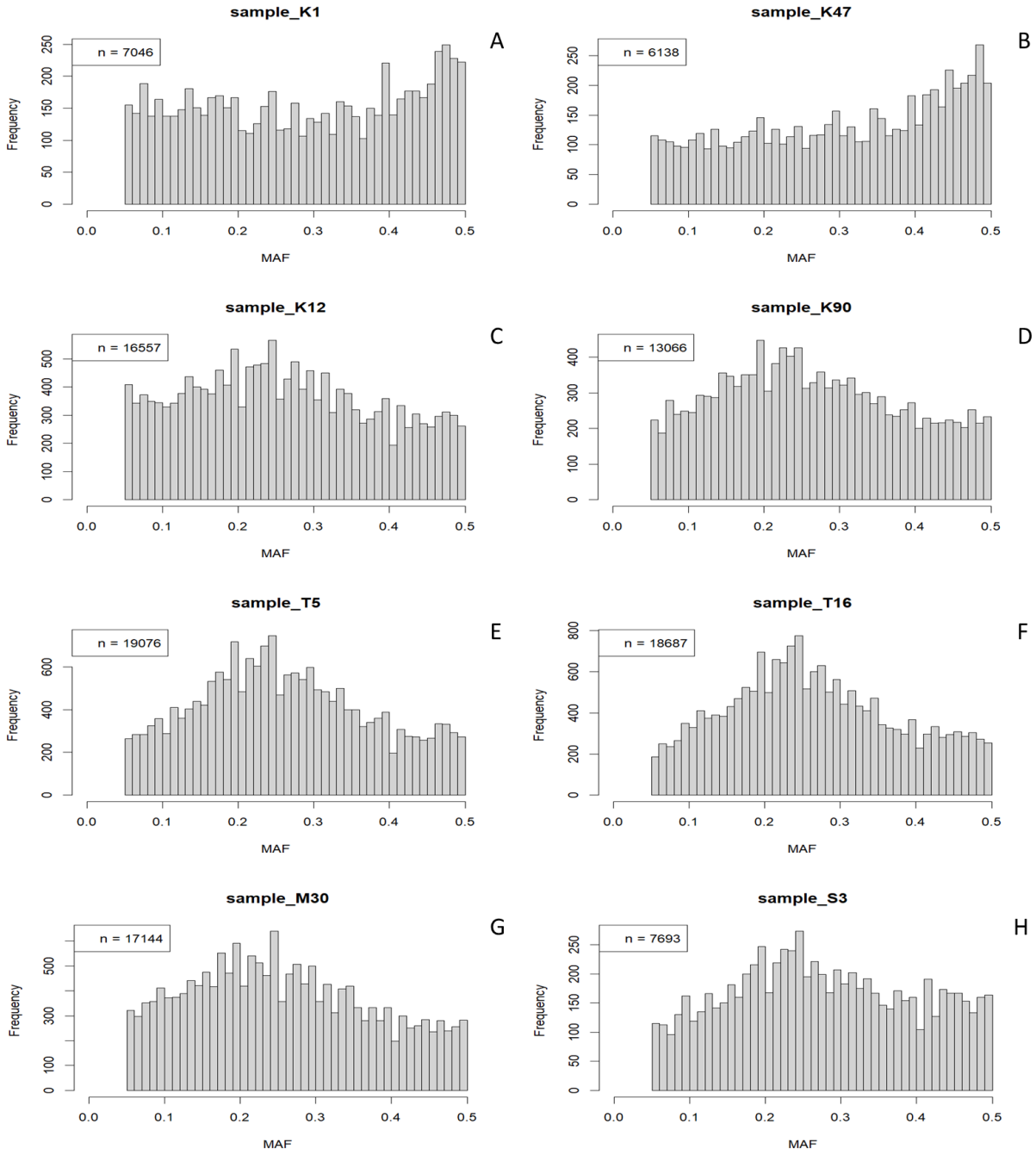

Figure 19. Minor allele frequency distributions of samples which originate from the KNP population (A-D), the NW population (E, F), the Manyoni population (G) and the St Lucia population (H). The number of loci used for each MAF plot is indicated on the top left of each plot (this value excludes loci with a MAF below 0.05)

To understand the relationship between ploidy and the apparent genetic distance between individuals, classical multidimensional scaling analysis was performed on the IBS distance matrix to visualise the genetic distance between individuals from all populations, as well as the degree of variability between tetraploid and diploid individuals (Figure 20). This revealed

that within KNP, all individuals which had been assigned as diploid via *in silico* ploidy assessment clustered separately from the tetraploid individuals within the population, with the diploid individuals displaying a lower degree of variability in genetic distance when compared to tetraploid individuals, which tended to be further separated from one another, i.e. a greater degree of genetic variation. The tetraploid samples from KZN_ST also clustered tightly together (Figure 20). Tetraploid individuals which originate from KNP appear more similar to tetraploid individuals which originated from the KZN_MA and NW populations, despite being collected from sites several hundred kilometres apart (Figure 20, Figure 5). Overall, this analysis suggested that tetraploid individuals exhibited a greater degree of genetic variation than did the diploids and suggested that variation in ploidy contributed to the pattern of relatedness seen in Figure 15.

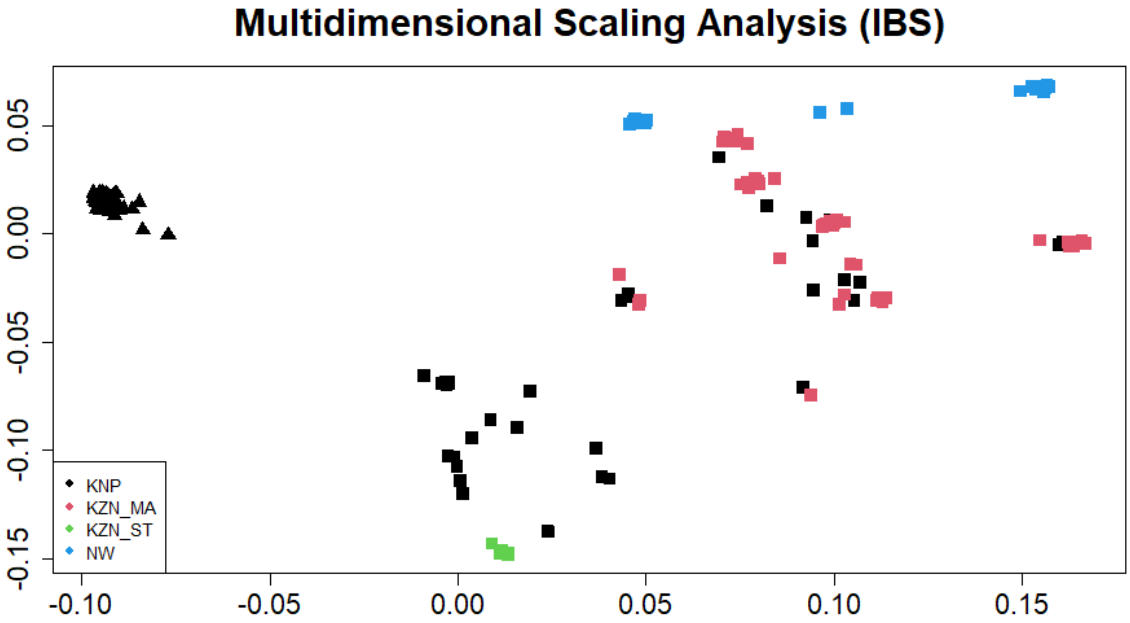

Figure 20. Multidimensional scaling analysis was performed on all the individuals within the sampling population to assess the degree of genetic variation between tetraploid individuals (square) and diploid individuals (triangle).

Assessing the phylogeny of all individuals within the sample space according to their estimated ploidy revealed that samples clustered strongly according to ploidy rather than geographical location (Figure 21). Overall, this suggested that assessing the genetic similarity between diploid and tetraploid individuals would be problematic as the degree of relatedness could not be confidently assessed between these individuals. Additionally, this suggested that the identification of clonal individuals within the replicate samples were taken to estimate the baseline ddRAD-seq error rate were from diploid individuals in the KNP.

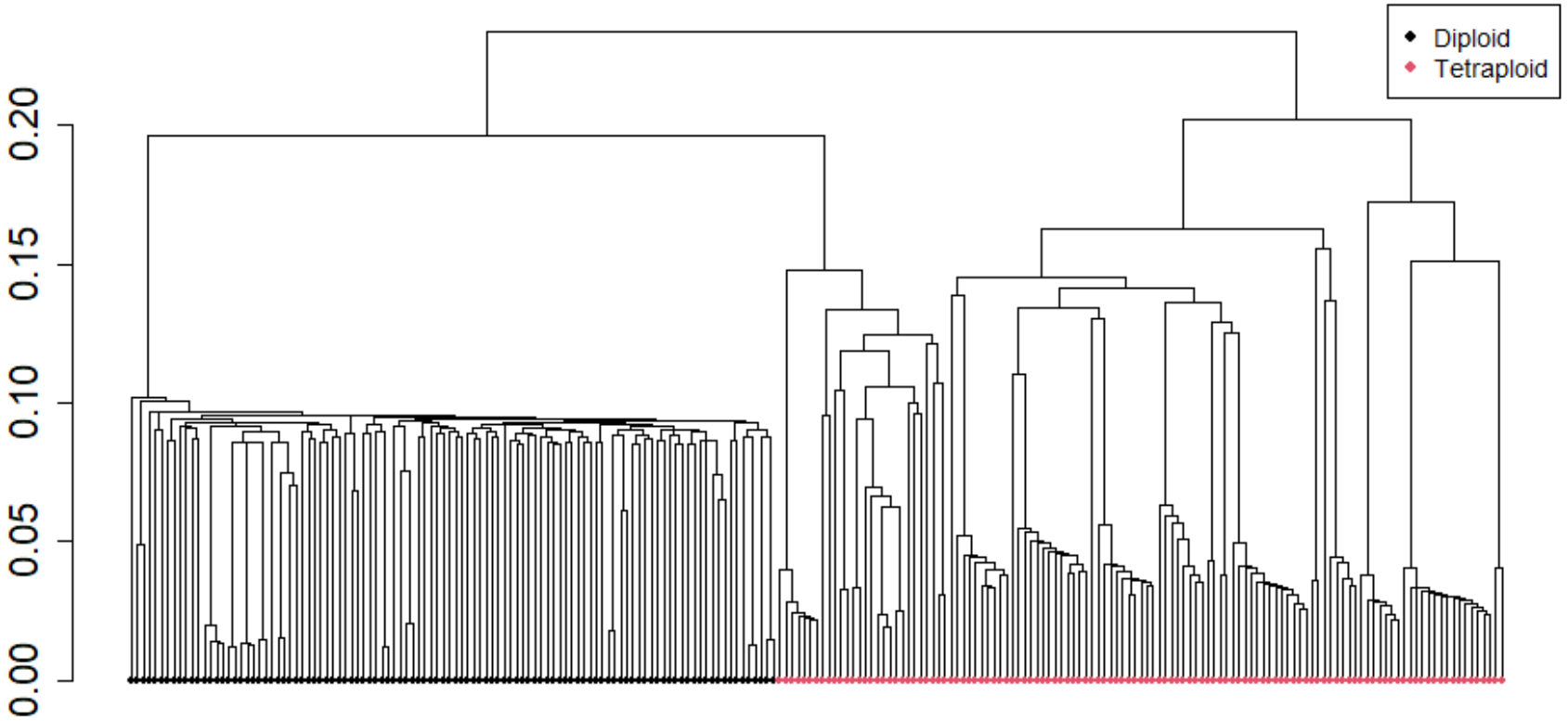

Figure 21. Dendrogram showing the genetic relatedness between individuals, based on the IBS score of an individual (y-axis). IBS analysis was performed on the total sample space, whereafter hierarchical clustering was performed to generate the dendrogram. Ploidy of the individual sample, as determined by MAF analysis, is indicated by colour.

Given the previous positive correlation between larger leaf size and higher ploidy levels reported in other plant species [58-60], a comparative analysis of average leaflet dimensions in diploid and tetraploid *D. cinerea* samples was performed. A Mann-Whitney test found a highly significant difference in both average leaflet width (W = 687.5, p-value < 2.2e-16) and average leaflet length (W = 655, p-value < 2.2e-16) (Figure 22), with both measures significantly larger in tetraploid versus diploid individuals, consistent with the trend reported by the aforementioned studies. Further investigation of the average width data revealed that sample K121 appeared to be an outlier to the diploid width data, however closer inspection of the MAF distribution confirmed it exhibited a diploid distribution.

