Metabarcoding pollen to identify allergenic species of grasses (Poaceae) and ragweed (*Ambrosia* spp.) across six monitored sites in South Africa

Siyavuya Sidla

Master of Science dissertation

Supervisor: Professor J. Peter

Co-Supervisors: Dr D. Berman, Dr N. Esterhuizen and Dr S. Pedretti

Division of Allergology and Clinical Immunology
Department of Medicine
Faculty of Health Sciences
University of Cape Town

DECLARATION

I, ……*Siyavuya Sidla*……, hereby declare that the work on which this dissertation/thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another `degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: … Signed by candidate ……

Date: ……6 February 2025…….

ACKNOWLEDGEMENTS

1. UCT institution and facilities

UCT institution for providing all the necessary facilities to carry the study and programmes that supported the course of the study.

2. UCT Lung Institute and DRC for funding

The financial aid that helped with stipends for the duration of this study was provided and made possible by the awards I received from the UCT Lung Institute and a fellowship from the Department of Medicine Research Council.

3. Supervisory and mentors

I am grateful for all the support I received from my lead supervisor, Prof Peter who provided me the opportunity and to conduct this study, Dr Berman leading the aerobiology side and Dr Pedretti that supported me with the DNA laboratory work, Dr Esterhuizen who also became a mentor and a friend provide me all the support and constantly checked up on my mental wellbeing. Dr Nindo and Dr Mugo who assisted with the bioinformatics analysis.

4. AI tools for writing

I acknowledge that I have at times used the OpenAI language model, ChatGPT to better my grammar in certain sections [1].

5. Colleagues and friends

The community of staff and students of the Allergology and Clinical Immunology at H52 and H47, who were always there to listen to my struggles and complaints and always available to help when I needed them the most.

6. Family

I would call my family just to hear my mom's, siblings and daughter's voices on days when the going got tough, hearing them kept me motivated. My partner, Ayanda, who held my hand and witnessed it all but made sure to push and lift me on days I felt knocked down by research.

Dedication

To my late sister, Aphelele, you believed in me more than I will ever believe in myself. This is for you mntakamama.

RESEARCH OUTPUTS

Presentations

- 49th Department of Medicine Science for Health Annual Research Symposium (UCT Groote Schuur Hospital) - Oral presentation
  Title: Metabarcoding pollen to identify allergenic species of grasses (Poaceae) and ragweed (Ambrosia) across six monitored sites in South Africa
  Date: 09&10 October 2024

- 37th Annual Conference of the South African Society for Atmospheric Sciences with ECERA and REACH Workshop - Oral presentation
  Title: Grass pollen trends as markers of climate change in Cape Town
  Date: 30 October - 03 November 2023

- 48th Department of Medicine Science for Health Annual Research Symposium (UCT Groote Schuur Hospital) - Poster presentation
  Title: Climate change induced shifts in grass pollen season length and concentrations
  Date: 11&12 October 2023

- Allergy Society of South African and South African Immunology Society (ALLSA/SAIS) Congress 2023 - Oral Presentation
  Title: Grass pollen seasons lengthening and strengthening in South Africa in response to climate change
  Date: 28 September - 01 October 2023

- 5th National Global Change Conference 2023 (UFS Bloemfontein) – Oral presentation
  Title: Metabarcoding grass pollen in South Africa: A possible biomarker for climate change
  Date: 30 January - 2 February 2023

Publication

Ambrosia (ragweed) pollen — A growing aeroallergen of concern in South Africa

Dorra Gharbi, Dilys Berman, Frank H. Neumann, Trevor Hill, **Siyavuya Sidla**, Sarel S. Cillers, Jurgens Staats, Nanike Esterhuizen, Linus Ajikah, Moteng E. Moseri, Lynne J. Quick, Erin Hilmer, Andri Van Aardt, Juanette John, Rebecca Garland, Jemma Finch, Werner Hoek, Marion Bamford, Riaz Y. Seedat, Ahmed I. Manjra, Jonny Peter

World Allergy Organization Journal, Volume 17, Issue 12

Table of Contents

Dissertation Summary

Background and aims

Aeroallergens contribute to the global burden of non-communicable diseases, causing allergic conditions such as rhinitis, conjunctivitis, and asthma. Pollen is the second most important contributor, and despite cross-reactivity between allergens, patients often experience distinct sensitivity patterns to different species. The occurrence of allergenic species varies with geography, seasonality, and their flowering periods are influenced by meteorological factors and climate. The South African Pollen Network (SAPNET) monitors aerospora in different biomes of the country, providing weekly reports on the prevalent species and their concentrations, raising awareness among allergy sufferers and healthcare providers. Grass pollen is one of the most abundant contributors to airborne allergies. However, current SAPNET methods of analysing pollen, using light microscopy, can only identify grasses up to the family level and ragweed (*Ambrosia* spp.) pollen to the genus level. This project aimed to introduce DNA metabarcoding techniques to classify and identify grass and ragweed species contributing to pollen allergies in various South African biomes.

Methods

Samples were selected from weeks with the highest grass pollen counts between October 2022-April 2023 and included samples from sites where *Ambrosia* pollen was identified (Durban and Potchefstroom). DNA was isolated from environmental pollen samples using the NucleoMag bead beating method, optimised to obtain a larger quantity of double-stranded DNA (dsDNA) and further purified with an additional OneStep PCR inhibitor removal step to improve purity. DNA metabarcoding analysis was performed using two taxonomic markers: ribulose-1,5-bisphosphate carboxylase (*rbcL*) and internal transcribed spacer 2 (ITS2), amplified with universal primer pairs in a two-step polymerase chain reaction (PCR) library preparation for Next Generation Sequencing (NGS) with MiSeq Illumina V2 systems. The sequence reads generated were assessed for quality, pre-processed, and assigned to taxa using two bioinformatics pipelines. The first followed the method for analysing metabarcoding dual-index data described by Sickel *et al*. (2015), while the second was adapted from the National Botanic Garden of Wales Plant Illumina pipeline developed by Ford and Jones (2021).

Results and Discussion

The DNA extraction protocol was optimised by altering the bead beating method using the samples from weeks with high grass pollen concentrations. From the improved method of extraction, we were able to obtain sufficient DNA yield with an average of 21.88 ng/µL. Additional purification improved the dsDNA A280/260 purity ratio to within the ideal range

for 73% of the samples, though it introduced salts that decreased the A230/260 ratio below the optimal range. Amplification of *rbcL* and ITS2 barcodes initially produced amplicons of 579 bp and 542 bp, respectively. However, to limit sequencing costs, an alternative set of ITS2 primers producing amplicons were explored. A total of 25 samples progressed to NGS: Cape Town (7), Kimberley (6), Bloemfontein (3), Johannesburg (3), Durban (3), and Potchefstroom (2), and six ITS2 samples were excluded during quality check due to short amplicon lengths below 100 bp.

The three taxonomic classification tools; UTAX, RDP, and BLAST were used to assign taxa, with a 50-read cutoff applied as the minimum read count, with confidence scores of ≥80% for UTAX and ≥0.8 for RDP. For BLAST classifications, only sequences with 99 - 100% identity was considered, using thresholds of ≥80% query coverage and an E-value of $\leq 1.00\times10^{-5}$. UTAX classified *rbcL* and ITS2 sequences identified 233 unique Poaceae species, including *Lolium perenne* (ryegrass), which was one of the 17 species identified by both barcodes. RDP classifications resulted in 48 Poaceae species, with *Poa annua* (annual bluegrass) and *Cynodon dactylon* (Bermuda grass) being two of the three species classified by both barcodes. BLAST classifications produced the largest biodiversity, identifying 307 unique Poaceae species with 19 species common to both barcodes including *Cynodon dactylon* and *Lolium perenne*. Additionally, *Paspalum notatum* (Bahia grass), *Phleum pratense* (Timothy grass), *Phragmites australis* (common reed) and *Stenotaphrum secundatum* (buffalo grass). Two ragweed species, *Ambrosia artemisiifolia* and *A. trifida* were positively identified from *rbcL* sequences with classifications using the three tools.

Conclusions

This pilot study successfully used pollen metabarcoding to identify known allergenic grasses in several South African regions, including Bermuda, rye, Timothy, common reed and annual bluegrass in Cape Town; Bermuda, rye, and annual bluegrass in Kimberley; Bahia and Bermuda grass in Johannesburg; weeping love and Bermuda grass in Bloemfontein; and *Ambrosia artemisiifolia* and *A. trifida* in Durban. The short length of DNA libraries limited sequence read length, which significantly influenced the classification of operational taxonomic units (OTUs). The low confidence in taxa assignments from UTAX classifications and the identification of species via BLAST that are not recorded in South Africa raised concerns regarding the reliability of UTAX as a taxonomic classification tool and the NCBI based reference databases, which are primarily developed based on Northern Hemisphere plant data and may misclassify endemic species. Future research should address these limitations

and include a larger number of SAPNET samples to better map the seasonality and distribution of Poaceae species across South African biomes.

Keywords: aeroallergens, pollen allergies, grass pollen, ragweed pollen, metabarcoding, Next-generation sequencing, DNA extraction, taxonomic classification, barcodes, taxa assignment, pollen monitoring, South African Pollen Network, *rbcL*, ITS2

Chapter 1

Introduction: Monitoring airborne pollen in South Africa

1.1. Pollen monitoring and sampling

Pollen monitoring in South Africa was pioneered by David Ordman, whose research spanning between 1945 to 1970, laid the foundation for aerobiology in the country. Ordman's studies systematically examined airborne pollen, identifying grasses as significant contributors to pollen allergies [2], [3], [4]. This early work highlighted the impact of grass pollen on respiratory health and established a basis for monitoring allergenic pollen.

For the longest time pollen monitoring in South Africa was conducted in few sites including Cape Town and Pretoria, with Cape Town having the longest pollen data dating from monitoring in the early 1970s and Pretoria since 1987 [5], [6]. However, with the establishment of the South African Pollen Network (SAPNET), multiple pollen monitoring stations have been deployed in major cities in diverse geographic regions – biomes, launching various sites of data collection across the country as shown by Esterhuizen *et al*. [7]. These stations currently use Burkard Hirst volumetric samplers to capture airborne pollen with their locations strategically selected based on factors like population density, pollen exposure risk, and biodiversity to ensure representative sampling of regional pollen loads [8]. SAPNET stations play a critical role in data collection, recording the concentrations and seasonal patterns of various pollen types across South Africa's different biomes. The continuous data generated by SAPNET has provided valuable insights into the allergenic pollen prevalent in each region and has enabled the development of pollen calendars, which track the seasonality of pollen types [7], [8].

1.2. Grasses in South Africa

Grasses, belonging to the Poaceae family have over 11 000 species distributed worldwide across diverse habitats. South Africa hosts approximately 10% of the total grass species, consisting of 967 indigenous, 329 endemic and 115 naturalized species contributing to the region's biodiversity [9], [10], [11]. Grasses are ubiquitous in many environments, from fields and lawn where species like Kikuyu grass and buffalo grass are common, to essential crops like maize and rice, which are staples in our diets[5], [12]. Poaceae species are classified as either annual, biennial, or perennial based on their flowering cycles and because they are

primarily wind-pollinated, they produce copious amounts of pollen that can be carried long distances through the air, impacting allergy sufferers [13]. During the flowering seasons of grass species, airborne pollen often triggers allergenic responses, contributing to respiratory diseases affecting about 20 million South Africans [8]. Although the seasonal patterns of individual grass species in South Africa are known, it remains unclear which specific species are most dominant at which point of the grass pollen season. Identifying the dominant species in each region throughout the flowering season is essential for pinpointing which grasses are primarily responsible for triggering allergies at different times [14], [15].

1.3. Distribution of grasses in South African biomes

Grasses classified into two main groups, C_3 and C_4, based on their distinct photosynthetic pathways for carbon dioxide (CO_2) assimilation which largely influences their adaptation to environmental conditions. In South Africa, the distribution, growth, and flowering of these grasses are primarily shaped by climatic and topographical factors, including latitude, altitude, and seasonal rainfall patterns, which vary across the country's diverse biomes. While C_3 grasses are well known to typically thrive in cooler and wetter climates in the Southern regions of the country, they can also be found in high lying areas of the country. C_4 grasses are more efficient in warmer and drier environments and are therefore predominantly in abundance in Northern areas and the Indian Ocean coastal belt [1], [2].