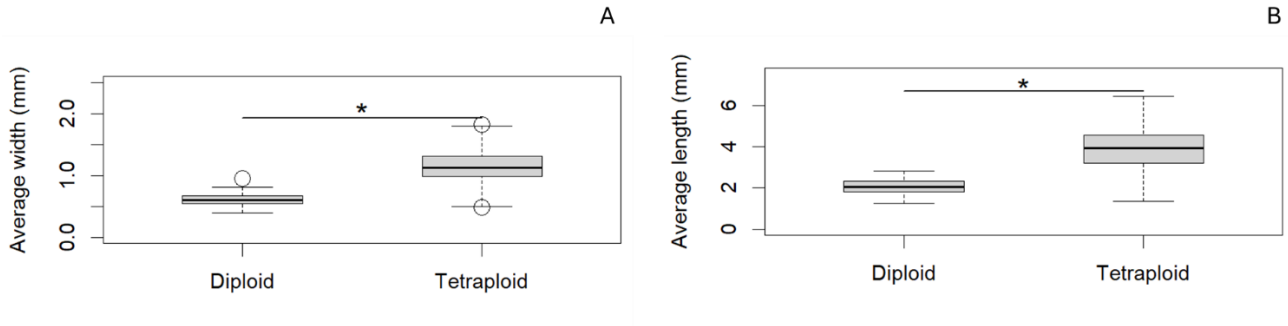

*Figure 22. Box plots of leaflet width (A) and length (B) in D. cinerea individuals identified as diploid or tetraploid by in silico analysis. A Mann-Whitney test revealed a highly significant difference in both measures between diploids and tetraploids. The boxplot displays the first and third quartiles, the black line within the boxplot shows the median of the data. The circles indicate outliers within the data. * Indicates a p-value < 2.2e-16.*

Verification of *in silico* ploidy estimates by flow cytometry

In silico ploidy estimates via allele frequency distributions provided a high throughput manner to estimate the ploidy of all the individuals sampled in this study. To increase confidence in these estimates, flow cytometry analysis of isolated plant nuclei from individuals predicted by *in silico* analysis to be either diploid or tetraploid was performed. Initially, the nuclei extraction procedure was first optimized on *S. lycopersicum* to generate a protocol that would allow for the extraction of intact nuclei from leaf tissue as well as the generation of sharp G1/G2 peaks. Additionally, the protocol would need to ensure that fluorescent background noise caused by debris would be reduced to a minimum. Once the protocol had been optimized, *D. cinerea* tissue was then interrogated to gain genome size estimates from fresh tissue of likely diploid and tetraploid individuals, on the basis of leaf size positively correlating with ploidy. Finally, ploidy was determined from dry tissue for selected individuals to validate the predictions of the *in silico* analysis.

Initial experiments were performed on fresh tissue samples as nuclei are more readily extracted from fresh tissue versus dried tissue. Two sources of *D. cinerea* plant material were

available, specifically a plant obtained from the KNP nursery and seeds from the NW province. Neither sample had been included in the ddRAD-seq experiment and so no predicted ploidy from MAF analysis was available, however, given the highly significant difference in leaflet size observed between diploid and tetraploid individuals (Figure 22) they were tentatively assigned as a either putative diploid (KNP leaflet size of 0.82±0.04 mm by width and 2.66±0.60 mm by length, henceforth referred to as "*D. cinerea* SL") and a putative tetraploid (NW leaflet size of 2.15±0.06 mm in width and 5.92±0.32 mm in length, henceforth referred to as "*D. cinerea* LL").

During preliminary experiments, three different nuclei extraction buffers were selected to assess their capability to produce distinct G1/G2 peaks of nuclei extracted from fresh *S. lycopersicum* tissue, namely, Galbraith buffer [99], Tris-MgCl buffer [100], and Woody Plant Buffer [89]. All three buffers resulted in successful nuclei extraction and staining, yielding sharp G1/G2 peaks for nuclei stained with propidium iodide (PI), while also exhibiting almost no low intensity particles (Figure 23), suggesting a low detection of debris. Comparison of all three histograms for PI intensity peaks showed considerable overlap between the independent runs (Figure 23A), with highly similar mean G1 peak PI intensities (Figure 23B)

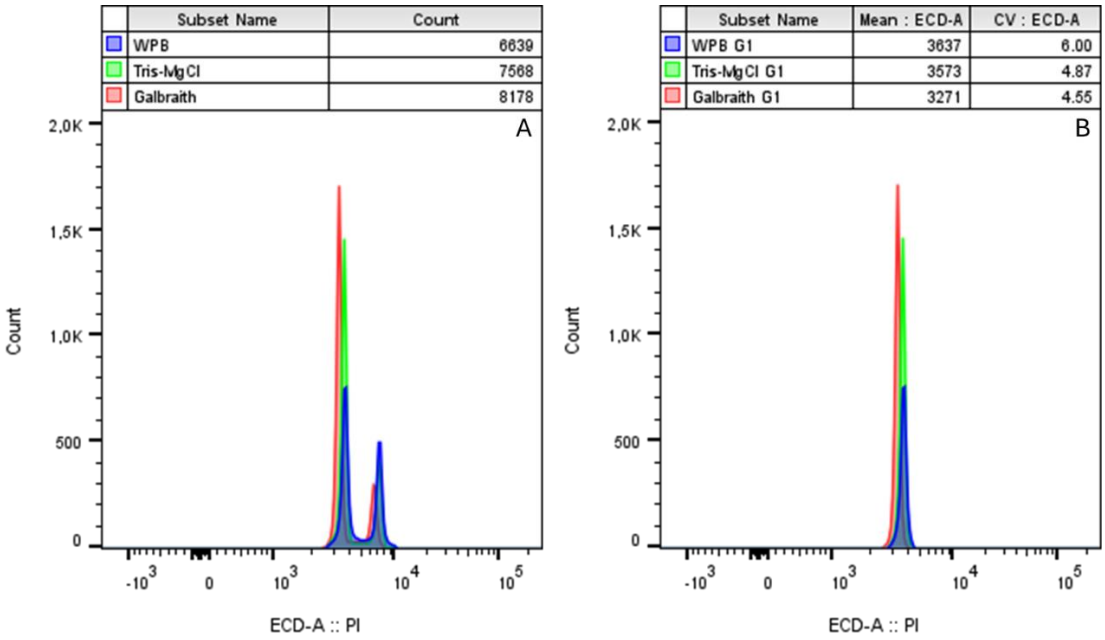

Figure 23. Frequency distributions of PI intensity (x-axis) provide an overview of both G1 and G2 peaks (A) while also allowing for the isolation of sharp G1 peaks (B). DNA was stained with propidium iodide.

D. cinerea is a non-model species with limited knowledge on its biochemistry, however, previous studies have shown that the species harbours several phenolic compounds [8, 9], a group of compounds that may generate issues during flow cytometry analysis [101]. To mitigate these effects, WPB was selected as the buffer of choice for initial experiments on *D. cinerea*, as it has been shown to be effective when used in recalcitrant woody plant species. It contains sodium metabisulfite which acted as a reducing agent and PVP-10 to bind phenolic

compounds, which sought to reduce the interference of phenolic compounds on with fluorescent staining [89]. To assess whether the WPB protocol would successfully detect stained intact nuclei extracted from *D. cinerea* tissue, flow cytometry was performed on fresh *D. cinerea* LL tissue. Results confirmed that the optimized protocol using WPB could produce a sharp G1 PI intensity peak in *D. cinerea* LL, with a smaller G2 peak also evident (Figure 24), with only a minor amount of debris (left shoulder of G1 peak) evident. The G1 PI peak identified had a mean PI intensity of 3930 and CV = 7.9 (Figure 24).

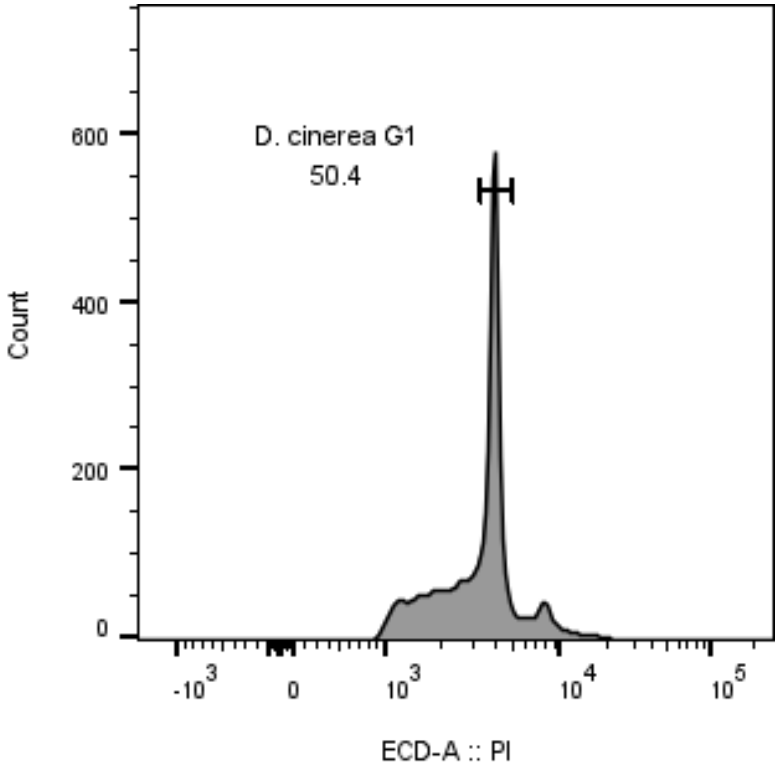

Figure 24. Frequency distribution of the number of events with a particular PI intensity (x-axis) for nuclei extracted from D. cinerea LL tissue using WPB extraction buffer. A clear G1 phase nuclei peak with mean PI of 3930 and CV of 7.9 was evident.

Having demonstrated that the WPB is appropriate for nuclei extraction from *D. cinerea*, an attempt was made to estimate the genome size of the *D. cinerea* LL plant, using *S. lycopersicum* as an internal standard. While at first glance only a single G1 peak is apparent (Figure 25A), closer inspection revealed a shoulder on the right-hand side of this peak, suggesting that the G1 peak from the two species overlaps to a significant degree. Visualising the nuclei in a side-scatter-area (SSC-A) vs PI intensity plot (Figure 25B) confirmed the presence of two distinct populations of nuclei in the G1 peak, as it provided a visualization of the distribution of PI intensity according to the difference in side-scatter light detected across two different nuclei populations. Thus the *D. cinerea* LL plant must have a 2C-DNA content very similar to that of *S. lycopersicum* (1.96 pg).

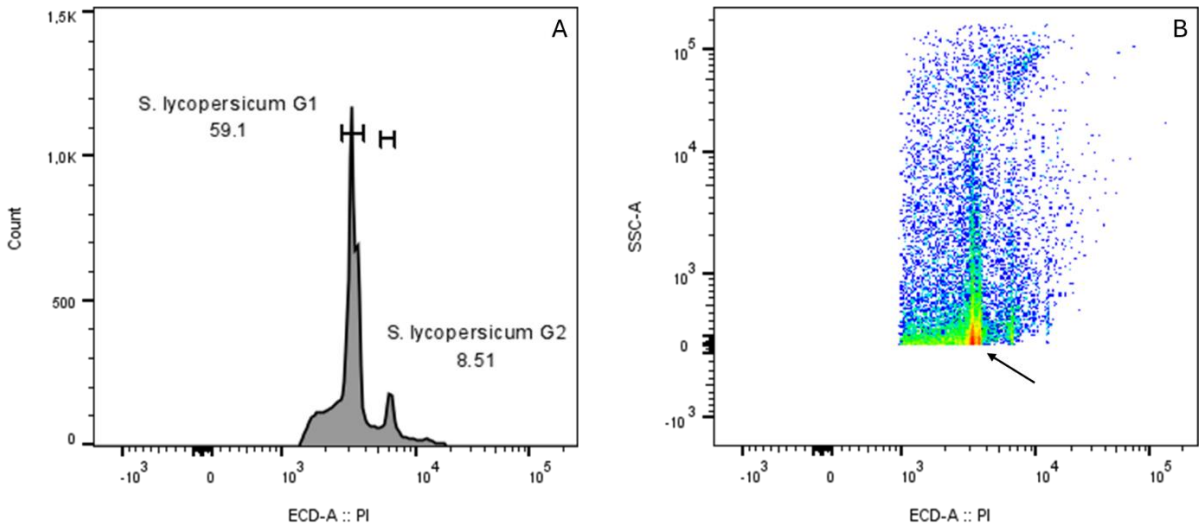

Figure 25. Frequency distribution of PI intensity values (x-axis) after internal standardization of D. cinerea LL with S. lycopersicum, which displays overlapping G1 peaks (A). Closer inspection of the overlapping region with a SSC-A vs PI intensity plot revealed two distinct populations (arrow) with highly similar mean PI intensity values (B).