Common examples of C_3 grasses include the common wild oat grass, quaking grass, perennial ryegrass, hare's tail grass, common reed and annual blue grass - grasses that are commonly found in the Cape fold and coastal areas of the Western Cape, and in the highveld interior regions in Gauteng and Free State, as illustrated in Figure 1.1 [10].

Grasses grouped as C_4 grasses are mostly found in warmer tropical and subtropical climates with higher rainfall and are found in lower altitudes regions of South Africa. Examples of these grasses include Bermuda grass, common thatching grass, weeping love grass, Bahia grass, Kikuyu grass, Johnson grass and buffalo grass. They are dominantly found in low veld regions of Mpumalanga and Limpopo, along the coastal region of KwaZulu-Natal, and in bushveld regions [16] and in the Grassland, Savanna, Fynbos, Nama-Karoo and Indian Coastal Belt biomes of our country - illustrated in Figure 1.2 [10].

As the severity of grass pollen allergies often varies with the flowering season of each species, researchers have hypothesised that climate change may be driving an increase in allergen content in certain species [17], [18]. Rising temperatures, changes in rainfall patterns, and

elevated CO_2 levels associated with climate change can influence both the timing and intensity of pollen production. Warmer temperatures may lead to longer flowering seasons, allowing some grasses to produce pollen for extended periods, while increased CO_2 levels boost pollen production and potentially increase the allergenicity of pollen grains [18], [19]. These changes may result in higher concentrations of airborne pollen, extended allergy seasons, and more intense allergic responses [17], [20].

1.4. Grass species behind allergies in South African biomes

The prevalence and severity of grass pollen allergies in South Africa vary geographically due to the diversity of grass species across the country's unique biomes. South Africa's regions support a rich variety of grass species, because of distinct climatic conditions and weather patterns which then impacts vegetation distribution and contribute to the country's biodiversity [21]. Different grass species dominate in different regions, leading to variations in allergenic pollen exposure across the biomes. For instance, allergenic pollen from *Cynodon dactylon* (Bermuda grass) is commonly observed in the northern regions, particularly in the Savanna and Grassland biomes, where warmer climates prevail. In contrast, species like *Lolium perenne* (perennial ryegrass) are more common in the Fynbos and Coastal biomes, which experience cooler temperatures [7].

SAPNET pollen data collected between 2019 - 2021 showed that four monitoring sites; Bloemfontein, Cape Town, Johannesburg, and Kimberley recorded the highest annual pollen indices for Poaceae (grasses), indicating a significant presence of grass pollen in regions encompassing the Fynbos, Grassland, Savanna, and Nama-Karoo biomes [7].

To support allergy awareness and management, we compiled a list of airborne allergenic grasses monitored in South Africa, available on The Real Pollen Count website (www.pollencount.co.za). This list includes both indigenous and naturalized species studied by Ordman, Potter, and Berman, and non-native species with confirmed presence in the Botanical Database of Southern Africa (BODATSA) maintained by SANBI [3], [5], [22]. Additionally, all species on this list are included in South African local Skin Prick Test (SPT) panels for allergy testing. Using information from the "Identification Guide to Grasses of Southern Africa" [11], we outlined each species' flowering period during the grass pollen season, as shown in Table 1.1. Furthermore, we mapped the distribution of these species from their regions of origin to their regions of occurrence in South Africa and grouped them based on their photosynthetic pathways, with C_3 grasses illustrated in Figure 1.1 and C_4 grasses in Figure 1.2.

Table 1.1. Grass species in Southern Africa monitored and associated with airborne pollen allergies with the corresponding flowering periods during grass pollen season and their regions of origin and distribution areas [11].

Species	Common name	Flowering period		Ecological distribution
		Start	End	
Avena fatua	Common wild oat grass	September	November	Naturalised from Europe
Briza maxima	Big quaking grass	July	December	Naturalised from Mediterranean region
Briza minor	Little quaking grass	September	December	Naturalised in warm temperate regions from the Mediterranean region
Cynodon dactylon	Bermuda grass	September	May	Worldwide in tropical and warm temperate regions
Eragrostis curvula	Weeping love grass	August	June	Native to East Africa, introduced throughout the tropics
Hyparrhenia hirta	Common thatching grass	September	June	Distributed across Africa to the Mediterranean and Pakistan
Lagurus ovatus	Hare's tail grass	October	November	Naturalised from Mediterranean
Lolium perenne	Perennial ryegrass	November	April	Naturalised from Europe, introduced globally
Paspalum notatum	Bahia grass	November	April	Native to tropical Africa and America; naturalised from South America
Pennisetum clandestinum	Kikuyu grass	August	April	Naturalised from east African highlands, introduced worldwide
Phragmites australis	Common reed	December	June	Cosmopolitan
Poa annua	Annual blue grass	January	December	Naturalised from Europe, Mediterranean region and eastwards to India and Central Asia; introduced worldwide
Sorghum halepense	Johnson grass	December	May	Naturalised from southern Eurasia and India, widespread in warm areas naturalised globally
Stenotaphrum secundatum	Buffalo grass	October	May	Native to Indian Ocean shores of Africa to Sri Lanka with introductions inland in Zimbabwe and East Africa

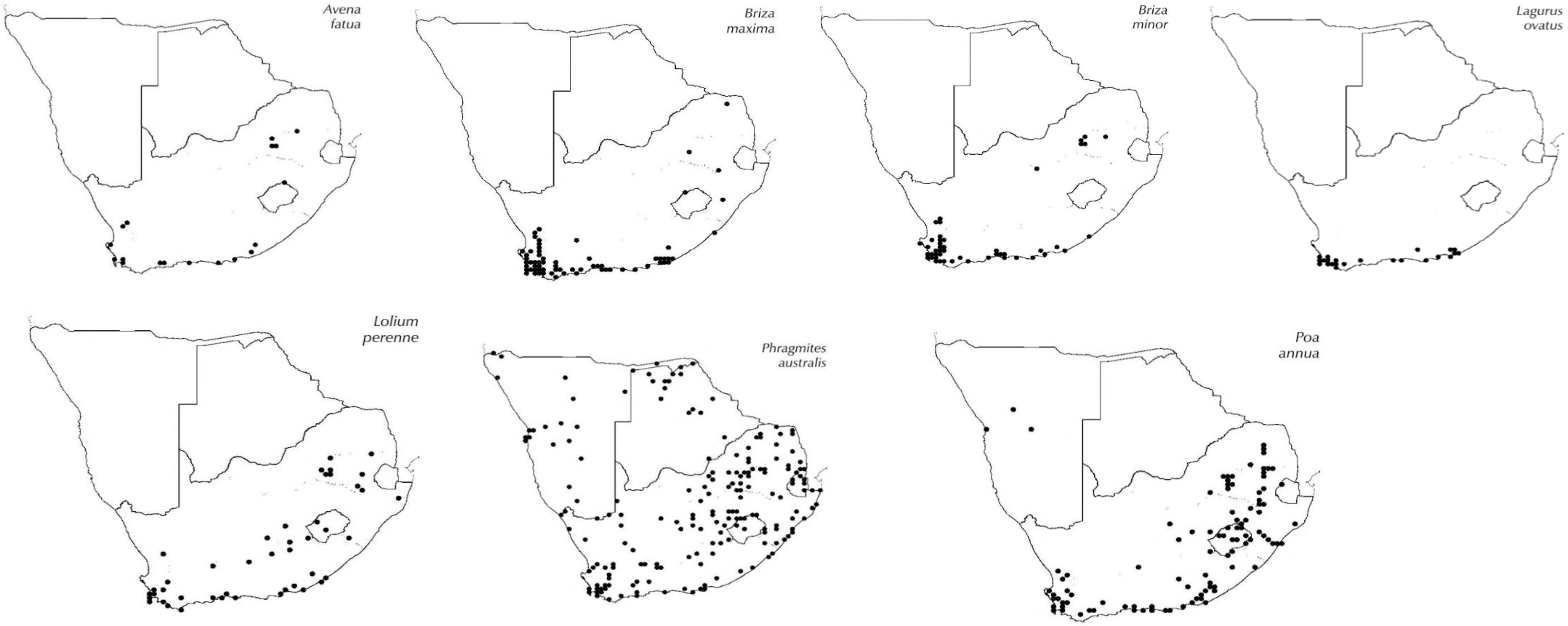


Figure 1.1. The regions of distribution of C_3 grasses associated with pollen allergies in across the country. *Avena fatua, Briza maxima /B. minor* and *Lagurus ovatus* occurring mainly in the coastal area of the Western Cape while *Lolium perenne, Phragmites australis* and *Poa annua* are widely spread across largely contributing to grass pollen allergies [10].

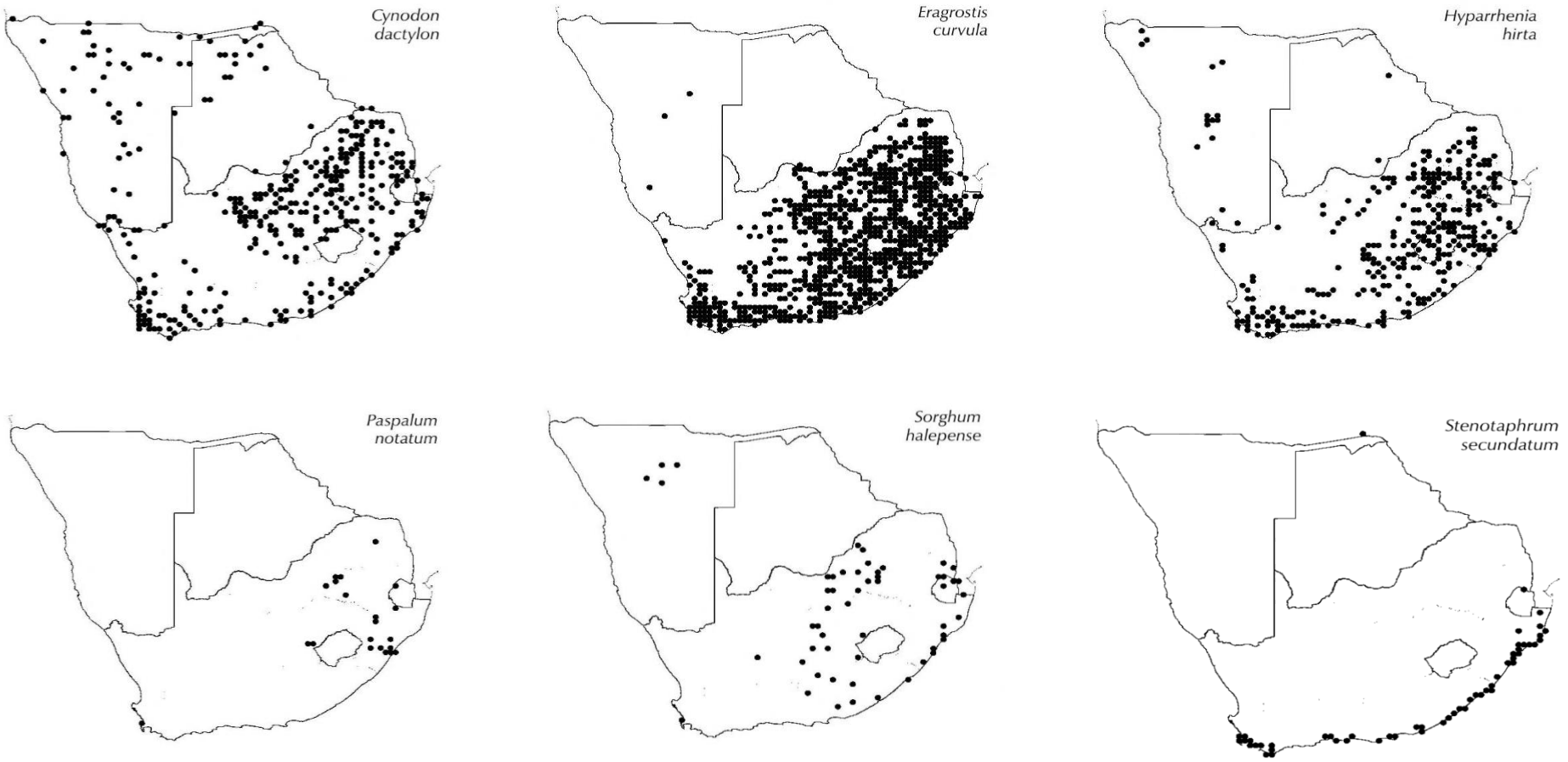


Figure 1.2. Distribution of C_4 grasses across the country; *Cynodon dactylon, Eragrostis curvula* and *Hyparrhenia hirta* are widely distributed while *Paspslum notatum* and *Sorghum halepense* mainly found in the Northern areas *and Stenotaphrum secundatum* found in the coastal belt [10].