In light of this, in the next experiment, the 2C-DNA content of the *D. cinerea* LL plant was instead estimated by external standardization with *S. lycopersicum*. This yielded sharp G1 peaks of very similar intensities for *D. cinerea* LL (mean PI = 3930, CV = 7.9) and *S. lycopersicum* (mean PI = 3071, CV = 3.83) (Figure 26), resulting in an estimated 2C-DNA content of 2.51 pg for *D. cinerea* LL equivalent to a genome size of 2.45 Gbp.

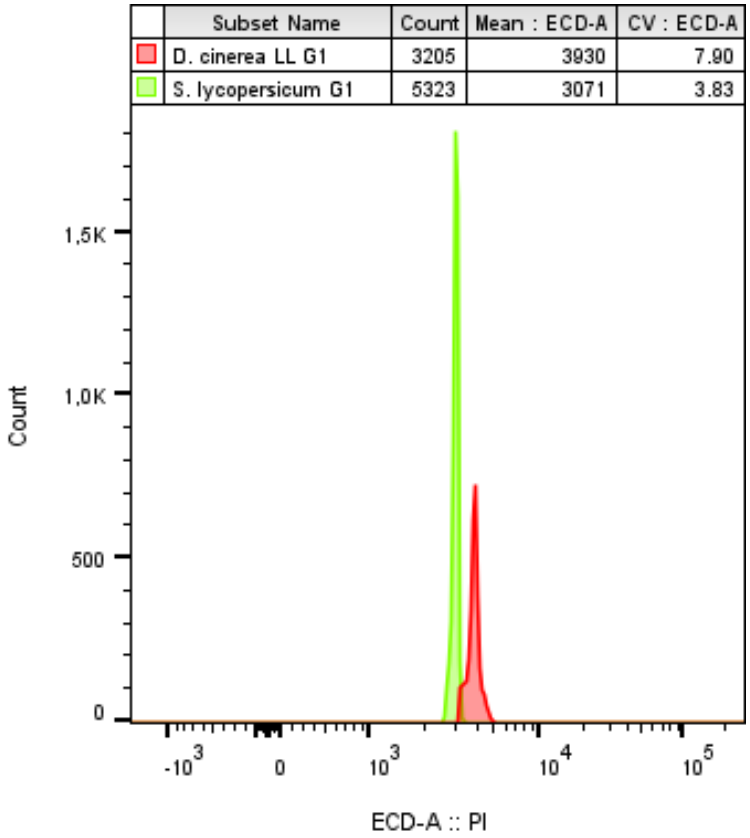

Figure 26. Frequency distribution of PI intensity values (x-axis), showing the external standardization of D. cinerea LL nuclei (red) with S. lycopersicum (green).

As the *D. cinerea* SL plant was predicted to be a diploid based on leaf size, the first experiment performed to estimate its genome size used *S. lycopersicum* as an internal standard. In contrast to the results obtained with *D. cinerea* LL (Figure 25), two distinct G1 peaks were observed (Figure 27), with mean PI intensities of 1627 (CV = 7.61) and 3001 (CV = 5.39) (Figure 27). As the peak at 3011 is almost identical in size to that obtained in the previous experiment for *S. lycopersicum* (Figure 26), the smaller G1 peak was attributed to *D. cinerea* SL, resulting in an estimated 2C-DNA content of 1.06 pg (1.04 Gbp), which is approximately 43% of the DNA content estimated for *D. cinerea* LL.

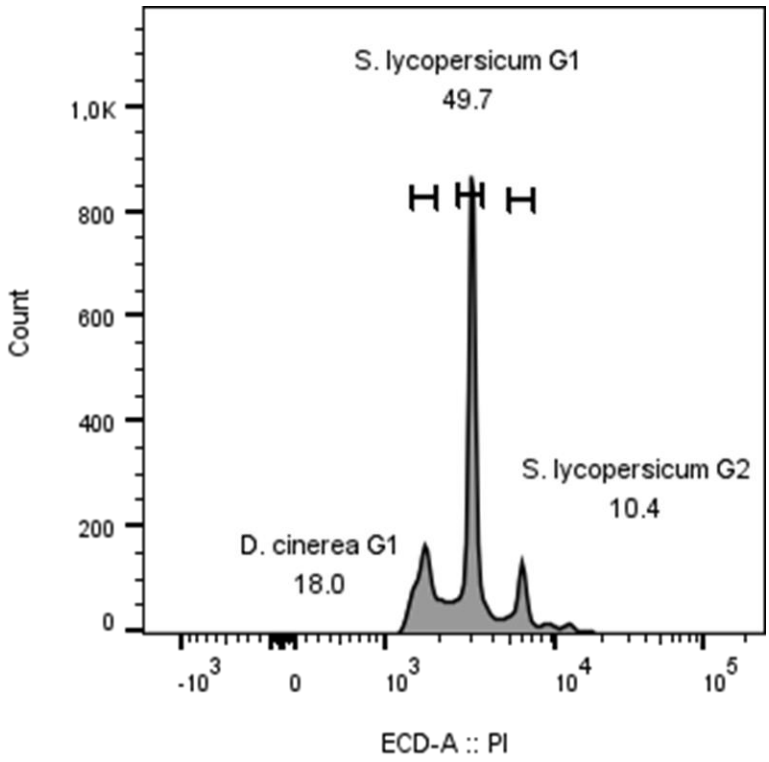

Figure 27. Frequency distribution of PI intensity, showing the internal standardization of D. cinerea SL nuclei with S. lycopersicum (Tomato).

Having established that the *D. cinerea* SL and LL plants display an approximately two-fold difference in estimated genome size, the final set of flow cytometry experiments was performed on dry leaf material from a subset of the samples used for ddRAD-seq. The G1 PI mean intensities of *D. cinerea* SL and *D. cinerea* LL were compared to a set of *in silico* predicted diploids and tetraploids respectively, in order to determine if the genome sizes were similar. Flow cytometry analysis of PI-stained nuclei extracted from three suspected diploid KNP individuals (K1, K47, K57) showed sharp G1 PI peaks for each sample, confirming that nuclei could still be extracted from dry leaflet *D. cinerea* tissue (Figure 28A). The G1 PI peaks overlapped strongly with each other and with the G1 peak of *D. cinerea* SL (Figure 28A), with an average PI intensity of 1663±17.83 across the four samples. Similarly, flow cytometry analysis of PI-stained nuclei extracted from suspected tetraploid individuals (K90, T5 and M30) revealed sharp G1 PI peaks for each sample analysed (Figure 28B) as well as a considerable overlap between resulting G1 peaks of dry leaflet samples and the fresh tissue G1 peak obtained from *D. cinerea* LL (Figure 28B). Notably the average PI intensity obtained from these four samples (3520.25±175.59) was approximately twice that obtained from the diploid samples (Figure 29). The flow cytometry analyses thus confirmed the existence of both diploid and tetraploid *D. cinerea* plants and validated the *in silico* determination of ploidy.

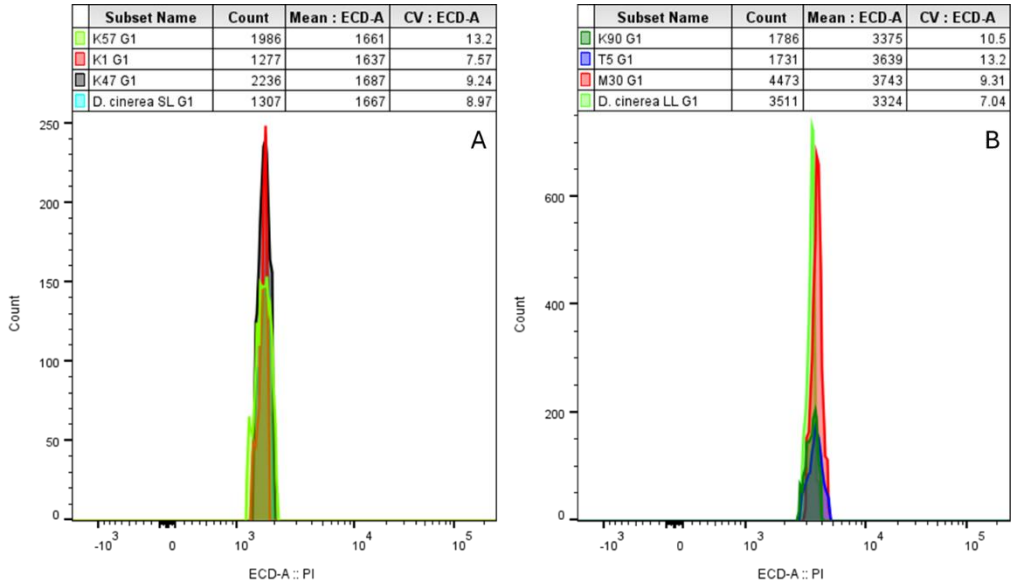

Figure 28. Frequency distributions of the PI intensity of extracted nuclei. Overlay image of the G1 peak of D. cinerea SL overlapped with samples predicted to be diploid via in silico ploidy analysis (A). Overlay image of the G1 peak of D. cinerea LL overlapped with the G1 peaks of expected tetraploids (B).

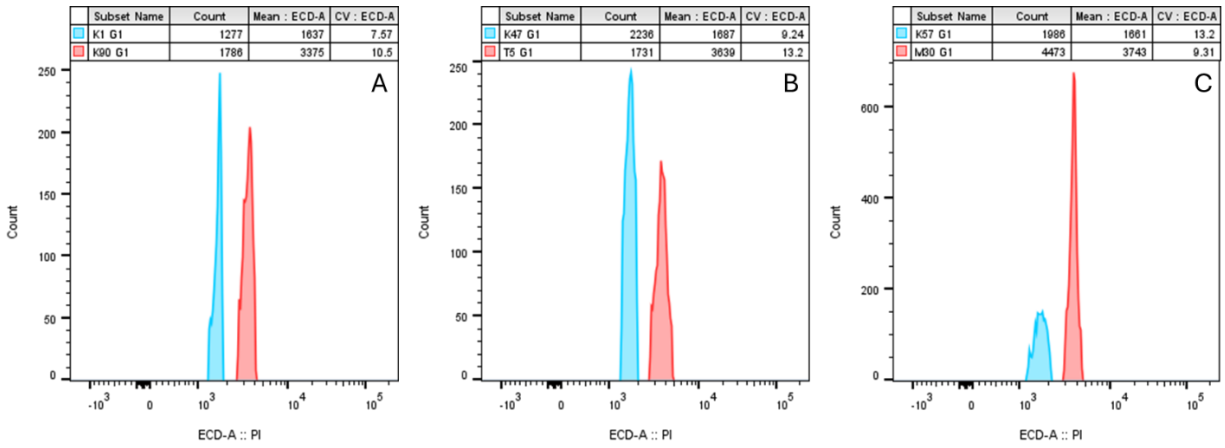

Figure 29. Comparison of the frequency distributions of PI intensity, displaying the G1 peaks of diploid individuals versus tetraploid individuals, namely K1 vs K90 (A), K47 vs T5 (B) and K57 vs M30 (C).

Clonality assessment in KNP diploids

Following the confirmation of intraspecific ploidy variation within the sampled *D. cinerea* population, it was possible to assess the genetic similarity between diploid individuals in KNP to identify potential clonal individuals within the population. In order to achieve this, a threshold value based on the maximum genetic distance achievable between known clonal individuals, i.e. sequencing error rate between the technical replicates, would be used to differentiate naturally occurring clonal individuals from non-clonal individuals, an approach which is similar to that has previously been used in other clonality studies [53, 102, 103]. This

analysis was restricted to the KNP diploids as technical replicates were only taken from this group of plants.

Identity-by-state (IBS) analysis was first performed on all diploid individuals from KNP, including the technical replicates, in order to assess relatedness between individuals as this would allow for a threshold value to be established at which clonal individuals could be distinguished from non-clonal individuals. This was followed by hierarchical clustering to generate a dendrogram for comparison. To identify clonal individuals within the population, the maximum genetic distance observed between any two technical replicates (Figure 30, solid line) was used as a threshold value, as this revealed the maximum genetic distance (i.e. sequencing error) possible between clonal samples. Analysis revealed that all technical replicates included in the analysis displayed a high and similar degree of genetic relatedness (Figure 30).