1.5. Grass pollen allergenicity

Grass pollen allergens are divided into eleven distinct groups based on their immunological traits, with Group 1 and Group 5 being the most responsible for triggering allergic reactions in approximately 95% of grass pollen allergy sufferers [23]. However, only pollen of C3 grasses and not C4 grasses have been found to contain the group 5 allergen [24], [25]. Cross-reactivity among grass pollen allergens is common, meaning people allergic to one grass species commonly react to others, complicating diagnosis and treatment [13], [23], [24], [26].

However, there are peoples whose serum IgE is not cross-reactive with sensitisation to only one species e.g., Bermuda and not rye, or vice versa. Upon compiling the data for Immunoglobulin E (IgE) sensitisation test using Allergy Explorer test from 319 patients tested in our local allergy clinic at University of Cape Town, as shown in Figure 1.3, it was noted that there is high prevalence of grass pollen sensitisation but also different patterns of sensitisation.

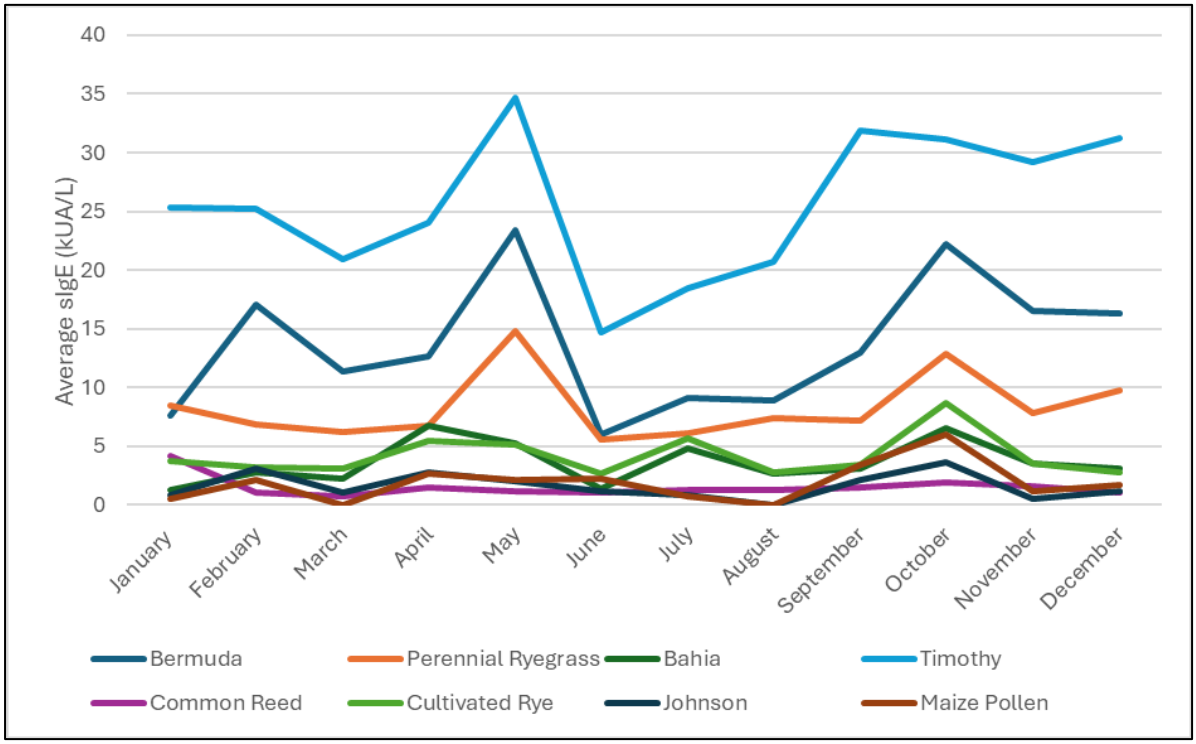


Figure 1.3. Summary of seasonal grass Pollen IgE levels from 319 patients tested in Allergy clinic in Cape Town.

Several management and treatment options available for grass pollen allergies including allergen avoidance strategies such as keeping windows closed during high pollen seasons. Additionally, pharmacotherapy with antihistamines and corticosteroids, and allergen-specific immunotherapy (AIT) are also used for treatment. However, even with the above-mentioned strategies there is a need to provide a more detailed understanding of the grass pollen seasons

across different regions of South Africa. This will allow patients with specific sensitisation patterns to tailor these therapeutic approaches, potentially improving their quality of life [26].

Furthermore, while some patients may show sensitisation to grass pollen (IgE positive) without significant clinical symptoms, the decision to initiate the often-complex AIT relies heavily on a clear history of clinically significant allergic symptoms aligned with IgE sensitisation patterns. A targeted, region-specific pollen calendar could therefore enhance both diagnosis and management, helping clinicians and patients determine the necessity and timing of immunotherapy more precisely [22], [27].

1.6. Challenges in grass pollen identification

SAPNET's pollen monitoring stations use light microscopy for pollen analysis, a technique that has limitations in distinguishing between different grass species within the Poaceae family [15]. Poaceae pollen grains are typically spheroidal to sub-oblate and are morphologically similar across species, which complicates species-level identification when using conventional light microscopy [28]. Phenological studies on Poaceae pollen indicate that key characteristics necessary for distinguishing subfamilies or genera only become discernible at magnifications above 600-1000x, generally requiring oil immersion microscopy [29], [30].

Pollen grain sizes within Poaceae range from approximately 20 µm to over 100 µm, with variations observed both between and within species. This size variation, however, has been found to be insufficient as a sole criterion for species identification [12]. Figure 1.4(A) shows an example of Poaceae pollen captured at the Cape Town station and visualised, illustrating two pollen grains collected at different times.

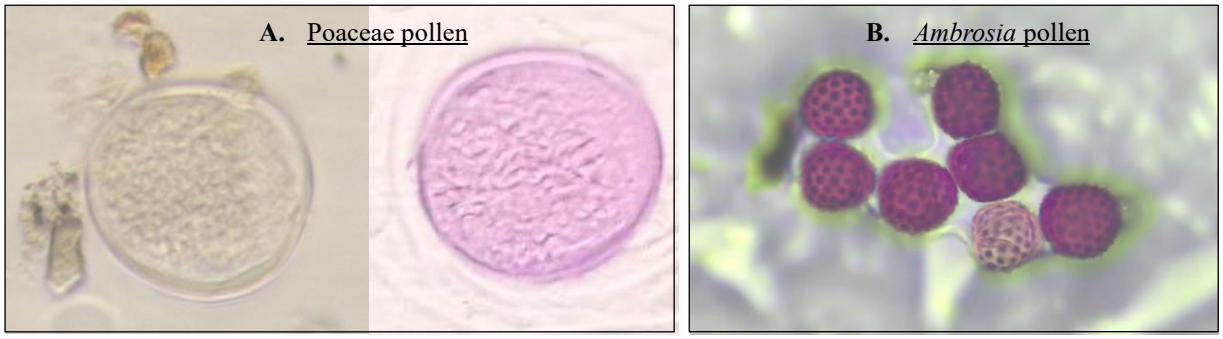


Figure 1.4. Grass and ragweed pollen captured in SAPNET spore traps and visualized under a light microscope. (A) Poaceae pollen grain collected from the Cape Town station: (left) unstained Poaceae pollen and (right) stained Poaceae pollen, both observed at 400x magnification. (B) Stained *Ambrosia* (ragweed) pollen collected from the Durban station, observed at 160x magnification.

The variation in colour between the pollen grains is due to the visualisation agent - fuchsin used as a dye during slide preparation to enhance visibility. While the pollen grains can be identified as belonging to the Poaceae family, distinguishing the exact species or attaining any discriminating characteristic beyond family level remains a challenge. This difficulty in species-level identification also applies to other allergenic taxa, such as ragweed (*Ambrosia* spp.), which are known to trigger respiratory allergies.

1.7. Ragweed species

Ragweed (*Ambrosia* spp.) consists of over 40 species, with *Ambrosia artemisiifolia* and *A. trifida* being the most common and widespread. South Africa is one of the countries outside North America and Europe where ragweed plants have been reported as illustrated in the global distribution map in Figure 1.5 as shared by Chen *et al*. 2018 [31].

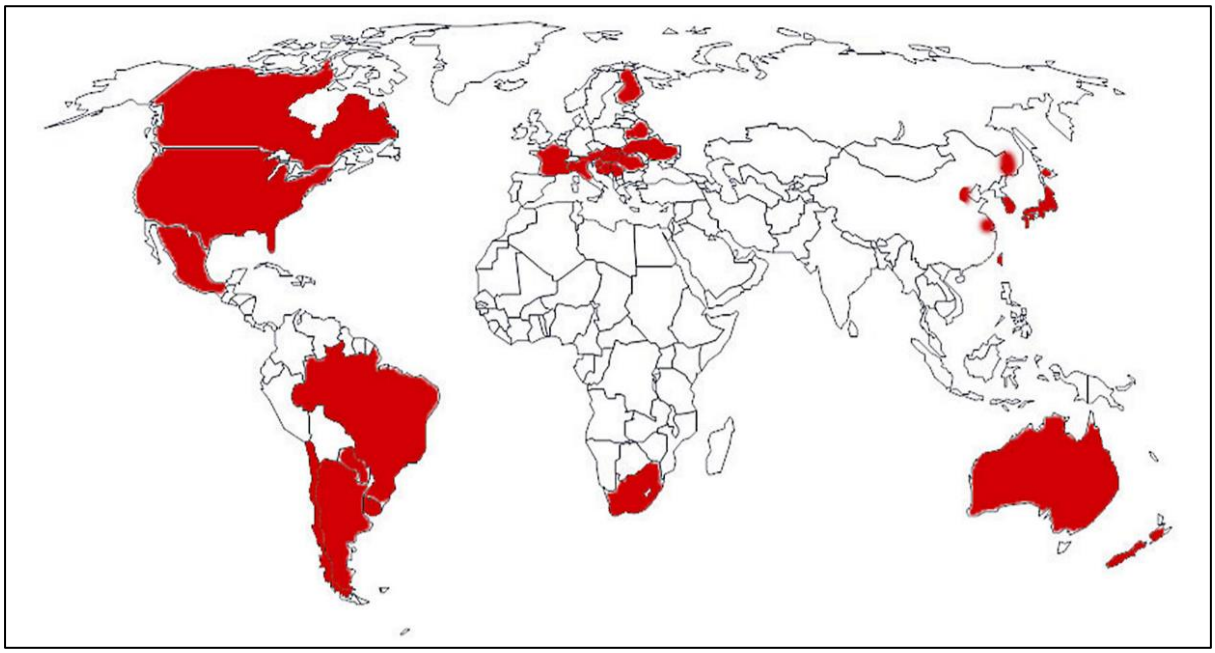


Figure 1.5. Worldwide distribution of ragweed pollen with the countries of reported occurrence marked in red. Illustration from Chen K. *et al*. 2018 [31].