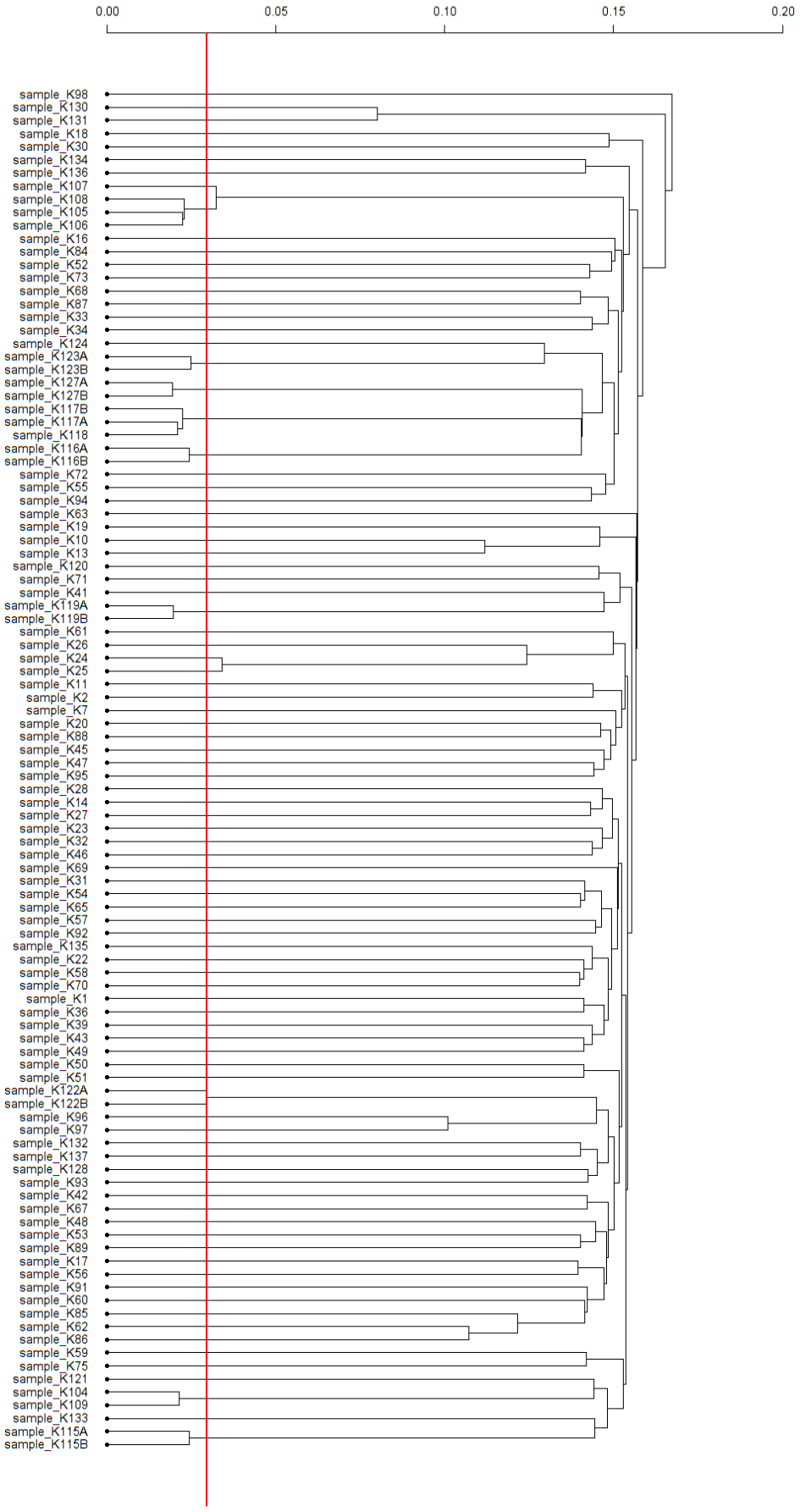
0.00
0.05
0.10
0.15
0.20
sample_K98
sample_K130
sample_K131
sample_K18
sample_K30
sample_K134
sample_K136
sample_K107
sample_K108
sample_K105
sample_K106
sample_K16
sample_K84
sample_K52
sample_K73
sample_K68
sample_K87
sample_K33
sample_K34
sample_K124
sample_K123A
sample_K123B
sample_K127A
sample_K127B
sample_K117B
sample_K117A
sample_K118
sample_K116A
sample_K116B
sample_K72
sample_K55
sample_K94
sample_K63
sample_K19
sample_K10
sample_K13
sample_K120
sample_K71
sample_K41
sample_K119A
sample_K119B
sample_K61
sample_K26
sample_K24
sample_K25
sample_K11
sample_K2
sample_K7
sample_K20
sample_K88
sample_K45
sample_K47
sample_K95
sample_K28
sample_K14
sample_K27
sample_K23
sample_K32
sample_K46
sample_K69
sample_K31
sample_K54
sample_K65
sample_K57
sample_K92
sample_K135
sample_K22
sample_K58
sample_K70
sample_K1
sample_K36
sample_K39
sample_K43
sample_K49
sample_K50
sample_K51
sample_K122A
sample_K122B
sample_K96
sample_K97
sample_K132
sample_K137
sample_K128
sample_K93
sample_K42
sample_K67
sample_K48
sample_K53
sample_K89
sample_K17
sample_K56
sample_K91
sample_K60
sample_K85
sample_K62
sample_K86
sample_K59
sample_K75
sample_K121
sample_K104
sample_K109
sample_K133
sample_K115A
sample_K115B

Figure 30. A dendrogram of the diploid D. cinerea plants sampled in KNP based on the genetic distance (IBS score) between individuals. The threshold value (solid red line) reflects the maximum amount of genetic distance observed between any two technical replicates due to sequencing error.

Following the establishment of a threshold value (henceforth referred to as the 'max threshold') which would act as a guideline for the assessment of clonality, technical replicates were subsequently removed, and IBS analysis was again performed on the remaining samples, followed by hierarchical clustering (Figure 31). When considering the max threshold three likely clonal pairs/groups were identified. Samples K104 and K109 were found to be clonal as they shared a high degree of genetic similarity, which fell below the max threshold value (Figure 31), as did samples 117 and 118 which were adjacent to each other in the field. Finally, a cluster of clonal individuals was identified which included samples K105, K106, and K108 as all of these samples displayed IBS values well below the max threshold value (Figure 31). Additionally, sample K107 was found to be closely related to members of this cluster, with a degree of genetic similarity just below the max threshold value which suggested that it is a potential clone (Figure 31). Similarly, samples K24 and K25 displayed a high degree of genetic similarity, just above the max threshold, and so also potentially represent clonal individuals.

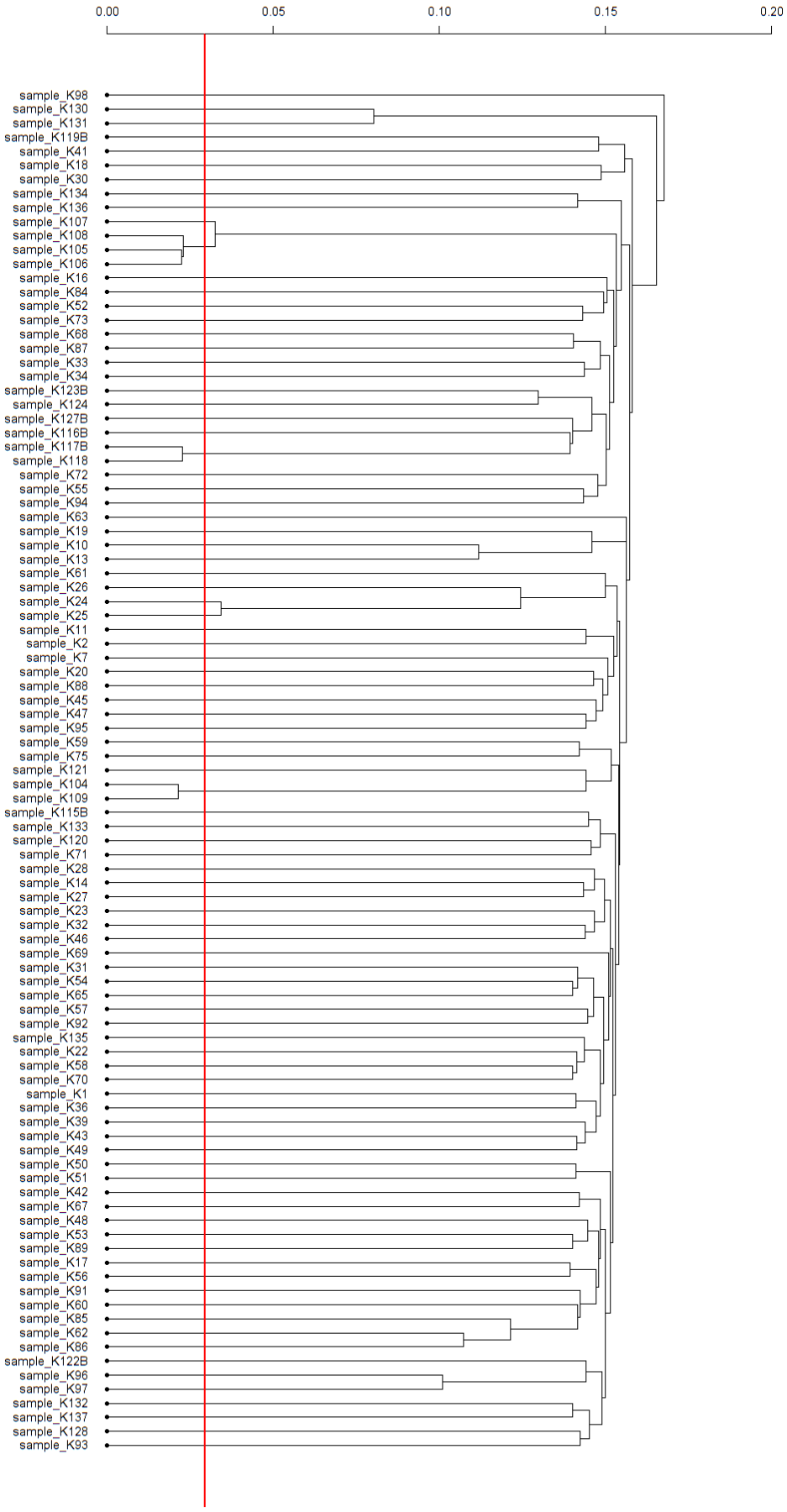
0.00
0.05
0.10
0.15
0.20
sample_K98
sample_K130
sample_K131
sample_K119B
sample_K41
sample_K18
sample_K30
sample_K134
sample_K136
sample_K107
sample_K108
sample_K105
sample_K106
sample_K16
sample_K84
sample_K52
sample_K73
sample_K68
sample_K87
sample_K33
sample_K34
sample_K123B
sample_K124
sample_K127B
sample_K116B
sample_K117B
sample_K118
sample_K72
sample_K55
sample_K94
sample_K63
sample_K19
sample_K10
sample_K13
sample_K61
sample_K26
sample_K24
sample_K25
sample_K11
sample_K2
sample_K7
sample_K20
sample_K88
sample_K45
sample_K47
sample_K95
sample_K59
sample_K75
sample_K121
sample_K104
sample_K109
sample_K115B
sample_K133
sample_K120
sample_K71
sample_K28
sample_K14
sample_K27
sample_K23
sample_K32
sample_K46
sample_K69
sample_K31
sample_K54
sample_K65
sample_K57
sample_K92
sample_K135
sample_K22
sample_K58
sample_K70
sample_K1
sample_K36
sample_K39
sample_K43
sample_K49
sample_K50
sample_K51
sample_K42
sample_K67
sample_K48
sample_K53
sample_K89
sample_K17
sample_K56
sample_K91
sample_K60
sample_K85
sample_K62
sample_K86
sample_K122B
sample_K96
sample_K97
sample_K132
sample_K137
sample_K128
sample_K93

Figure 31. A dendrogram of the diploid D. cinerea plants sampled in KNP based on the genetic distance (IBS score) between individuals, with all technical replicates removed. The threshold value (solid red line) reflects the maximum amount of genetic distance observed between any two technical replicates due to sequencing error.

Estimation of heterozygosity and inbreeding coefficient

Although the assessment for clonality within the *D. cinerea* KNP diploid population successfully identified several individuals which exhibited a similar degree of genetic identity to known clonal samples (technical replicates), it is also conceivable that such individuals could arise via self-fertilisation of highly inbred (homozygous) plants. To exclude this possibility, diversity and inbreeding statistics were calculated for both the KNP diploid population SNP data in STACKS populations program [44]. Based on 39 299 SNPs derived from 98 samples, with no missing data present, the observed heterozygosity was reported as 0.236, while the expected heterozygosity was reported as 0.162. Observed homozygosity was reported as 0.782, with expected homozygosity amounting to 0.838. Nucleotide diversity was found to be 0.163. The Fis value of -0.108 is not consistent with a high degree of inbreeding in the KNP diploids (Table 6), instead indicating a slight excess of heterozygotes in the population. Additionally, the assessment of self-pollination revealed that none of the 15 inflorescences bagged with pollinator exclusion bags were able to produce seed pods, suggesting that the species is not capable of self-fertilisation. Furthermore, there was an increase in both the expected heterozygosity and observed heterozygosity reported for tetraploid individuals in comparison to diploid individuals from within the KNP population, though these values must be interpreted with caution due to inability to distinguish between homologous and homeologous SNPs.

Table 6. Population diversity statistics for all sampled populations based on 39 299 SNPs: N = number of samples in population; private = number of private alleles of a population; Obs_Het = observed heterozygosity; Obs_Ho = observed homozygosity; Exp_Het = expected heterozygosity; Exp_Ho = expected homozygosity; Pi = nucleotide diversity; Fis = inbreeding coefficient.

Pop ID	Private	Nu	Obs_Het	Exp Het	Pi	Fis
KNP diploids	1001	98	0.23558	0.17293	0.17382	-0.14171
KNP tetraploids	12357	39	0.42344	0.29861	0.30249	-0.28834

Discussion

D. cinerea is one of several woody plant species that pose a major threat to natural savanna ecosystems in Southern Africa, by contributing to the rapid woody plant encroachment (WPE) of these biomes. *D. cinerea* in particular is considered to be a major driver of this process across South Africa [24]. Several removal efforts have been made to control the spread of *D. cinerea*, including in KNP, however there is no clear consensus on how to effectively reduce the encroachment of this species in grassland habitats. It is known that woody plant encroachment is affected by several factors, such as the degree of herbivory, fire frequency and rainfall, however it is unclear what the relative contributions of sexual and asexual reproduction are to this process. This is mostly due to the fact that the previously employed methods used to assess the extent of clonality in *D. cinerea*, and other woody plant species, are highly laborious and time consuming [4]. The rapid advancement of NGS technologies in the past decade have led to the advent of several high throughput marker discovery techniques, such as ddRAD-seq [43], which allow for the generation of hundreds to thousands of SNP markers for studies to assess the relatedness between individuals of a population to assess population structure. This study sought to design a workflow which would allow for the detection of clonality within *D. cinerea* to facilitate the investigation on the effects of clonality on WPE.

Population structure does not correlate completely with geographic location

To assess genetic diversity across a large part of range of *D. cinerea* in South Africa, three major sites were sampled. In the Mpumalanga province, samples were collected across the southern area of the Kruger National Park (KNP), spanning an area of over 500 km^2. In the Kwazulu-Natal province, the samples were collected from two sites, namely from within the St. Lucia area (KZN_ST) and the Manyoni Private Game Reserve (KZN_MA). Finally, samples were also collected from the Rustenburg area located in the North West province (Figure 5). Identity-by-state (IBS) analysis performed on the SNPs generated by ddRAD-seq revealed that the plants sampled did not cluster entirely on the basis of geographical location as might have been expected (Figure 15). Samples originating from the NW, KZN_MA and KZN_ST collection sites did cluster together, which indicated that these individuals appeared to be more closely related to individuals from within the same geographical area rather than individuals from a different collection site. However, 17 samples originating from KNP appeared to be more closely related to individuals from the KZN_MA population, while an additional 28 samples originating from KNP appeared to be more closely related to individuals from the KZN_ST population (Figure 15), despite the KNP collection area being located over 300 km away from the two sites in KZN (Figure 5).

This finding contradicts observations that have been reported in previous studies. A study on sixteen *Liriodendron* populations, containing *L. tulipifera* and *L. chinense*, found across the Jiangxi Province in China found a clear genetic divergence between *L. chinense* populations originating from the eastern and western parts of the province, which were separated by geographical barriers in the form of large plains and differences in altitude. Phylogenetic

analysis revealed that all sixteen populations clustered into three major clades, with *L. chinense* populations further separating into two distinct clades consisting of either the eastern or western populations [104].

Within the South African range of *D. cinerea*, four subspecies have previously been described on the basis of morphological variations that exist between individuals across the area [24]. Several taxonomic variants have also been proposed, however, there is disagreement on the validity of these taxa in South Africa (Ross 1974) and no genetic evidence has been generated to support their existence. Additionally, the classification of wild *D. cinerea* individuals into the various taxonomic variants of these subspecies requires careful consideration of multiple morphological characteristics [3], which was not possible here as the majority of individuals sampled were not flowering. However, leaflet material did provide an opportunity to assess the distribution of leaflet size within the sampling population, as one of the main morphological characteristics used to assign an individual to the subspecies *nyassana* is whether the fresh leaflet width is greater than 2 mm. In this study, leaflet measurements were made on dry material, which is preferable as the water content of the leaves varies over the course of the day.

Assessment of the leaflet measurements revealed a bimodal distribution of both leaflet length and width, which suggested that there appeared to be two groups of individuals which differed in their average leaflet size (Figure 16). This difference in leaflet size within a population of the same species has been previously observed in other studies, and can result from a change in ploidy, a phenomenon often referred to as the "Gigas effect". For example, a study in *Buddleja macrostachya* revealed a significant difference in absolute leaf size, fruit length and flower length between hexaploidy and dodecaploid individuals, suggesting that that some morphological traits tend to increase in size as a result of an increase in ploidy and that these traits could serve as indicators of ploidy levels [59]. Another study by Souza et al. (2023) sought to confirm chromosome duplication in artificially polyploidized clones of *Eucalyptus grandis* x *E. urophylla*, which could potentially increase the performance of these clones in an industrial context, by inducing the Gigas effect in these individuals. This study reported a significant increase in leaf area in polyploid individuals in comparison to diploid individuals, reporting a 46% increase in leaf area in polyploidized clones [105].

D. cinerea exhibits intraspecific variation in ploidy

To determine whether the failure of the KNP samples to cluster by geographical location (Figure 15) was due to differences in ploidy among the individuals sampled, *in silico* ploidy estimation was performed on each individual sample using the ddRAD-seq derived SNPs. This revealed that *D. cinerea* in KNP consisted of a mixed ploidy population, made up of diploid and tetraploid individuals. This distinction was made on the basis of the MAF distribution around modes expected either diploid or tetraploid individuals i.e. a distribution of MAF around 0.5 for diploid individuals and a distribution of MAF around 0.25 for tetraploid individuals (Figure 19). This approach has also been used in previous studies. For example. while investigating the degree of clonality in *S. uniplumis* and *S. ciliata* grasses in fairy circles, Kappel et al (2020) used allele frequency distributions of SNPs derived from ddRAD-seq data to assess the ploidy of samples within the sampled circles. This confirmed that the grasses

surrounding these fairy circles exhibited intraspecific variation in ploidy [52]. PLOIDYFROST, a reference-free ploidy estimation tool also makes use of this concept in its implementation. The pipeline consists of several steps, which ultimately result in the depiction of allele frequency distributions from input data for which the average log-likelihood is calculated to assess which ploidy level is best fit for the data [86].

In the present study, of the 225 samples collected (including technical replicates), 106 samples were determined to be diploid while the remaining 119 samples were found to be tetraploid. It was also observed that all the samples collected from the NW, KZN_ST and KZN_MA collection sites were apparent tetraploids. Interestingly, only within the KNP were a mixture of diploid and tetraploid individuals detected.

Flow cytometry confirms *in silico* ploidy estimations.

A possible limitation of the approach used for *in silico* estimation of sample ploidy may be that the assignment of ploidy was performed via visual interpretation of the allele frequencies, which may leave room for erroneous interpretation of allele frequency plots where the distribution of alleles around a mode is not clear. It was therefore necessary to provide additional evidence in order to validate *in silico* ploidy estimates via flow cytometry (FCM) analysis of the available dry leaflet tissue. To achieve this, suspected diploid and tetraploid individuals form each population were selected for genome size comparison, via FCM, in order to infer ploidy. While ploidy was not directly measured in this experiment, the expectation was that the tetraploid individuals should have DNA contents twice that of the diploids.

FCM experiments were initially conducted on *D. cinerea* fresh samples that were putatively identified as diploid (small leaf, SL) or tetraploid (large leaf, LL) based on leaflet measurements. FCM analysis of *D. cinerea* LL plant gave a genome size estimate of 2.45 Gbp (Figure 26), while *D. cinerea* SL plant had a genome size less than twice as large at 1.04 Gbp (Figure 27). This is consistent with a whole genome duplication in the LL plant which has been a frequent occurrence in the history of plant genomes [106]. This conclusion is further supported by the fact that the genome size estimate of 1.04 Gbp for the SL plant was similar to the 0.9 Gbp estimate from GenomeScope analysis of K47, (Figure 11), an individual determined to be diploid via *in silico* analysis (Figure 19B). Together this suggested that the *D. cinerea* SL variant was a diploid individual as the genome sizes were similar, while the LL plant would be tetraploid.

Additional flow cytometry experiments were then performed on dry leaf material from a subset of the samples used for ddRAD-seq. This was done in order to compare the G1 PI mean intensities of *D. cinerea* SL and *D. cinerea* LL to a set of *in-silico* predicted diploids and tetraploids respectively. For the both the *D. cinerea* SL and *D. cinerea* LL variants, this analysis revealed that the G1 peaks of the fresh tissue samples overlapped almost entirely with those from the dry leaflet tissue of *in silico* diploid and tetraploid samples respectively (Figure 28). This method took an approach similar to the guidelines proposed by Sliwinska et al (2022) [78]. These guidelines suggest that while external standardization can be used for ploidy estimation, a plant of the same species with a known ploidy level should be used as a standard. Similarly, comparison of the G1 PI peaks from fresh tissue to the resulting G1 PI peaks of dry

leaflet tissue of known DNA-ploidy indicate that the samples have the same genome size and by extension the same ploidy. It should however be noted that the assessment of DNA-ploidy via FCM is an indirect method of inference, and should be complimented by chromosome counting [78].

Intraspecific variation in ploidy complicates analysis of genetic relatedness

The genetic comparison of individuals that differ in ploidy can be challenging. This is due to the fact that polyploid individuals contain a larger number of chromosome copies than diploid individuals, which could lead to a higher mutation rate in polyploid individuals in comparison to diploid individuals. Additionally, the effects of genetic drift are also reduced in polyploid populations in comparison to a diploid population. This combination of effects can thus result in a higher level of genetic diversity on polyploid populations [107]. For example, in their assessment of clonality in *Posidonia australis*, Edgeloe et al. (2022) reported an increase in observed heterozygosity and nucleotide diversity in polyploid populations in comparison to the single diploid population found throughout the study [53].

The comparison between diploids and allopolyploids via the use of SNP data is also complicated by the difficulties in distinguishing between homeologous SNPs (polymorphic sites occurring across subgenomes within and between individuals) from allelic SNPs (polymorphic positions occurring within a single subgenome between individuals) [108, 109]. However, some bioinformatic approaches have previously attempted to assess the relationships between diploid and polyploid species using RAD-seq data. A study by Wagner et al. (2020) sought to elucidate the evolutionary history of different species within the genus *Salix*, a species known to have a range of ploidies (2n to 8n), as well as to assess whether RAD-seq data could be reliable in the phylogenetic assessment of species across different ploidies [110]. In order to elucidate the phylogenetic relationship between 35 *Salix* species of mixed ploidies the study utilized ipyrad, a toolkit oriented around the assembly and analysis of data generated from RAD-seq based techniques [111]. An advantage of this software was that it allows for the generation of consensus sequences during the assembly process, which includes variant sites represented by IUPAC ambiguity codes. This is particularly useful when assessing sequences with more than two alleles, a problem often presented by polyploid individuals granted the increased allele number of polyploids in comparison to diploids. This approach allowed for the study to resolve the phylogenetic relationships in Salix to a high degree of accuracy [110].

Additionally, the use of SNPs derived from high throughput genotyping methods, such as RAD-seq, in polyploids is largely impeded by the lack of both suitable bioinformatic tools and statistical analyses which are able to cater for more than two alleles [108]. While Wagner et al. (2020) did not generate a novel tool to resolve this issue, the study did contribute possible solutions by presenting a workflow which aimed to adapt the existing SNiPloid pipeline for the categorization of SNPs presented in the study [110]. SNiPloid is a tool used to infer the SNP genome origin, while also distinguishing homoeologous SNPs from interspecific SNPs in allopolyploids, however, this tool utilizes RNA data as input in order to perform the analysis

[109]. In their assessment of SNP composition of tetraploid species in the dataset, Wagner et al. (2020) adapted the SNiPloid pipeline to allow for analysis of RAD-seq data by making use of the concatenated RAD loci of a putative diploid parent species to act as a "pseudoreference" on to which the tetraploid species of interest and the second putative diploid parent species would be mapped. Furthermore, in order to elucidate the genetic structure of each sample, the allelic information of polyploid species would be reduced to pseudo-diploid consensus sequences. This approach allowed for both diploid and polyploids to be included in the same STRUCTURE analysis, setting a framework which could potentially be utilized by other similar studies [110].