Ambrosia pollen has been observed in pollen at Durban since the early weeks of the establishment of SAPNET in 2019, and recently it has been identified at other SAPNET sites as well. Captured and visualised *Ambrosia* pollen from Durban station is shown in Figure 1.4(B). Most recently, *Ambrosia* pollen grains were identified at the Potchefstroom site in the North West province during extreme windy period from 03 – 16 April 2023. Although the

pollen was confirmed to be of the *Ambrosia* genus based the specific morphology of spikes on the surface of the pollen grains, it could not be identified to the species level upon analysis with light microscope.

The pollen from various species of *Ambrosia* can be differentiated by examining the surface texture and size of pollen, and the distance between the spikes which can be observed using Scanning Electron Microscopy (SEM) [32]. However, we found SEM pollen analysis to be challenging when analysing environmental pollen samples, due to the nature of current collection of pollen on a coated sticky tape of Burkard Hirst volumetric spore straps.

The inability to identify *Ambrosia* pollen to species level at SAPNET poses a significant threat to patients with pollen allergies in South Africa as even low concentrations such as 10 pollen grains per cubic meter (pg/m^3), can trigger allergic reactions [33], [34]. Furthermore, gaps in the records of the South African National Biodiversity Institute (SANBI) 'Red list of South African Plants' detailing the distribution of *Ambrosia* species in the country complicates effective diagnosis and treatment strategies, making it challenging for allergists to accurately identify the specific species responsible for allergic reactions [3], [4], [5]. Addressing these challenges is essential, as unidentified pollen sources continue to impact public health. In response to the growing need within the healthcare sector and the goal of improving quality of life for allergy sufferers, establishing a reliable method for identifying grass and ragweed pollen species in South Africa is a pressing priority [35].

1.8. Allergenicity of invasive ragweed species (*Ambrosia* spp.)

Invasive species of ragweed (*Ambrosia* spp.) have garnered increasing attention due to their potent allergenic properties and adverse effects on human health. These plants, originally native to North America, have spread to various regions worldwide, posing a significant threat to allergy sufferers due to their highly allergic nature [31], [34]. Several species of ragweed, especially *Ambrosia artemisiifolia* (common ragweed) and *Ambrosia trifida* (giant ragweed), have become invasive in regions far from their native habitats [36]. Human activities, including transportation and trade, have facilitated the spread of ragweed seeds, leading to their establishment in Europe, Asia, and other continents as illustrated in Figure 1.5 from the worldwide distribution of ragweed described by Chen *et al*. 2018 [31]. Climate change and altered environmental conditions are also contributing to further promote the spread of invasive ragweed species. Increased temperatures, extended growing seasons, and elevated carbon

dioxide levels have the potential to enhance the allergenicity and invasive potential of ragweed [33], [36], [37]. While the presence in South Africa is not as extensive, the potential for these plants to become more widespread poses a significant threat due to the nature of abundant release of their pollen [36], [38].

Ragweed pollen along with grass pollen is a leading cause of allergic rhinitis (hay fever) in many regions [39], [40]. Exposure to ragweed pollen can result in exacerbation of asthma in susceptible individuals and can significantly impact an individual's quality of life, leading to reduced productivity, missed work or school days, and a decreased sense of well-being during peak pollen seasons [31], [40], [41].

The detection of *Ambrosia* species in SAPNET spore traps have raised a call for alert, because if this is a result of climate change events it means that these invasive species have begun to establish themselves, raising the need to accurately identify the pollen.

Chapter 2

Literature Review: The need for pollen DNA metabarcoding in South Africa

2.1. Introduction

Pollen DNA metabarcoding emerged as a method enabling aerobiologists and environmental scientists to monitor flowering patterns and geographical distributions of plant species throughout South Africa. DNA based approaches of pollen analysis deliver broader identifications which surpass the conventional methods, and over the years these techniques have enabled better large-scale analysis [42]. These methods of analysis have been established and are widely used by different pollen monitoring stations in the Northern Hemisphere.

Pollen metabarcoding became prominent with the development of universal primer pairs capable of amplifying taxonomic gene regions 'barcodes' across multiple species samples [43], [44]. Unlike animals, plants do not have a standard DNA barcode, leading to the extensive studies on several barcode candidates including chloroplast DNA regions (e.g., *rbcL*, *matK*, *trnL*) and nuclear ribosomal DNA Internal transcribed spacers (e.g., ITS, ITS1, ITS2) [45], [46]. The adaptation of this DNA-based method has found wide application in pollen analysis and has aided in investigating plant-pollinator interactions, mapping of flora diversity, and the monitoring of airborne allergenic pollen [47], [48], [49], [50], [51].

2.2. Airborne pollen DNA metabarcoding

Metabarcoding is dedicated to achieving precise and meaningful outcomes characterised by a high degree of reproducibility. Researchers are actively improving the methodology's efficiency and continually refining its analytical capabilities [50]. However, the handling and processing of environmental samples pose limitations. These challenges include the risk of introducing contaminants, the potential for cross-contamination, and the susceptibility of DNA to degradation and fragmentation during extraction processes [52], [53].

The application of metabarcoding lies in the inherent biases associated with PCR amplification. Specifically, the preferential amplification of certain regions across diverse plant groups [54]. Universal barcode primers designed for species-level resolution often exhibit limitations in their applicability and this often becomes evident when attempting to differentiate between closely related species in downstream analyses [55]. There are ways that have been developed over the years to help in improving the major steps of the methodology.

The experimental studies on the application of pollen DNA metabarcoding have shown that the method relies on four key components: i) DNA extracted with high integrity for subsequent sequencing; ii) The choice of a barcode; iii) Barcode library reference databases iv) Sequencing and bioinformatic analytic methodologies [56].

2.2.1. Extracting DNA from environmental pollen

Different pollen collection methods have been shown to have an impact on the DNA extraction process, influencing both the extraction method applied and the amount of DNA obtained. While most studies on pollen metabarcoding for taxonomic identification have focused primarily on honey pollen material [53], [56], [57], several pollen monitoring groups have adopted this method for aerobiological samples due to its complex nature and requirement of small amount of material [58]. When using environmental pollen monitoring samples where total biological material is frequently low, the selection and optimisation of extraction methods that are efficient to maximise extract pollen DNA quantity are critical. The method should effectively disrupt the tough outer layer of the pollen grains while preserving the integrity of the DNA. Techniques such as pulverizing, bead beating, cell lysing with proteinases or devices have been explored and adopted as the first step of cell lysis in the extraction method [53], [56].

The extraction and purification of DNA from pollen have mostly relied on commercial kits based on either silica gel spin column technology or magnetic bead technology. While the comparability of the two methods remains uncertain, a study by Leontidou in 2018 reported that the magnetic bead-based DNA extraction yielded a greater quantity of DNA [53]. We evaluated the various techniques that have been used to extract DNA from airborne pollen samples to weigh the efficiency of their method retrieving DNA with high quality and quantity from airborne pollen as shown in Table 2.1.

Table 2.1. Various techniques that have been explored for extracting DNA from environmental pollen samples, pros and cons of the application of each method.

DNA extraction technique	**Method**	**Advantages**	**Disadvantages**
Mechanical disruption with steel beads – Nucleomag kit [53], [59]	High energy agitation with beads	High DNA yield and effective for mechanically disrupting the walls	Requires specialised equipment

DNeasy Plant mini kit with glass/steel beads [53], [60]	Liquid nitrogen freezing and bead beating	Improved DNA yield with frozen samples and enhances lysis efficiency	Requires liquid nitrogen and freezing step adds complexity
Qiagen Qiacube system [53], [61]	Silica column-based DNA extraction	Relatively fast automated method	Lower DNA yield with PCR inhibitors impurities
Magnetic beads-based DNA extraction – Nucleomag kit [53], [62]	Magnetic beads-based DNA extraction	Higher DNA quality and yield and less prone to contamination	Requires multiple wash steps and may result to sample loss

Although extracting high quantity DNA is desirable the quality is mostly important as it directly impacts amplification and processing of DNA libraries for sequencing. Therefore, it is important to choose an extraction method that not only maximises the yield but also ensures high quality DNA for downstream processing.

2.2.2. DNA barcode selection

The second key component of metabarcoding is the selection of a barcode. This is more complex in plants since they lack a universally established barcode unlike animals, which commonly use the mitochondrial Cytochrome c Oxidase Subunit I gene (COI) as a maternally inherited standard barcode for broad species identifications [63]. This is due to the differences in their reproduction system as well as the widely dispersed taxonomic information within plant genomes, found in both chloroplast and nuclear ribosomal DNA regions [56]. The Consortium for the Barcode of Life (CBOL) Plant Working Group has proposed various taxonomic gene sequences for barcoding plant material, with a primary emphasis on plastid genome sequences such as *rbcL* and *matK*, which have been extensively explored as barcodes [46], [64].

When metabarcoding different plant groups, the use of a combination of barcodes is usually needed to address and account for genetic variability [65], [66], [67]. This offers several key advantages including: i) resolution enhancement, as different barcodes may vary in their ability to different between closely related species [68] and ii) accounting for intraspecific variation, considering that individuals of the same genera may exhibit slight genetic differences, as seen in the case of *Ambrosia* spp. [43], [69]. Therefore, careful considerations are essential when

selecting barcode combinations for each study, with a focus on the unique benefits and discriminatory capabilities offered by each barcode [45].

Among the plant barcodes commonly used in metabarcoding for species differentiation, *psbA-trnH* has been found to offer high inter-specific divergence and reliable amplification across diverse plant taxa but is limited by lower species-level identification accuracy [65]. *matK* has been found to show significant sequence variation in certain plant groups, but its low inter-specific divergence and limited barcoding gap reduce its effectiveness in distinguishing closely related species [70]. Similarly, *rbcL* provides consistent amplification and low intra-specific variation, making it suitable for broader taxonomic resolution, though it struggles with closely related species [55]. The ITS DNA Internal transcribed spacer was found to show higher inter-specific divergence in some taxa but is limited by low amplification efficiency in diverse plant groups [71]. ITS2 has been found to excel with its combination of high inter- and intra-specific divergence, making it ideal for species-level identification but encounters amplification challenges in certain plants [72].

For species taxonomic identification of grasses, an ideal barcode combination is one that offers a balance in high resolution at the species level with reliable amplification across various grass taxa and based on the studies that have investigated species within Poaceae through metabarcoding pollen, the best combination includes ITS2 due to its high inter-specific divergence and reliable species-level resolution and *rbcL* for broader taxonomic coverage as it amplifies consistently [73], [74], [75]. This selection aligns with the needs of grass species identification, where closely related species often require precise differentiation.

2.2.3. DNA barcode reference databases

The third critical component necessary for species identification from DNA is the availability of a reliable reference database to match the sequences of the plant barcodes. Reference databases enable accurate taxonomic identification and comparison of sequence data which is critical for biodiversity assessment. Plant DNA reference databases have been developed over the years; however, their reliability vary with the source of the reference sequences. For example, reference databases based on the International Nucleotide Sequences Database Collaboration (INSDC), which includes GenBank, the European Nucleotide Archive (ENA), and the DNA Data Bank of Japan (DDBJ), have been associated with inconsistencies in data quality [76]. One of the main concerns with these repositories is that they often allow sequence submissions without requiring links to voucher specimens or raw sequence data. This lack of quality control can lead to the inclusion of sequences with incorrect species identifications or

low-quality data, ultimately reducing the reliability of the database for accurate species identification [77].

The most used standardised reference databases for *rbcL* and ITS2 barcodes have been developed based on nucleotide information from the Barcode of Life Database (BOLD), which offers higher quality sequences due to its stricter quality control measures [60], [78].

Some of the widely used reference databases in metabarcoding of plants include the work of Sickel for ITS2 sequences and Bell for *rbcL* sequences which both require to be integrated into the bioinformatic pipeline as developed by Sickel [78], [79]. These databases are limited for use within their standardised pipeline. Dubois further built on this work, creating independent references databases based on ITS2 and *rbcL* NCBI sequences for use and application across other pipelines [80].

2.2.4. Bioinformatics pipeline for analysis of sequenced DNA library

The final important element is a bioinformatic pipeline which facilitates the handling and interpretation of the large amounts of sequence data generated in DNA metabarcoding experiments. Several bioinformatics pipelines have been published and found application in the analysis of pollen metabarcoding. These include the Quantitative Insights into Microbial Ecology (QIIME) pipeline that is highly complex and not user friendly to non-bioinformaticians [60], [80]. DADA2 is another pipeline compatible for use in metabarcoding data and its strict model offers high accuracy but is associated with high loss of sequences which often impacts biodiversity assessment [74], [81]. The Ribosomal Database Project (RDP) has also found application in plant metabarcoding, especially after the establishment of ITS2 as a plant barcode. RDP is a tool that is fast and efficient, but its database is primarily focused on microbial sequences, limiting its effectiveness in plant metabarcoding analysis [72], [82]. USEARCH is another tool which particularly clusters sequences into Operational Taxonomic Units (OTUs). While USEARCH has been found to be efficient in metabarcoding, its reliance on fixed clustering thresholds can oversimplify biodiversity by merging distinct species into a single unit, potentially masking species-level differences [78], [83].

2.2. Conclusion and future prospects

Because of our country's unique biodiversity content, a thorough identification approach of the airborne monitored pollen allergen is a necessity to map out and identify all the species present precisely. Metabarcoding presents as such a tool that will enable for the precise species level detection and identification of contributors to pollen allergies that are otherwise difficult to be

identified currently in SAPENT stations. This study was carried on a fraction of SAPNET sites as is a pilot study aimed at developing standardised method that will enable for application of DNA metabarcoding. Its success could potentially lead to its application in most pollen monitoring sites in South Africa enabling us to diversify and identify these species from the various biomes.

2.3. Study aim

The main aim of the study was to identify allergenic pollen of grasses the Poaceae family and the ragweed species of the *Ambrosia* genus. The pollen obtained from the environmental pollen monitoring samples were used for DNA metabarcoding analysis.

2.4. Study objectives

- The isolation of double stranded DNA with high quality and sufficient quantity from environmental pollen samples obtained from the monitored areas rich in grass pollen and sites with *Ambrosia* pollen.
- Amplification of taxonomic gene regions (barcodes) with standard universal primers from DNA extracted from different pollen present in the environmental sample.
- Constructing a comprehensive local refence database of the grass species of the Poaceae family and ragweed species of the *Ambrosia* that are found in the local regions, by means of comparisons of the obtained sequences with the verified barcode sequences.

2.5. Research questions

- To what extent does DNA metabarcoding accurately identify the most prevalent allergenic pollen species within the Poaceae family during high pollen seasons in monitored areas of South Africa?
- To what degree does the quality and quantity of DNA extracted from airborne pollen samples impact the success rate of metabarcoding identification for specific allergenic species?
- Can DNA metabarcoding effectively detect and confirm the presence of *Ambrosia* allergenic species pollen in selected regions of South Africa?
- How reliable is DNA metabarcoding in differentiating between different species of grass and ragweed pollen in complex environmental samples?

- What are the seasonal flowering patterns of allergenic grass pollen, and how do these patterns vary throughout the seasons in correlation with weather patterns and pollen emission periods?
- Can the gaps in the local reference database of grass and ragweed pollen be addressed to improve future identification and monitoring of these allergens through DNA metabarcoding?

2.6. Scope and structure of thesis

The main purpose of this study was to introduce a DNA-based analytical approach to classify and identify pollen from allergenic grasses within the Poaceae family, and *Ambrosia* species to overcome the challenges imposed by the light microscope current method of analysis. This was achieved through the following stages:

2.6.1. Site selection criteria

The study was conducted through analysis of pollen sampled from spore traps placed in six South African cities monitored by SAPNET, that is: Bloemfontein, Cape Town, Durban, Johannesburg, Kimberley and Potchefstroom. The selection of pollen sampling locations for this study was guided by pollen calendars developed using data spanning from 2019 to 2021 [6]. Specifically, these calendars indicated that Bloemfontein, Cape Town, Johannesburg, and Kimberley had the highest annual pollen concentrations of Poaceae pollen during the seasons of data collection. The primary objective was to gain a precise understanding which species were present in different regions of the country during peak grass pollen periods.

To investigate the species of ragweed pollen, we looked at the observation of pollen data from Durban and Potchefstroom sites where *Ambrosia* pollen was identified by light microscopy.

2.6.2. Sampling pollen for DNA analysis

Pollen samples that were used for this study were sampled using Hirst type volumetric spore traps placed at least 3m above the ground. The instrument operates on the principle of impaction and volumetric collection. These traps use a clock driven rotating drum that supports a collection Melenix tape coated with a suitable adhesive to capture airborne particles, including pollen grains. The device continuously aspirates air at the flow rate of 10 litres/minute through an orifice or sampling inlet for air over a period of seven days. At the end of the sampling period the tape collected was longitudinally sectioned to reserve half of the

collected matter while using the other half for microscopic identifications and counting to allow for the quantitative measurement of pollen concentrations.

2.6.3. DNA analysis method overview

Accurate molecular identification of various allergenic grass and ragweed pollen was achieved by exploiting distinctions in their DNA sequences of taxonomic markers. To differentiate between multiple species of airborne pollen particles, often mixed with other airborne particles, we employed metabarcoding. This approach relied on two crucial taxonomic markers: ribulose bisphosphate carboxylase (*rbcL*) and the Internal transcribed spacer 2 (ITS2).

Our methodology involved extracting the DNA from pollen samples collected weekly from the air and subsequently, amplifying and sequencing the taxonomic sequences of the species present in the mixture. The sensitivity of these markers is paramount for the successful identification of all species because by amplifying multiple copies of a species' taxonomic sequences, even those present in low quantities can be accurately identified. The species sequences generated are aligned with corresponding sequences of the species found in the genomic reference database to enable for identifications of the species found in each of the various regions. The overall flow of this work is illustrated in Figure 2.1.

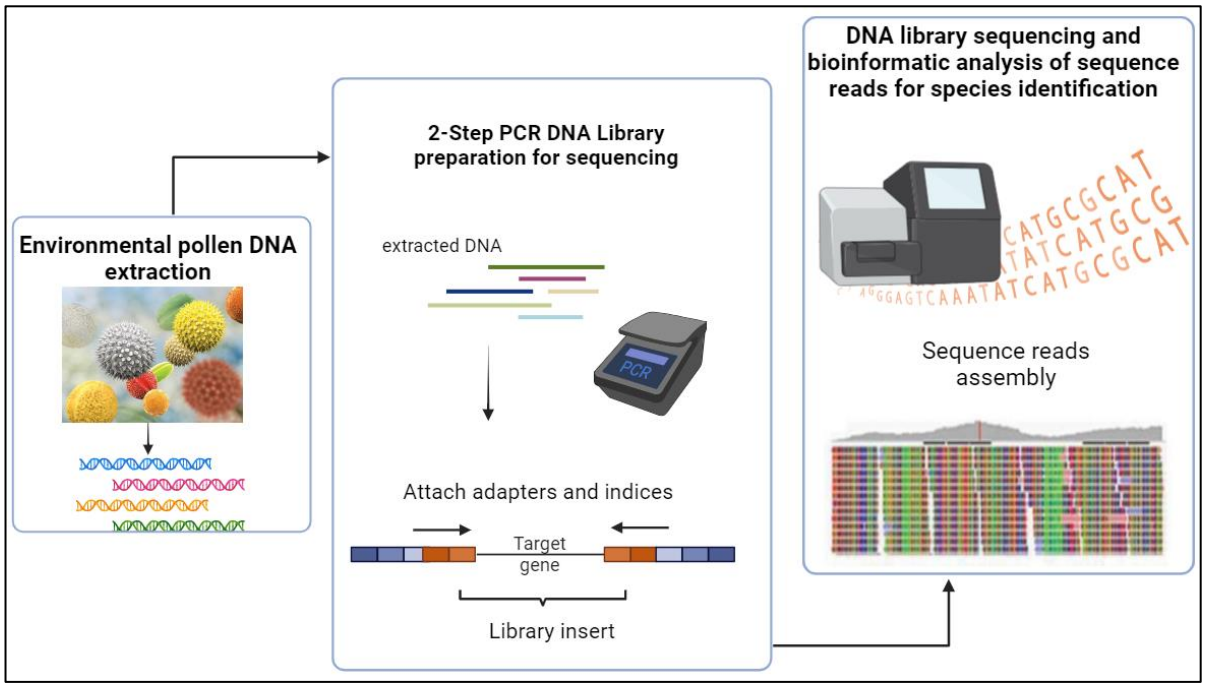


Figure 2.1. Workflow of the DNA metabarcoding methodology. Created with BioRender.com.

2.6.4. Expected outcomes and potential impact of study

This research can be applied to improve public health such as allergen monitoring and management, contributing to early warning systems, aiding in allergy diagnostics, environmental health and the monitoring of allergenic invasive species. Monitoring of grass

and ragweed pollen using DNA metabarcoding can provide valuable information to allergy sufferers and healthcare professionals. Allergen forecasting systems can be developed, similar to weather forecasts, to alert individuals with pollen allergies about high pollen count periods, allowing them to take proactive measures to manage their symptoms.

DNA metabarcoding can enable the early detection of invasive or highly allergenic pollen species, helping authorities implement timely measures to control their spread and mitigate allergen exposure and can improve the accuracy of allergy diagnostics by identifying specific pollen sources responsible for allergic reactions in patients. Healthcare providers can use this information to develop personalized allergy management plans and recommend targeted allergen immunotherapy. The findings from this study have the potential to contribute to future research and it may contribute to a deeper understanding of the seasonal distribution of grass and ragweed pollen. It can aid in identifying trends and correlations between pollen exposure and disease exacerbations, helping researchers and clinicians develop more effective treatments.

Understanding the distribution and dynamics of grass and ragweed pollen through DNA metabarcoding can contribute to assessments of the impact of climate change on allergen exposure. Additionally, it would broaden our understanding of the impact of urbanisation and inform land-use planning to reduce the prevalence of allergenic plants in urban environments. Monitoring the presence and spread of invasive ragweed species through DNA metabarcoding is important for preventing the expansion of allergenic plants into new areas. Integrating the findings of grass and ragweed pollen DNA metabarcoding research into public health strategies requires collaboration among scientists, healthcare professionals, policymakers, and the public and can lead to more informed decision-making, improved allergy management, and ultimately better health outcomes for individuals affected by pollen allergies.

Pollen DNA extraction experiments

3.1. Pollen sampling and material retrieval for DNA analysis

The pollen samples used in this study were collected from environmental monitoring stations across South Africa, trapping pollen using 7-day Hirst volumetric spore traps (Burkard, UK) shown in Figure 3.1(A). At the end of week, each collection tape was halved for analysis with both microscope and DNA. Esterhuizen *et al*. (2023) further explains the preparation tape for microscopy analysis and details the count method for the classification of pollen grains to identify the allergenic pollens in the weekly sample from each region [7]. The half of the collection tapes reserved for DNA analysis were stored in sterile boxes and transported to the laboratory from all collection sites, as illustrated in Figure 3.1(B).

Sample storage method was developed based on methods described by Campbell *et al*. (2022) and Nunez *et al*. (2017) as they used similar samples from a Hirst-type trap for DNA-based analysis [84], [85]. However recent advances in pollen monitoring for DNA barcoding make use of cyclone automated samplers and that has enabled the direct collection of pollen into sterile safe lock tubes simplifying the sample preparation process [58], [86], [87]. The future application of DNA metabarcoding as a method of analysis at SAPNET aims to adopt this method of sampling with cyclone automated samplers.