Other polyploid oriented genotyping software has also been developed. For example, the comprehensive allopolyploid genotyper (CAPG) tool allows for a probabilistic approach to SNP calling in allotetraploids. CAPG utilizes whole genome sequencing data (WGS) to aligns reads to subgenomic references, whereafter CAPG then identifies the major and minor allele for each site and calculates possibilities for potential genotypes [112]. While it has been shown to outperform popular genotyping tools such as GATK, it assumed reference genomes are available for the species of interest, a large limitation when considering that majority of the biologically relevant species outside of crops are non-model and do not have WGS data available. It is also for this reason that CAPG was not used in this study, as no reference genome is available for *D. cinerea*.

With the advent of high throughput sequencing technologies, long read sequencing techniques, such as Oxford nanopore and PacBio, have shown a rise in popularity as they are well suited to their application in the generation of phased plant genomes [113]. Phased genome assemblies differ from commonly used haploid reference genomes by providing a distinction between paternal and maternal haplotypes. In polyploid plant genomes, accurate SNP detection is challenging as the use of haploid genome assemblies can result in the ambiguous mapping of homoeologous loci, leading to a high false positive rate in SNP detection [114]. The use of a phased genome during SNP calling is advantageous as it allows for the distinction between true SNPs and SNPs that are only present on a single subgenome, allowing studies to filter these homoeologous SNPs for downstream analysis [115]. Furthermore, the use of phasing in plant genomes also harbours the potential to construct accurate phylogenies and better inform the evolutionary history of a polyploid species. For example, Eriksson et al (2018) sought to assess whether allele phasing would help to elucidate the evolutionary origin of two polyploid species in the genus *Medicago*, which revealed that *M. arborea* and *M. strasseri* most likely arose from the hybridization of two parental species, allowing these species to be recognised a allotetraploids [116]. The application of a phased genome approach would greatly benefit future studies which aim to elucidate the phylogeny of *Dichrostachys* and whether *D. cinerea* arose due to a hybridization event or whether a genome duplication had occurred along the evolutionary history of this species, resulting in lineage(s) of this species being autotetraploid.

Given the dependence on software of most studies to calculate population statistics and generate phylogenies, careful consideration should be exercised as to which software is most suitable for both the data types and the ploidies of the organisms considered. While a large

number of software do exist for population genetic analysis, the majority of these software assume the data originated from a diploid organism [117]. However, in recent years there has been a rise in the implementation of polyploid oriented analysis software, such as the user friendly software Genodive [118]. As previously discussed, polyploid genomes are far more complex than diploid genomes due to the increased chromosome number, differences in genetic diversity and the presence of homoeologous SNPs present on only a single subgenome. As a result, this could lead to a high false positive rate in SNP identification in polyploids when a diploid genome is assumed by the genotyping software, as this software may only make use of a haploid reference genome which does not consider the assignment of SNPs to their respective subgenome [114]. This could lead to a high number of false SNPs being considered during downstream phylogenetic analysis. As these SNPs may not accurately depict the true natural variation that occurs between these species or individuals, the resulting phylogeny may also not provide an accurate depiction of the evolutionary history of the species of interest. Prior studies have also struggled with these issues and failed to implement software that was more suitable for polyploid individuals. For example, in their assessment of the population structure of mixed cytotype species *P. australis*, Edgeloe et al. (2022) utilize the STACKS software to perform *de novo* genome assembly and SNP calling on the available ddRAD-seq data for all samples [53]. As the STACKS software assumed diploidy [44], no assessment had been made as to whether the resulting SNP-set included a high degree of false SNPs which may have been homoeologous, which could bring into question how accurate the phylogeny presented by the study is.

In addition to difficulties in ascertaining the nature of SNPs (allelic versus homeologous) in *D. cinerea*, the subsequent clustering of individuals during phylogenetic analysis may also be affected by the presence of mixed cytotypes within a sampled population. It is known that within mixed cytotype populations, individuals of higher ploidy levels which descended from a single autopolyploidization event tend to cluster together, regardless of their population of origin [119], similar to what is seen with the KNP tetraploid individuals in this study (Figure 21). This provides some preliminary evidence that the tetraploid variant observed across the species range may have arisen due to a genome duplication event at some point along the species' evolutionary timeline, as opposed to a hybridization event between two parent species. However, the inference of whether the observed polyploids were autopolyploid or allopolyploid was outside the scope of this study, and the data presented here only provide evidence for speculation. Future studies would need to design an experiment with a similar approach to what has been used by Wagner et al. (2020), who utilized software such as HyDe and SNiPloid in order to assess whether individuals originated from a hybridization event [110].

ddRAD-seq derived SNP markers allow the detection of clonality in *D. cinerea*

Due to the issues associated with determining whether SNPs called in tetraploids map to homologous or homeologous chromosomes, and the lack of technical replicates necessary to determine the background sequencing error rate from sites other than the KNP the analysis performed to detect potential clones was restricted to the KNP diploids. In order to distinguish

between clonal and non-clonal individuals, an approach that has successfully been used in other studies to assess clonality was utilized. This entailed the determination of the maximum amount of genetic distance present between any pair of technical repeats as this would provide an estimate of the apparent baseline genetic distance that arises due to somatic mutations and sequencing error.

This approach has been widely utilized in studies which sought to detect clonality using high throughput SNP data. For example, in order to identify naturally occurring clonal individuals in *Montastraea cavernosa* using SNP markers generated via the 2bRAD approach, the study calculated IBS levels for technical replicates collected in-field to determine the threshold at which naturally occurring clones are expected to occur. This allowed for the detection of two naturally occurring clonal individuals within the dataset [102]. Another study made use of this approach in order to assess for clonality in *Orbicella faveolata* where the IBS distance observed in technical replicates was used as a threshold value in order to differentiate between clonal and non-clonal individuals. This again allowed researchers to identify several clonal colonies within the sampling site [103]. Similar to these studies, the use of an IBS derived threshold value on the basis of the maximal genetic distance expected to be present between technical replicates revealed three groups of clonal individuals (present in two of the three transects performed), two pairs and one group of three (Figure 31), plus three additional potential clones (an additional member, 107, of the group formed of 105, 106 and 108, plus another pair of individuals, 104 and 109).

Furthermore, assessment of the genetic distances present between technical replicates showed that these were highly similar across all technical replicates (Figure 30), confirming that there was a low and consistent degree of genetic distance between a sample and its technical replicate due to sequencing error and somatic mutations. This also suggests that the threshold value provided to allow for the detection of clonality was a robust measure of sequencing error/somatic mutations, as the high degree of genetic similarity between clonal samples and technical replicate samples was clearly and consistently distinguishable from the rest of the sample population.

Previous findings made my Wakeling & Bond (2007) observed a mean proportion of 0.55 root suckers per site across 11 sites from a total of 370 individuals [4]. However, it is important to note that in this study the samples were localised in 25m x 4m plots and so are obviously not comparable to the sampling undertaken here. However, given the results obtained here revealed that clonal individuals can be detected, it is proposed that the use of ddRAD-seq to detect clonality in *D. cinerea* could be utilized in a similar study to Wakeling & Bond (2007), in order to provide a more accurate and less labour-intensive approach. This genetic approach could avoid certain pitfalls which were preset in the study by Wakeling & Bond (2007), such as the loss of the connecting roots of plants of the same genet, as the leaflet material used in this study was shown to be a robust source of genetic information to allow for the detection of clonality. Additionally, the method of sampling in this study is much more efficient as it only took several days of sampling to procure over 200 *D. cinerea* samples.

While it was suggested by Wakeling & Bond (2007) that fire and herbivory do not seem to play a major role in clonality [4], it should be noted that in order to gain a clearer understanding of

how clonality plays a role in encroachment, environmental factors do need to be taken into account. However, since this the effects of environmental data were not considered in this study, this study cannot comment on whether clonality does play a role in the woody plant encroachment of South African savanna biomes by *D. cinerea*, but rather provides a framework in the detection of clonality in order to facilitate future studies which aim to take a more holistic approach in the assessment of whether clonality does play a role in woody plant encroachment.

Finally, a preliminary assessment of self-pollination in *D. cinerea* was conducted by isolating the resulting 15 inflorescences of individuals. This revealed that no seed pods had been generated as a result, providing preliminary evidence that *D. cinerea* is not capable of self-fertilisation. This is in agreement with the population statistics generated from the KNP diploid population, which indicated a higher level of observed heterozygosity that what was expected for the population, complimented by a low inbreeding coefficient (Table 6). Overall, this strongly suggests that the clonal individuals detected within the transects were indeed the result of asexual propagation rather than an individual which had propagated from selfing of a highly homozygous individual.

The association between leaflet size and ploidy may provide a basis for the delineation of *D. cinerea africana* and *nyassana* subspecies

To assess whether there was a difference between leaflet sizes of samples of a specific ploidy, leaflet measurements and *in silico* ploidy estimates were considered in a Mann-Whitney test. This revealed a significant difference in both leaflet width and leaflet length between diploid and tetraploid individuals, with a notable increase in leaflet size being observed for tetraploid individuals (Figure 22). This could indicate that the Gigas effect has taken place within *D. cinerea*. This hypothesis corresponds to similar findings that have been made by previous studies having also observed a notable increase in leaf size due to an increase in ploidy [59, 60, 105]. If this were the case, it would suggest that leaflet measurements may provide a rapid method for the assessment of ploidy within the species, which would help minimize the need for time consuming experimental procedures in order to determine the ploidy of the organism in future studies. This would also allow for the selection of diploid individuals during the sample collection phase of future studies, saving researchers time and storage in the field, where these resources are often limiting factors.

With the data made available from this study, some insights can be provided into the taxonomical uncertainty of this species. During this study, *in silico* analysis of MAF distributions for all samples revealed that *D. cinerea* exists as a mixed cytotype species consisting of both diploid and tetraploid individuals in South Africa (Figure 19). It is also clear from the data of this study that there is a significant difference in average leaflet dimensions between diploid and tetraploid individuals (Figure 22). While there does appear to be a bimodal distribution of leaflet dimensions (Figure 16), there is still an overlap between these leaflet dimensions (Figure 22), suggesting that while there does appear to be two different groups of individuals which could be separated on the basis of leaflet size, there is still continuous variation observed between these groups. This is in agreement with observations

made by Ross (1974) which suggested that while clear leaflet differences could be observed between subsp. nyassana and other subspecies, there appears to be more or less continuous variation among the leaflet sizes of subspecies [3]. However, granted that there is an association between leaflet size and ploidy in *D. cinerea*, it can be speculated from this study that subsp. *nyassana* may have arisen from a whole genome duplication event, leading to the induction of the Gigas effect in leaflet size. This then suggests that in order to end the taxonomic debate on which subspecies within the South African range can be considered true, a future study considering a broad range of morphological traits, together with measures of ploidy and genetic information could be performed.

Limitations and Future experiments

As this study revealed that *D. cinerea* consisted of a mixed cytotype population across the species range, it was not possible to assess whether clonality could be detected in tetraploid individuals. This was due to the fact that no technical replicates were collected for tetraploid populations as the initial assumption during the tissue collection phase was that all *D. cinerea* consisted of the same ploidy. This was accompanied by the fact that this study was unable to identify potentially inaccurate homoeologous SNPs in tetraploid individuals as no whole genome information was available prior to this study. In hindsight, technical replicates should have been collected for all populations where sampling had occurred and is one of the leading limitations of this study, as majority of the samples collected could not be used to assess for clonality. This does however create opportunities for future studies to assess whether the detection of clonality is plausible within mixed cytotype populations.

Another limitation of this study was that the SNP genotyping software used, dDocent, is suitable for only diploid individuals [49]. It was initially chosen due to its highly user-friendly design and detailed online documentation, which was desirable granted the workflow was intended to be established for use in labs which did not necessarily have a strong bioinformatic background but were still interested in answering biologically relevant questions in the field of conservation biology. It has been retained as the software of choice for this study due to both time constraints on the project and the fact that the study question could still be addressed via the analysis of the diploid population. However, it is recommended that in future studies that aim to utilize the workflow presented here, another variant calling software be used and that more whole genome data be generated in order to allow for the generation of a phased genome.

Since this study sought to establish a framework for the detection of clonality in non-model woody plant species, biologically important questions which relate to clonality in these plant species can now be addressed using a genetic approach, such as whether clonality plays a major role in the success of woody plant encroachment within the Savanna. To my knowledge, very few studies have sought to address the question of whether clonality is a contributing factor as to why *D. cinerea* so rapidly encroaches on areas of the Savanna ecosystem. Several treatment methods have also been applied to *D. cinerea* in hopes of minimizing the spread of this species in areas where it encroaches on the natural biome [16, 25]. However, the question as to whether clonality may be the reason these treatments prove largely ineffective has not been answered.

References

1. Blaser, W.J., et al., *Woody encroachment reduces nutrient limitation and promotes soil carbon sequestration.* Ecology and Evolution, 2014. **4**(8): p. 1423-1438.
2. Utaile, Y.U., et al., *Effect of Dichrostachys cinerea encroachment on plant species diversity, functional traits and litter decomposition in an East-African savannah ecosystem.* Journal of Vegetation Science, 2021. **32**(1): p. e12949.
3. Ross, J., *A note on Dichrostachys cinerea in South Africa.* Bothalia, 1974. **11**(3): p. 265-268.
4. Wakeling, J. and W. Bond, *Disturbance and the frequency of root suckering in an invasive savanna shrub, Dichrostachys cinerea.* African Journal of Range and Forage Science, 2007. **24**(2): p. 73-76.
5. Cheek, M. *Dichrostachys cinerea (L.) Wight & Arn.* 2009; Available from: https://pza.sanbi.org/dichrostachys-cinerea.
6. Valero-Jorge, A., et al., *Mapping and Monitoring of the Invasive Species Dichrostachys cinerea (Marabú) in Central Cuba Using Landsat Imagery and Machine Learning (1994–2022).* Remote Sensing, 2024. **16**(5): p. 798.
7. Heuzé, V. *Sicklebush (Dichrostachys cinerea).* 2015 October 7, 2015; Available from: https://www.feedipedia.org/node/298.
8. Banso, A. and S. Adeyemo, *Evaluation of antibacterial properties of tannins isolated from Dichrostachys cinerea.* African Journal of Biotechnology, 2007. **6**(15).
9. Abou Zeid, A.H., M.S. Hifnawy, and R.S. Mohammed, *Phenolic compounds and biological activities of Dichrostachys cinerea L.* Med Aromat Plant Sci Biotechnol, 2009. **3**: p. 42-49.
10. Maphosa, V., J.L.N. Sikosana, and V. Muchenje, *Effect of doe milking and supplementation using Dichrostachys cinerea pods on kid and doe performance in grazing goats during the dry season.* Tropical Animal Health and Production, 2009. **41**(4): p. 535-541.
11. Yayneshet, T., L.O. Eik, and S.R. Moe, *Feeding Acacia etbaica and Dichrostachys cinerea fruits to smallholder goats in northern Ethiopia improves their performance during the dry season.* Livestock Science, 2008. **119**(1): p. 31-41.
12. Aganga, A. and C. Motshewa, *Nutritive value of urea molasses blocks containing Acacia Erubescens or Dichrostachys cineria as natural protein sources.* Journal of Animal and Veterinary Advances, 2007. **6**: p. 1280-1283.
13. Pedroso, D.T. and M. Kaltschmitt, *Dichrostachys cinerea as a possible energy crop—facts and figures.* Biomass Conversion and Biorefinery, 2012. **2**: p. 41-51.
14. Baskaran, P., et al., *Characterization of new natural cellulosic fiber from the bark of dichrostachys cinerea.* Journal of Natural Fibers, 2018. **15**(1): p. 62-68.
15. Brenan, J.P. and R.K. Brummitt, *The variation of Dichrostachys cinerea (L.) Wight & Arn.* 1965.
16. Utaile, Y.U., et al., *Assessing removal methods for controlling Dichrostachys cinerea encroachment and their impacts on plant communities in an East-African savannah ecosystem.* Applied Vegetation Science, 2023. **26**(1): p. e12720.
17. Roques, K., T. O'connor, and A.R. Watkinson, *Dynamics of shrub encroachment in an African savanna: relative influences of fire, herbivory, rainfall and density dependence.* Journal of Applied Ecology, 2001. **38**(2): p. 268-280.
18. *Most Common Savanna Plants*. Available from: https://myplantin.com/blog/savanna-plants.
19. Scholes, R.J. and S.R. Archer, *Tree-grass interactions in savannas.* Annual review of Ecology and Systematics, 1997. **28**(1): p. 517-544.
20. Venter, Z.S., M.D. Cramer, and H.J. Hawkins, *Drivers of woody plant encroachment over Africa.* Nature Communications, 2018. **9**(1): p. 2272.
21. Ward, D., et al., *Spatial analysis reveals facilitation in young clonal trees and competition in older trees during re-invasion of encroaching trees in an African savanna.* Plant Ecology, 2022. **223**(10): p. 1167-1180.

22. Bussa, B. and S. Shibru, *Effects of Sicklebush (Dichrostachys cinerea (l.) wight and arn. shrub) encroachment on floristic and vegetation structure in semi-arid savannah of southern Ethiopia.* Journal of Environment and Earth Science, 2020. **10**: p. 1-11.
23. Ward, D., M.T. Hoffman, and S.J. Collocott, *A century of woody plant encroachment in the dry Kimberley savanna of South Africa.* African Journal of Range & Forage Science, 2014. **31**(2): p. 107-121.
24. O'connor, T.G., J.R. Puttick, and M.T. Hoffman, *Bush encroachment in southern Africa: changes and causes.* African Journal of Range & Forage Science, 2014. **31**(2): p. 67-88.
25. Strydom, T., et al., *High-intensity fires may have limited medium-term effectiveness for reversing woody plant encroachment in an African savanna.* Journal of Applied Ecology, 2023. **60**(4): p. 661-672.
26. Reuter, Jason A., D.V. Spacek, and Michael P. Snyder, *High-Throughput Sequencing Technologies.* Molecular Cell, 2015. **58**(4): p. 586-597.
27. Slatko, B.E., A.F. Gardner, and F.M. Ausubel, *Overview of Next-Generation Sequencing Technologies.* Current Protocols in Molecular Biology, 2018. **122**(1): p. e59.
28. Imelfort, M. and D. Edwards, *De novo sequencing of plant genomes using second-generation technologies.* Briefings in bioinformatics, 2009. **10**(6): p. 609-618.
29. Vurture, G.W., et al., *GenomeScope: fast reference-free genome profiling from short reads.* Bioinformatics, 2017. **33**(14): p. 2202-2204.
30. van Lieshout, N., et al., *De novo whole-genome assembly of Chrysanthemum makinoi, a key wild chrysanthemum.* G3 Genes|Genomes|Genetics, 2021. **12**(1).
31. Al-Samarai, F.R. and A.A. Al-Kazaz, *Molecular markers: An introduction and applications.* European journal of molecular biotechnology, 2015. **9**(3): p. 118-130.
32. Botstein, D., et al., *Construction of a genetic linkage map in man using restriction fragment length polymorphisms.* Am J Hum Genet, 1980. **32**(3): p. 314-31.
33. Oshaghi, M.A., A.R. Chavshin, and H. Vatandoost, *Analysis of mosquito bloodmeals using RFLP markers.* Experimental Parasitology, 2006. **114**(4): p. 259-264.
34. Kalia, R.K., et al., *Microsatellite markers: an overview of the recent progress in plants.* Euphytica, 2011. **177**(3): p. 309-334.
35. Alves, S.I.A., et al., *What are microsatellites and how to choose the best tool: a user-friendly review of SSR and 74 SSR mining tools.* Frontiers in Genetics, 2024. **15**.
36. Abdurakhmonov, I.Y., *Introduction to microsatellites: basics, trends and highlights.* Microsatellite markers, 2016. **1**: p. 13.
37. Vignal, A., et al., *A review on SNP and other types of molecular markers and their use in animal genetics.* Genetics selection evolution, 2002. **34**(3): p. 275-305.
38. Davey, J.W., et al., *Genome-wide genetic marker discovery and genotyping using next-generation sequencing.* Nature Reviews Genetics, 2011. **12**(7): p. 499-510.
39. Baird, N.A., et al., *Rapid SNP discovery and genetic mapping using sequenced RAD markers.* PloS one, 2008. **3**(10): p. e3376.
40. Tsujimoto, M., et al., *Genet assignment and population structure analysis in a clonal forest-floor herb, Cardamine leucantha, using RAD-seq.* AoB PLANTS, 2019. **12**(1).
41. Wang, S., et al., *2b-RAD: a simple and flexible method for genome-wide genotyping.* Nature methods, 2012. **9**(8): p. 808-810.
42. Toonen, R.J., et al., *ezRAD: a simplified method for genomic genotyping in non-model organisms.* PeerJ, 2013. **1**: p. e203.
43. Peterson, B.K., et al., *Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species.* PloS one, 2012. **7**(5): p. e37135.
44. Catchen, J., et al., *Stacks: an analysis tool set for population genomics.* Mol Ecol, 2013. **22**(11): p. 3124-40.

45. Severn-Ellis, A.A., et al., *Genotyping for species identification and diversity assessment using double-digest restriction site-associated DNA sequencing (ddRAD-seq).* Legume genomics: Methods and protocols, 2020: p. 159-187.
46. Danecek, P., et al., *The variant call format and VCFtools.* Bioinformatics, 2011. **27**(15): p. 2156-8.
47. Danecek, P., et al., *Twelve years of SAMtools and BCFtools.* GigaScience, 2021. **10**(2).
48. Lu, F., et al., *Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol.* PLoS genetics, 2013. **9**(1): p. e1003215.
49. Puritz, J.B., C.M. Hollenbeck, and J.R. Gold, *dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms.* PeerJ, 2014. **2**: p. e431.
50. Villano, C., et al., *Genetic diversity and signature of divergence in the genome of grapevine clones of Southern Italy varieties.* Frontiers in Plant Science, 2023. **14**: p. 1201287.
51. Amor, M.D., J.C. Johnson, and E.A. James, *Identification of clonemates and genetic lineages using next-generation sequencing (ddRADseq) guides conservation of a rare species, Bossiaea vombata (Fabaceae).* Perspectives in plant ecology, evolution and systematics, 2020. **45**: p. 125544.
52. Kappel, C., et al., *Fairy circles in Namibia are assembled from genetically distinct grasses.* Communications Biology, 2020. **3**(1): p. 698.
53. Edgeloe, J.M., et al., *Extensive polyploid clonality was a successful strategy for seagrass to expand into a newly submerged environment.* Proceedings of the Royal Society B: Biological Sciences, 2022. **289**(1976): p. 20220538.
54. Mable, B.K. and S.P. Otto, *The evolution of life cycles with haploid and diploid phases.* BioEssays, 1998. **20**(6): p. 453-462.
55. Comai, L., *The advantages and disadvantages of being polyploid.* Nature Reviews Genetics, 2005. **6**(11): p. 836-846.
56. Kolář, F., et al., *Mixed-Ploidy Species: Progress and Opportunities in Polyploid Research.* Trends in Plant Science, 2017. **22**(12): p. 1041-1055.
57. Schneider, D.J., R.A. Levin, and J.S. Miller, *Reproductive isolation between diploid and tetraploid individuals in mixed-cytotype populations of Lycium australe.* American Journal of Botany, 2023. **110**(2): p. e16133.
58. Castañeda-Nava, J.J., et al., *EVALUATING THE CORRELATION OF PLOIDY LEVEL, LEAF SIZE, STOMATA CHARACTERISTICS AND TUBER WEIGHT IN Dioscorea spp. POPULATIONS FROM JALISCO, MÉXICO.* Tropical and Subtropical Agroecosystems, 2023. **26**(2).
59. CHEN, G., W.-B. SUN, and H. SUN, *Morphological characteristics of leaf epidermis and size variation of leaf, flower and fruit in different ploidy levels in Buddleja macrostachya (Buddlejaceae).* Journal of Systematics and Evolution, 2009. **47**(3): p. 231-236.
60. Zhang, Y., et al., *Ploidy and hybridity effects on leaf size, cell size and related genes expression in triploids, diploids and their parents in Populus.* Planta, 2019. **249**(3): p. 635-646.
61. Cires, E., et al., *Genome size variation and morphological differentiation within Ranunculus parnassifolius group (Ranunculaceae) from calcareous screes in the Northwest of Spain.* Plant Systematics and Evolution, 2009. **281**(1): p. 193-208.
62. Augusto Corrêa Dos Santos, R., G.H. Goldman, and D.M. Riaño-Pachón, *ploidyNGS: visually exploring ploidy with Next Generation Sequencing data.* Bioinformatics, 2017. **33**(16): p. 2575-2576.
63. Yoshida, K., et al., *The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine.* eLife, 2013. **2**: p. e00731.
64. McKinnon, K.M., *Flow Cytometry: An Overview.* Curr Protoc Immunol, 2018. **120**: p. 5.1.1-5.1.11.
65. Dolezel, J., J. Greilhuber, and J. Suda, *Flow cytometry with plant cells: analysis of genes, chromosomes and genomes*. 2007: John Wiley & Sons.