A. Hirst-type volumetric spore trap

B. Pollen samples preparation for DNA analysis

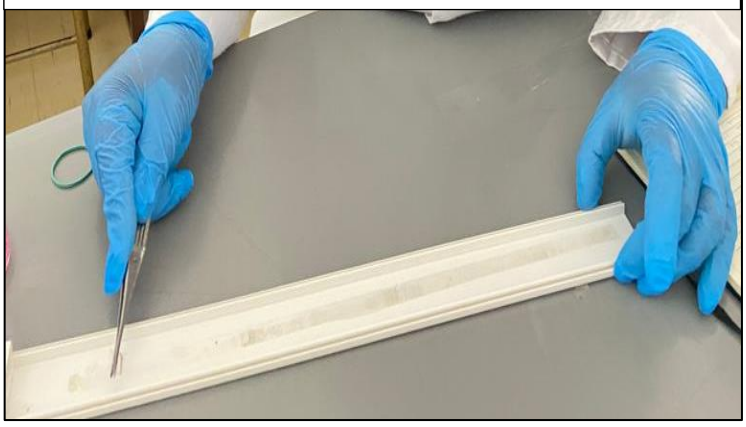


Figure 3.1. (A) Hirst-type volumetric spore trap, Burkard manufacturing (B) Preparation and transfer of pollen collected in Melinex tape for DNA analysis.

Initially, samples were sourced from a library of SAPNET samples collected over the years of which most were tapes sectioned into seven parts and mounted on microscope slides to mimic each day's pollen content in a 7-day week. The process of retrieving the pollen material involved heating the slides to loosen the cover slips mounted with fuchsin, resuspending the tape and cover slit in nuclease-free water, and transferring the solution to 50 ml falcon tubes (Thermo Fisher Scientific (RRID:SCR_008452)). The resuspended tape was centrifuged at 10,000 rpm for 10 minutes and the supernatant discarded leaving a volume of less than 1 ml. The pollen-containing retained solution was transferred to a 1.5 ml nuclease-free Eppendorf tube for DNA analysis. Samples processed using this method included those from Ermelo, Calvinia, Port Elizabeth and Durban.

However, this approach was highly time-consuming, as each DNA sample required a week's worth of material, necessitating the use of seven microscope slides per sample which was not feasible for a larger sample size. Additionally, the process was prone to contamination during various stages of pollen material retrieval. To overcome this, the longitudinal sectioning of the collection tapes was implemented for samples collected from Kimberley, Bloemfontein, Johannesburg, Potchefstroom and Durban for the recent 2022/2023 pollen season.

To further increase sample quantity, an additional spore trap was set up in Cape Town dedicated solely collecting pollen for DNA analysis, providing full tapes of collection for the DNA extraction. The samples used for DNA extraction were recorded for the site they were obtained from, and the period of collection. The summary of all the samples that were used and extracted DNA from is shown in Figure 3.2.

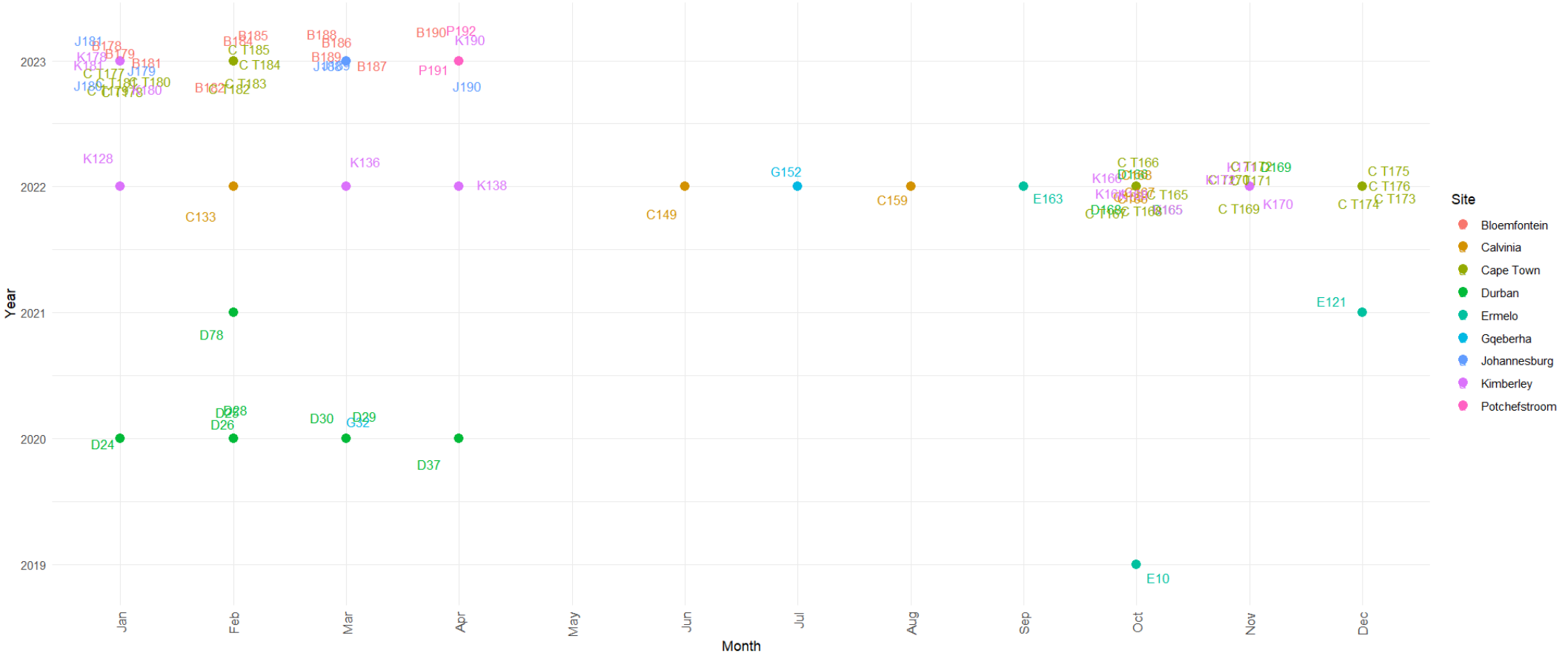


Figure 3.2. Overall samples collected during 2021-2023 pollen season for the DNA extraction, each unique point represents the collection point (i.e., week and year the samples were collected from each site). Created with RStudio [88].

3.2. Isolating DNA from pollen using bead-based DNA extraction technology by with NucleoMag microbiome, Macherey Nagel

Bead-based DNA isolation technology primarily looks at lysing the cells with a mechanical lysis process referred to as bead-beating. The steps that were followed for the extraction can be briefly explained with the steps in Figure 3.3 following the manufacturer's protocol.

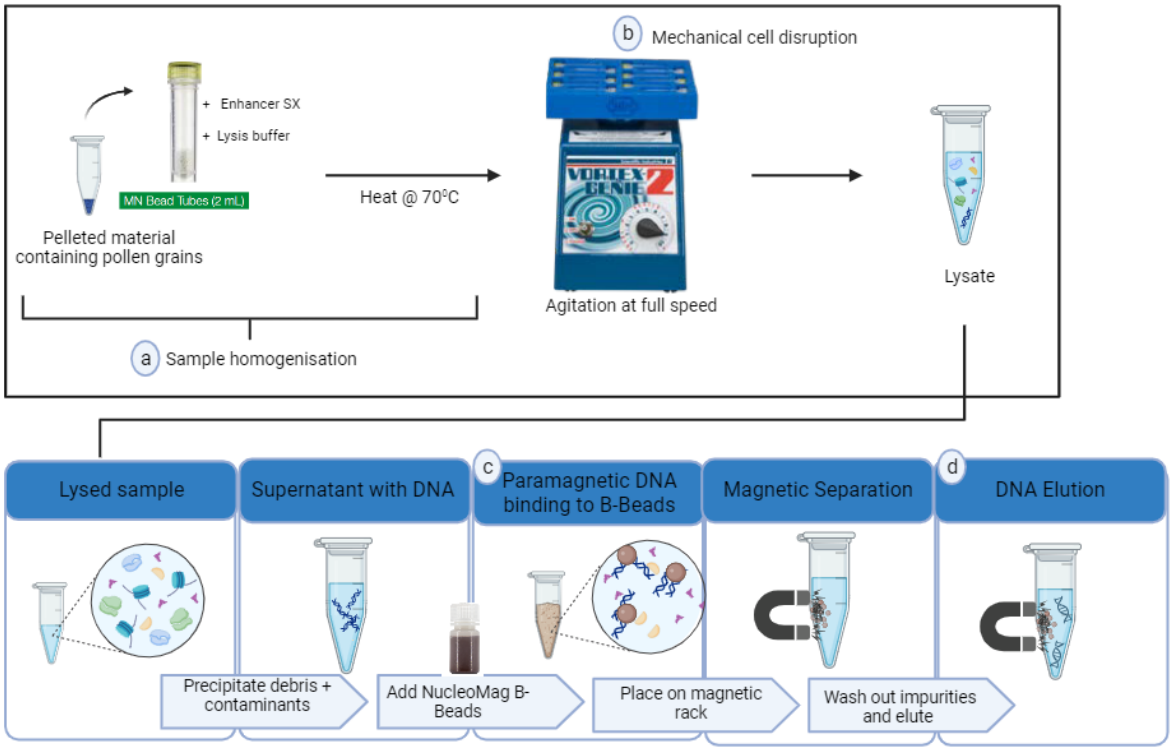


Figure 3.3. Summary of the bead-based DNA extraction process with the NucleoMag Microbiome kit. The key steps highlighted a) sample homogenisation, b) mechanical cell disruption, c) DNA binding, and d) DNA elution, were optimised for a higher DNA yield (Macherey–Nagel, Düren, Germany).

In the first step, up to 500 µl solution containing pollen was transferred to 2 ml MN bead tubes with 0.6 - 0.8mm ceramic beads Type A (Macherey–Nagel, Düren, Germany), and added lysis MI1 buffer and enhancer SX solution as per manufacturer's instruction. The mixture was heated in a 70°C water bath to homogenise the solution and activate the chemical lysis activity. Post heating the tubes were placed on a rubber foam adapter on a vortex and vortexed at full speed to mechanically lyse the grains and after, the lysate was transferred to another Eppendorf tube.

The basic principle behind this reaction is that the main substance of the lysis buffer, that is, sodium chloride (NaCl) dissociates to form sodium and chloride ions that help maintain the ionic strength and osmotic balance of the solution. Since the DNA molecule consists of a highly

negative phosphate backbone, the sodium ions helped to shield the negative charges preventing the strands from repelling each other and the presence of this salt also helps to reduce the solubility of proteins, resulting in their aggregation and precipitation from the solution. Ethylendinitrilo tetraacetic acid (EDTA), a substance contained in the enhancer SX solvent, is a chelating agent that binds metal ions, that is, calcium ions i.e., Calcium ions (Ca^{2+}) released by cellular components, is as an inhibitor that prevents them from acting as cofactors for DNA degrading enzymes (DNases).

The lysate constituting of DNA, cell debris and particulate matter collected along with the pollen was precipitated with the addition of MIC buffer and the solution incubated on an ice bath of 2-8°C for 10 minutes. The centrifugation of the mixture pellets the cell debris and protein bound contaminants leaving behind the supernatant containing DNA.

The supernatant was transferred, and the precipitant discarded. For every 500 µl DNA solution, 8.3% v/v binding buffer (NucleoMag B-beads/MI2) was added. The ethylene glycol diacetate solvent present in the mixture aids in coating the magnetic B-beads to promote the binding of DNA. The protocol suggests that the reaction occurs instantaneously as the DNA containing solution is introduced in the solution of B-beads, but this is the overall rate determining step.

This was followed by several wash steps: first being a wash with MI3, a triethanolamine containing buffer that helps solubilise all the other insoluble material surrounding the DNA bound immobilised beads, this was followed by two consecutive wash steps with MI4 buffer containing guanidine hydrochloride compound that is a chaotropic protein denaturant to ensure that the DNA bound beads are rid of protein contaminants.

A final wash with 70% ethanol to help aggregate the DNA molecules prior to aid in preventing degradation in storage. The last step of the reaction is the release of the DNA from the b-beads, and this was achieved by suspension of the DNA-bead complex in MI5 elution buffer and stored at -18°C.

The chemical compositions and concentrations of the buffers and solutions that were used in the extraction reaction are explained in the NucleoMag microbiome manual. The dsDNA yield that we obtained from this extraction was unsatisfactory and below the required quantity for subsequent analysis, therefore, this method was optimised.

3.3. DNA extraction method optimisation

A series of optimisation experiments were conducted using samples with moderate to high grass pollen concentration ranging from 10 - 87 pollen grains per cubic meter (pg/m^3). Each key step in the DNA extraction process was extended to evaluate the effects of increasing reaction durations as summarised in Table 3.1. For optimising based on sample homogenisation, times were lengthened from 5 to 10, and 20 minutes to improve solution mixing (standard, Reaction 2, and Reaction 8). For optimising focusing on mechanical cell lysis (bead-beating) times were extended from 10 to 15 and 30 minutes (standard, Reaction 2, and Reaction 9) to ensure complete cell disruption. DNA binding duration was increased from 5, 10, 20 and 30 minutes to find the optimal binding period of DNA to paramagnetic B-Beads (standard, Reaction 3, Reaction 10 and Reaction 12). Finally, DNA elution times were extended from 5 up to 60 minutes (standard, Reaction 4, Reaction 11, Reaction 14 and Reaction 15) to maximise DNA recovery. These optimisations helped identify the ideal protocol for maximising DNA yield when extracting DNA using Nucleomag bead-beating method with paramagnetic beads.

3.4. DNA quantity assessment

A high sensitivity Qubit 3.0 (Invitrogen). fluorometer assay was used to quantify DNA extracts and measure the yield of extracted double stranded (dsDNA). The Qubit assay is highly sensitive and can detect from as little as but not less than 0.005 ng dsDNA in a 1 μL solution [89].

3.5. DNA quality assessment

A spectrophotometer was used to measure the purity of the samples extracted. Ultraviolet (UV) absorbance technology is used mainly in the detection and reporting of the impurities of the samples. The purity of the samples was checked using this instrument, which gives the absorbance purity ratios at 230 and 280 nm of solution against the standard absorption wavelength of 260 nm for DNA. An indication of a good quality DNA samples is the purity ratios of A230/260 and A280/260 at a range of 1.8 - 2.2 and 1.7 - 1.9 respectively. A value obtained outside of this range is an indication of impurities such as organic compounds resulting from a carryover of carbohydrates, protein contaminants, and residual phenol from buffers and solutions used in the extraction process [51], [90].

3.6. Sample purity optimisation

Enzyme treatment was incorporated into the sample preparation step of the DNA extraction process to improve the purity of DNA extracted [91]. The DNA samples obtained were further purified with a PCR inhibitor removal kit prior to PCR.

3.6.1. Enzyme treatment with Proteinase K and RNase A

The concept of enzyme treatment, as proposed by Leontidou *et al*. 2019 [53], was considered to achieve the purest form of the samples after extraction. In the pursuit of enhancing the purity of the samples, enzyme treatment was integrated into the optimisation experiments, beginning at Reaction 10 in Table 3.1. During the sample preparation phase for DNA isolation, the samples underwent treatment with Proteinase K and Ribonuclease A (RNase A) (Invitrogen).

Proteinase K, a serine protease known for its potent enzymatic capabilities, excels in breaking down and removing proteins. Lysis reagents contain denaturing agents that effectively unravel contaminating proteins, making them more susceptible to digestion by Proteinase K. To implement this process, Proteinase K was introduced into the sample preparation solution at concentrations ranging from 50 to 1000 µg/mL. Its presence quickly deactivated DNase, effectively preventing unwanted DNA degradation during the sample homogenisation phase.

RNase A was also added at concentrations ranging from 1 to 100 µg/mL, serving a critical role in the elimination of ribonucleic acids (RNA) or potential RNA contamination. It should be noted that the generative nucleus of pollen cells has been suggested to participate in RNA synthesis, implying that among the cellular components released during lysis, there might be traces of ribonucleotides. RNase A helped prevent RNA contamination by absorbing or digesting these ribonucleotides present in the solution ensuring the purity and integrity of the DNA sample for subsequent analyses.

3.6.2. Purification for PCR inhibitors removal

Following the DNA extraction process, the samples were subjected to an additional purification step to thoroughly eliminate any potential PCR inhibitors. To achieve this, all samples were purified using a OneStep PCR Inhibitor Removal Kit (Zymo Research), following the manufacturer's recommended instructions. This kit uses column matrix filtration technology, allowing for the efficient removal of various naturally occurring compounds, such as polyphenols, humic and fulvic acids, and tannins, which are typically found in plant materials.

3.7. Results

3.7.1. DNA yield optimisation results

The primary objective of the optimisation reactions was to manipulate the reaction times for each of the key steps to obtain a larger DNA yield. Although this helped us identify the limits up to which the method could be pushed to attain a high yield from samples of this nature, it came at a cost of sacrificing or preventing 31% of the total samples from further progressing to downstream analysis. The effect of the changes that took place is summarised in Table 3.1, and the trends and effect of changes drawn in Figure 3.4 based on the DNA yield obtained with each change.

Table 3.1. Reaction conditions manipulated in the optimisation experiments aimed at evaluating the effect of lengthening the reaction periods on DNA yield from the four major steps of the bead-based DNA extraction (Nucleomag kit, Macherey-Nagel).

Reaction	Sample homogenisation (min)	Mechanical lysis (min)	DNA binding (min)	DNA elution (min)	Cumulative processing (min)	dsDNA concentration (ng/µL)
Standard	5	10	5	5	25	0,830
Individual step optimisation						
1	10	10	5	5	35	5,840
2	5	15	5	5	30	3,350
3	5	10	10	5	30	4,120
4	5	10	5	10	30	4,630
Combinatorial steps optimization - change applied to subsequent step						
5	10	15	5	5	35	6,570
6	10	15	10	5	40	5,200
7**	10	15	10	10	45	17,600
Combinatorial change to subsequent step (double the period)						
8	20	15	10	10	55	0,170
9	20	30	10	10	70	<0,005
10	20	30	20	10	80	0,560
11	20	30	20	20	90	0,570
Lengthen DNA binding and elution periods						
12	10	15	30	10	65	7,060
13	10	15	30	20	75	12,300
14	10	15	30	30	85	15,200
15	10	15	30	60	115	20,600

The numbers in red are the changes in time applied to each optimisation reaction.
**Indicates the newly established standard reaction conditions to which the subsequent optimisations were based on.

Increasing homogenisation time initially improved yield but reached a plateau at 10 min, with yields decreasing at longer durations. Extending lysis time to 15 minutes improved DNA

extraction efficiency, but 30 minutes was damaging. Increasing binding time to 30 minutes maximised DNA concentration when combined with a moderate homogenisation and lysis time. Extended elution times consistently improved DNA concentration, with the highest yield at 60 minutes, and therefore rest of the samples that were extracted for the course of this study followed this reaction conditions with the processing time approximated at 115 minutes.

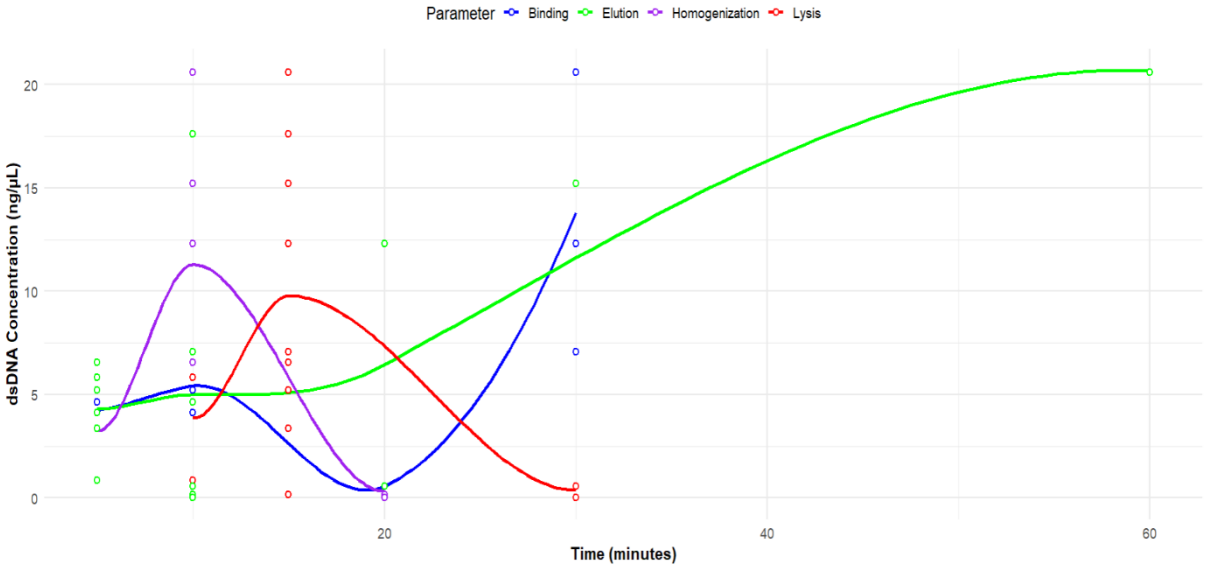


Figure 3.4. Trends observed in the DNA yield with the optimisation reaction for each of the parameters i.e., lengthening of sample homogenisation, mechanical lysis and DNA binding and elution periods investigated. Created with RStudio [88].

Optimising DNA yield required extending the extraction duration to five times longer than the manufacturer's recommended protocol. However, this was not considered a limitation, as the extended process was properly planned and carried to ensure consistent and thorough extraction.

3.7.2. Sample purity results

Although there was a slight improvement in the purity of the samples with the addition and incorporation of RNase A and Proteinase K enzymes during extraction the samples were further purified post extraction. The purity post extraction was compared with the purification profile post the additional purification step as shown in Figure 3.5. Following the extractions, only samples that had low purity based on A260/280 ratio aligned to the optimal range indicative of better quality. However, only one sample (Cape Town, week 174) had optimal purity that aligned for A260/230 ratio post extraction, but they all further declined post additional purification.

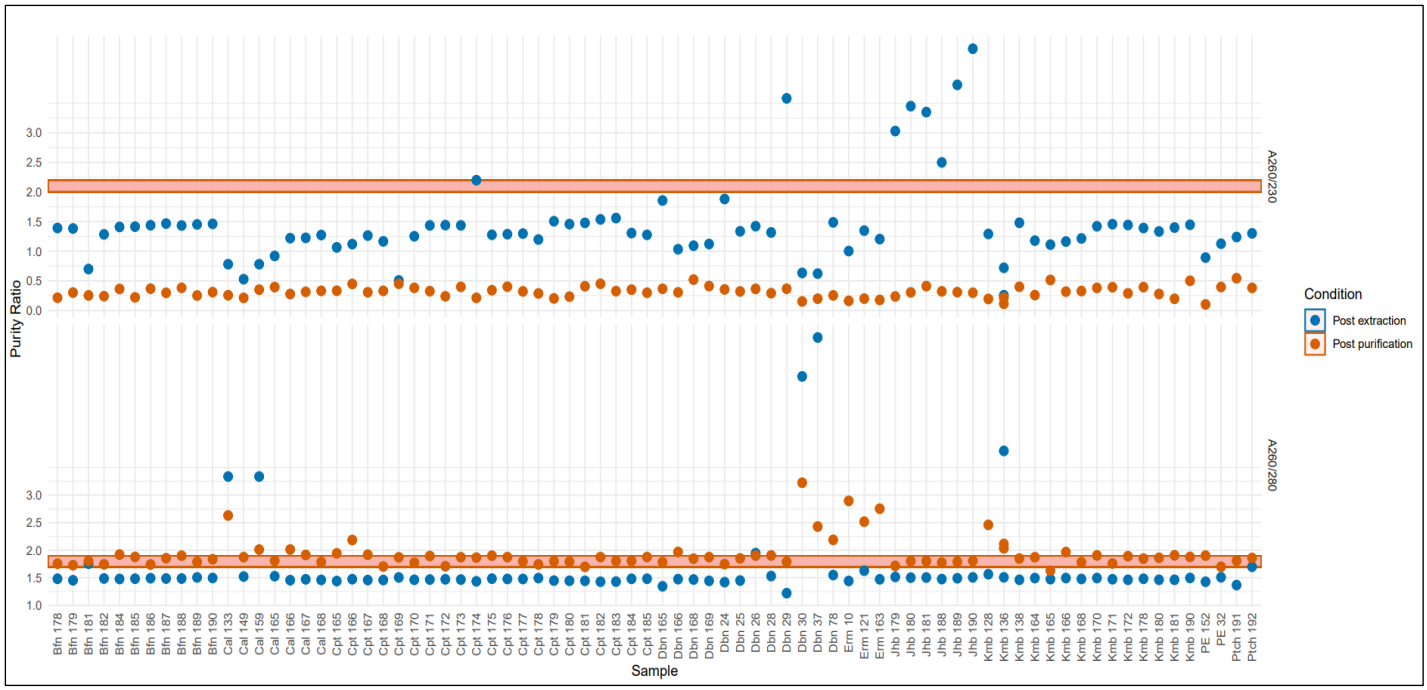


Figure 3.5. Purity ratios of all the samples measured after extraction, and after additional purification, the red line indicates the ideal purity range the samples expected to fall into. Created with GraphPad Prism version 10.0.0.