66. Telford, W.G., *Overview of Lasers for Flow Cytometry*, in *Flow Cytometry Protocols*, T.S. Hawley and R.G. Hawley, Editors. 2018, Springer New York: New York, NY. p. 447-479.
67. Jyoti, T.P., S. Chandel, and R. Singh, *Flow cytometry: Aspects and application in plant and biological science.* Journal of Biophotonics, 2024. **17**(3): p. e202300423.
68. Dolezel, J. and J. Bartos, *Plant DNA flow cytometry and estimation of nuclear genome size.* Ann Bot, 2005. **95**(1): p. 99-110.
69. Hagenbeek, D. and C.D. Rock, *Quantitative analysis by flow cytometry of abscisic acid-inducible gene expression in transiently transformed rice protoplasts.* Cytometry, 2001. **45**(3): p. 170-179.
70. Galbraith, D.W., *Analysis of Plant Gene Expression Using Flow Cytometry and Sorting*, in *Flow Cytometry with Plant Cells*. 2007. p. 405-422.
71. Doležel, J., et al., *Chromosome Analysis and Sorting*, in *Flow Cytometry with Plant Cells*. 2007. p. 373-403.
72. Doležel, J., et al., *Flow cytometric analysis and sorting of plant chromosomes.* The Nucleus, 2023. **66**(3): p. 355-369.
73. Kubaláková, M., et al., *Analysis and sorting of rye (Secale cereale L.) chromosomes using flow cytometry.* Genome, 2003. **46**(5): p. 893-905.
74. Heller, F., *DNA measurement of Vicia faba L. with pulse cytophotometry.* Ber Dtsch Bot Ges, 1973. **86**(5-9): p. 437-441.
75. Greilhuber, J., E.M. Temsch, and J.C.M. Loureiro, *Nuclear DNA Content Measurement*, in *Flow Cytometry with Plant Cells*. 2007. p. 67-101.
76. Suda, J. and P. Trávníček, *Reliable DNA ploidy determination in dehydrated tissues of vascular plants by DAPI flow cytometry—new prospects for plant research.* Cytometry Part A, 2006. **69**(4): p. 273-280.
77. Fomicheva, M. and E. Domblides, *Mastering DNA Content Estimation by Flow Cytometry as an Efficient Tool for Plant Breeding and Biodiversity Research.* Methods Protoc, 2023. **6**(1).
78. Sliwinska, E., et al., *Application-based guidelines for best practices in plant flow cytometry.* Cytometry Part A, 2022. **101**(9): p. 749-781.
79. Healey, A., et al., *Protocol: a simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species.* Plant methods, 2014. **10**(1): p. 1-8.
80. Ranallo-Benavidez, T.R., K.S. Jaron, and M.C. Schatz, *GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes.* Nature communications, 2020. **11**(1): p. 1432.
81. Andrews, S. *FastQC: A quality control tool for high throughput sequence data.* 2010; Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
82. Li, W. and A. Godzik, *Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.* Bioinformatics, 2006. **22**(13): p. 1658-1659.
83. Chong, Z., J. Ruan, and C.-I. Wu, *Rainbow: an integrated tool for efficient clustering and assembling RAD-seq reads.* Bioinformatics, 2012. **28**(21): p. 2732-2737.
84. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
85. Garrison, E. and G. Marth, *Haplotype-based variant detection from short-read sequencing.* arXiv preprint arXiv:1207.3907, 2012.
86. Sun, M., et al., *ploidyfrost: Reference-free estimation of ploidy level from whole genome sequencing data based on de Bruijn graphs.* Mol Ecol Resour, 2023. **23**(2): p. 499-510.
87. Zheng, X., et al., *A high-performance computing toolset for relatedness and principal component analysis of SNP data.* Bioinformatics, 2012. **28**(24): p. 3326-8.
88. Knaus, B.J. and N.J. Grünwald, *vcfr: a package to manipulate and visualize variant call format data in R.* Molecular Ecology Resources, 2017. **17**(1): p. 44-53.
89. Loureiro, J., et al., *Two new nuclear isolation buffers for plant DNA flow cytometry: a test with 37 species.* Annals of botany, 2007. **100**(4): p. 875-888.

90. Khan, S., et al., *Protocol for isolation of genomic DNA from dry and fresh roots of medicinal plants suitable for RAPD and restriction digestion.* African Journal of Biotechnology, 2007. **6**(3): p. 175.
91. Quiñones, K.J.O., et al., *Liquid-nitrogen-free CTAB DNA extraction method from silica-dried specimens for next-generation sequencing and assembly.* Methods X, 2024. **12**.
92. Wang, X.-D., Z.-P. Wang, and Y.-P. Zou, *An improved procedure for the isolation of nuclear DNA from leaves of wild grapevine dried with silica gel.* Plant Molecular Biology Reporter, 1996. **14**(4): p. 369-373.
93. Aboul-Maaty, N.A.-F. and H.A.-S. Oraby, *Extraction of high-quality genomic DNA from different plant orders applying a modified CTAB-based method.* Bulletin of the National Research Centre, 2019. **43**(1): p. 1-10.
94. Tibbits, J.F.G., et al., *A rapid method for tissue collection and high-throughput isolation of genomic DNA from mature trees.* Plant Molecular Biology Reporter, 2006. **24**(1): p. 81-91.
95. Choi, H.-K., et al., *Development of nuclear gene-derived molecular markers linked to legume genetic maps.* Molecular Genetics and Genomics, 2006. **276**(1): p. 56-70.
96. Wang, Y.-H., et al., *Plastomes of Mimosoideae: structural and size variation, sequence divergence, and phylogenetic implication.* Tree Genetics & Genomes, 2017. **13**(2): p. 41.
97. Desai, A., et al., *Identification of Optimum Sequencing Depth Especially for De Novo Genome Assembly of Small Genomes Using Next Generation Sequencing Data.* PLOS ONE, 2013. **8**(4): p. e60204.
98. Li, F.W. and A. Harkess, *A guide to sequence your favorite plant genomes.* Appl Plant Sci, 2018. **6**(3): p. e1030.
99. Galbraith, D.W., et al., *Rapid flow cytometric analysis of the cell cycle in intact plant tissues.* Science, 1983. **220**(4601): p. 1049-1051.
100. Pfosser, M., et al., *Evaluation of sensitivity of flow cytometry in detecting aneuploidy in wheat using disomic and ditelosomic wheat–rye addition lines.* Cytometry, 1995. **21**(4): p. 387-393.
101. LOUREIRO, J., et al., *Flow Cytometric and Microscopic Analysis of the Effect of Tannic Acid on Plant Nuclei and Estimation of DNA Content.* Annals of Botany, 2006. **98**(3): p. 515-527.
102. Sturm, A.B., et al., *Population genetic structure of the great star coral, Montastraea cavernosa, across the Cuban archipelago with comparisons between microsatellite and SNP markers.* Scientific Reports, 2020. **10**(1): p. 15432.
103. Manzello, D.P., et al., *Role of host genetics and heat-tolerant algal symbionts in sustaining populations of the endangered coral Orbicella faveolata in the Florida Keys with ocean warming.* Global Change Biology, 2019. **25**(3): p. 1016-1031.
104. Zhong, Y., et al., *RAD-Seq Data Point to a Distinct Split in Liriodendron (Magnoliaceae) and Obvious East–West Genetic Divergence in L. chinense.* Forests, 2019. **10**(1): p. 13.
105. Souza, T.d.S., et al., *Assessment of the cytogenetics and leaf anatomy of synthetic polyploids of Eucalyptus clones.* Crop Breeding and Applied Biotechnology, 2023. **23**(1): p. e43762316.
106. Thriveni, V., et al., *Impact of Polyploidy on Crop Improvement and Plant Breeding Strategies: A Review.* Journal of Advances in Biology & Biotechnology, 2024. **27**(11): p. 714-725.
107. Meirmans, P.G., S. Liu, and P.H. van Tienderen, *The analysis of polyploid genetic data.* Journal of Heredity, 2018. **109**(3): p. 283-296.
108. Clevenger, J., et al., *Single nucleotide polymorphism identification in polyploids: a review, example, and recommendations.* Molecular plant, 2015. **8**(6): p. 831-846.
109. Peralta, M., et al., *SNiPloid: A Utility to Exploit High-Throughput SNP Data Derived from RNA-Seq in Allopolyploid Species.* International Journal of Plant Genomics, 2013. **2013**(1): p. 890123.
110. Wagner, N.D., L. He, and E. Hörandl, *Phylogenomic relationships and evolution of polyploid Salix species revealed by RAD sequencing data.* Frontiers in Plant Science, 2020. **11**: p. 1077.
111. Eaton, D.A. and I. Overcast, *ipyrad: Interactive assembly and analysis of RADseq datasets.* Bioinformatics, 2020. **36**(8): p. 2592-2594.

112. Kulkarni, R., et al., *CAPG: comprehensive allopolyploid genotyper.* Bioinformatics, 2023. **39**(1): p. btac729.
113. Michael, T.P. and R. VanBuren, *Building near-complete plant genomes.* Current Opinion in Plant Biology, 2020. **54**: p. 26-33.
114. Clevenger, J.P. and P. Ozias-Akins, *SWEEP: A Tool for Filtering High-Quality SNPs in Polyploid Crops.* G3 Genes|Genomes|Genetics, 2015. **5**(9): p. 1797-1803.
115. Clevenger, J.P., et al., *Haplotype-Based Genotyping in Polyploids.* Frontiers in Plant Science, 2018. **9**.
116. Eriksson, J.S., et al., *Allele phasing is critical to revealing a shared allopolyploid origin of Medicago arborea and M. strasseri (Fabaceae).* BMC Evolutionary Biology, 2018. **18**(1): p. 9.
117. Meirmans, P.G., *Analyzing Autopolyploid Genetic Data Using GenoDive*, in *Polyploidy: Methods and Protocols*, Y. Van de Peer, Editor. 2023, Springer US: New York, NY. p. 261-277.
118. Meirmans, P.G., *genodive version 3.0: Easy-to-use software for the analysis of genetic data of diploids and polyploids.* Molecular Ecology Resources, 2020. **20**(4): p. 1126-1131.
119. McAllister, C.A. and A.J. Miller, *Single nucleotide polymorphism discovery via genotyping by sequencing to assess population genetic structure and recurrent polyploidization in Andropogon gerardii.* American Journal of Botany, 2016. **103**(7): p. 1314-1325.

Appendix A

Table A1. Assembly statistics obtained after analysing the WGS assembly generated using the MEGAHIT assembly software.

Statistic	**WGS assembly**
Nucleotide distribution (A:C:G:T %)	32.75 : 17.69 : 17.66 : 31.91
Scaffold N50 (bp)	311 545
Scaffold L50	952
GC content (%)	35.35
Total scaffold count	2 408 056
Total contig count	2 408 056
Max Scaffold length (bp)	89 422
Max contig length (bp)	89 422
Sequence total (bp)	1 496 764 653

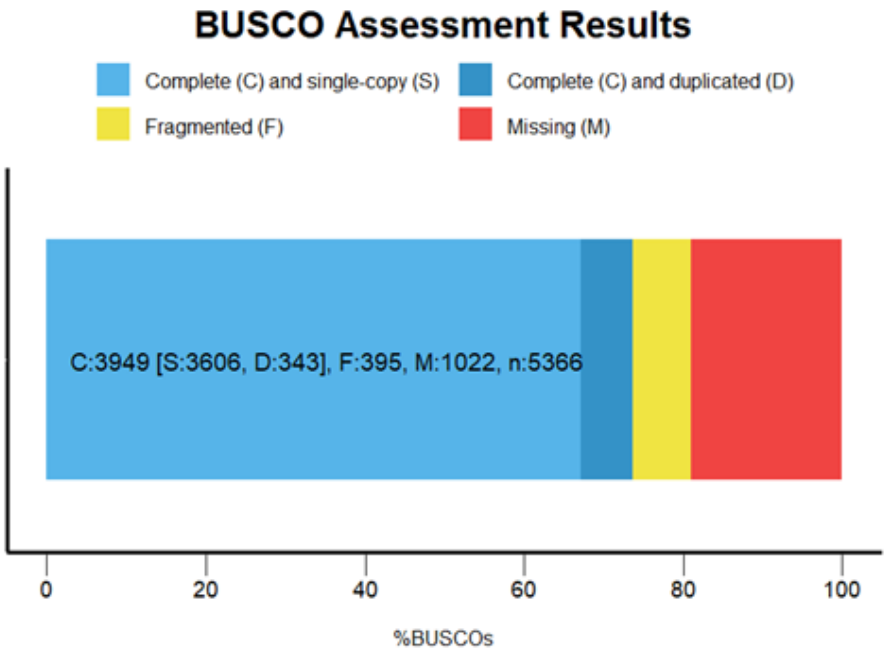

Figure A1. BUSCO scores obtained from the analysis of the assembly generated from sample K47 WGS data using the MEGAHIT assembly software.