3.7.3. Correlation analysis

To determine whether there was correlation between the concentration of DNA obtained from the sample and the Poaceae/*Ambrosia* pollen concentration, we ran a correlation analysis using the concentration values of dsDNA and grass/ragweed pollen concentrations quantified with light microscope.

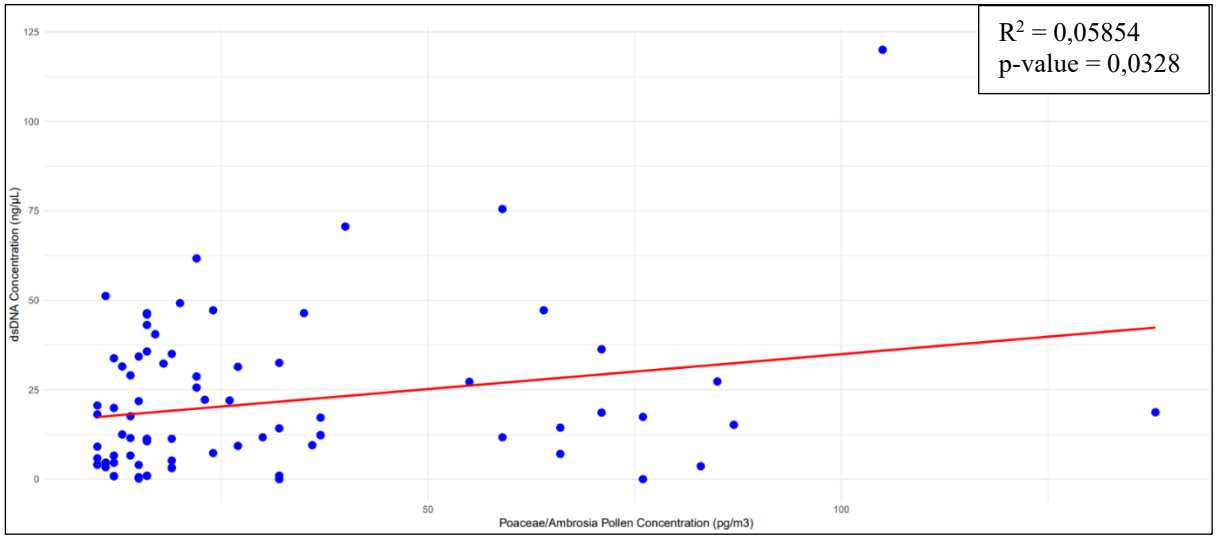


Figure 3.6. Correlation analysis performed indicating statistically significant correlation in the dsDNA obtained with respect to the targeted pollen contained. Created with RStudio.

The pollen concentration that was used for correlation focused solely on grasses or ragweed which limited the scope of our analysis since the DNA from several other plants' pollen in the samples may also have been extracted. Because the 'extremely high count' samples did not necessarily produce 'extremely high dsDNA concentration', there was no positive correlation when looking at the two variables. There was no clear indication of the influence of the targeted pollen quantity to the final DNA yield obtained as evident by $R^2 < 95\%$.

Despite this, correlation analysis (p-value < 0.05) revealed a positive and statistically significant relationship between grass/ragweed pollen grain concentration (quantified microscopically) and the dsDNA concentration. Although the R^2 value indicated that only 5.85% of the variability in dsDNA could be explained by the abundance of pollen, the statistically significant p-value suggest that this relationship, while modest, is meaningful. This emphasises the relevance of pollen grain concentration as a possible predictor of the amount of DNA extracted.

3.7.4. Summary of samples extracted

dsDNA was extracted from 79 environmental pollen samples with varying grass and ragweed pollen concentrations. Table 3.2 provides detailed information on the extraction sequence and the identity of each sample, including the monitoring site, collection time point during the season (week number), and grass/ragweed pollen concentrations. Additionally, the table tracks the DNA yield and purity for each sample, highlighting the progression from the initial extractions using the standard protocol to the final samples extracted.

Table 3.2. Summary of pollen samples extracted according to the number of samples obtained from each site with their corresponding pollen concentration and the interquartile range of the concentration of DNA extracted.

Site	Number of samples	Mean concentration (pg/m^3)	Mean dsDNA (ng/µL)	Median dsDNA	Max dsDNA
Bloemfontein	11	23.82	8.08	4.60	22.00
Calvinia	7	18.00	22.50	28.70	51.20
Cape Town	21	34.43	33.57	32.50	120.00
Durban	12	23.67	20.30	17.40	46.40
Ermelo	3	37.33	9.56	0.83	27.30
Johannesburg	6	51.50	11.31	12.00	18.70
Kimberley	15	30.67	21.27	17.40	61.70

Port Elizabeth	2	15.50	5.88	5.88	6.57
Potchefstroom	2	46.50	22.10	22.10	25.60

Out of the 79 samples from which DNA was extracted, 6 were from sectioned tapes mounted on microscope slides, 52 from longitudinally halved tapes, and 21 were full tape samples obtained from the Cape Town Spore trap that was specifically operated for DNA analysis.

Post extraction, all the samples with a yield less than 1 ng/µL were excluded from further analysis. From the total of samples with pollen differently sourced, the rate of successful DNA extraction with adequate DNA was 79% from longitudinally halved tapes, 95% from full tapes, and 17% from sectioned tapes mounted on microscope slides. A total of 17 samples was excluded based on A260/280 optimal purity ratio as shown in Figure 3.6.

Because of high sensitivity of the subsequent PCR reactions to the presence of protein contaminants in DNA samples, priority for assigning optimal purity was based on the A280/260 ratio as a key indicator of ideal purity to ensure the suitability for further processing.

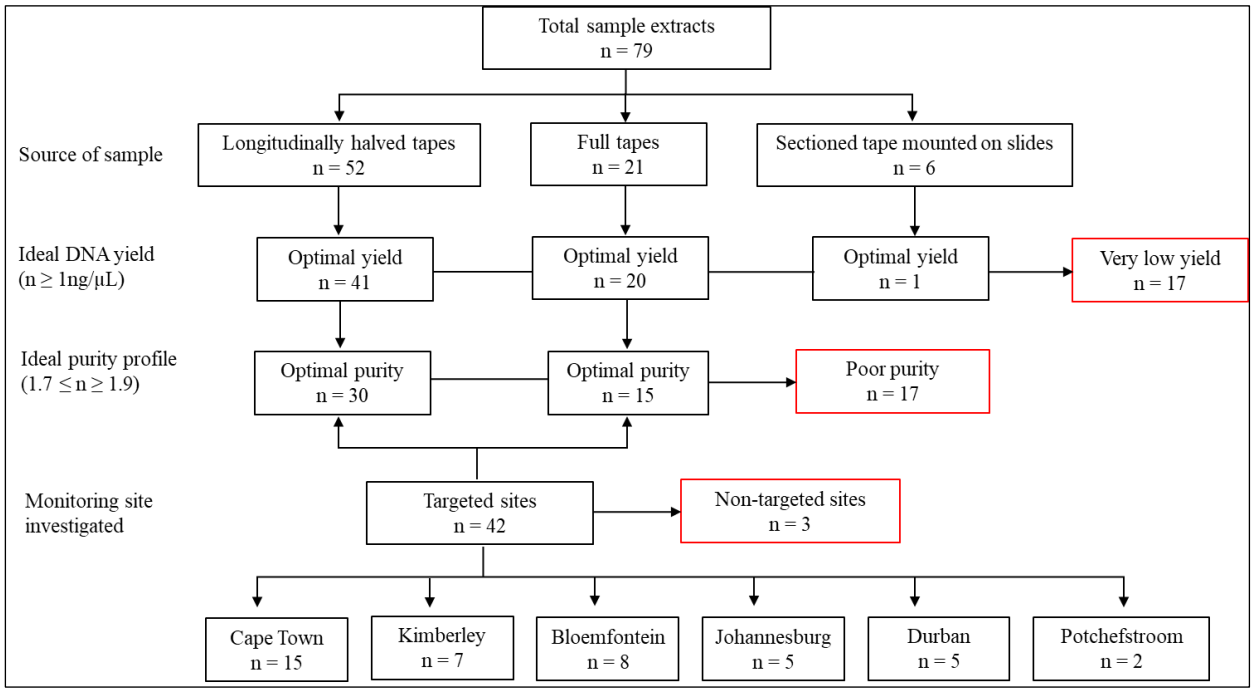


Figure 3.6. Workflow for sample exclusion following DNA extraction. The 79 sample extracts were obtained from pollen collected on Melinex tapes, which were either longitudinally halved, left intact (full), or mounted on slides. Exclusions (Outlined in red) were based on inadequate DNA yield, suboptimal purity, or non-targeted collection sites.

3.8. Discussion

The optimisation of the pollen DNA extraction process aimed to improve the bead-based method using the Nucleomag kit, following the manufacturer's recommended instructions. Given that this method was found to be most useful when extracting DNA from environmental pollen samples, this effort focused on improving both the yield and quality of the extracted DNA. It was essential to refine the method to maximise DNA recovery due to the complexity of our pollen monitoring samples, which often is collected along with other biological matter and low quantities of airborne particles.

3.8.1. Double stranded DNA yield optimisation

The strategy that was used to improve the yield, primarily focused on altering the duration of the key steps. Initially, these steps were individually altered to assess their impact on the final yield of dsDNA. In the individual step optimisation approach, each step was extended by 5 minutes. The initial optimisation, which featured an extended homogenisation period, immediately yielded a noteworthy improvement, that was a seven-fold increase in dsDNA concentration compared to yield of the standard reaction. Similar patterns of increased dsDNA concentration were observed when lysis, binding, and elution periods were extended. While the observed increases in the first four reactions did not follow a specific pattern across the evaluated key steps, they provided valuable insights when comparing with the standard reaction conditions. These reactions were performed individually with different samples, making it challenging to draw direct comparisons and determine the optimal conditions. As the achieved yield was still below 10 ng/µL, further adjustments were made through combinatorial optimisations.

The combinatorial optimisation approach aimed to incorporate changes introduced in one reaction to subsequent reactions. The resulting three reactions of this approach (Table 3.1, reactions 5-7) revealed a synergistic positive impact leading to a remarkable increase in yield 21-fold over the standard reaction. Additional experiments were conducted to further decide on the best combination of conditions to maximise the yield.

In this second set of combinatorial optimisation reactions, the period lengths were doubled compared to the "newly established" conditions in Table 3.1, reaction 7. This included a homogenisation period of 20 minutes, which was the longest duration among all the reactions. This extended period allowed the pollen material to remain in a higher-temperature environment within a chemical lysis setting for an extended duration. This prolonged exposure had a significant impact on the yield, which saw a drastic reduction, falling well below the

yield achieved in the standard reaction. This effect was particularly pronounced when mechanical lysis was also increased. Indeed, in addition to the extended application of heat, the grains underwent an intensified bead-beating process that not only damaged the cell walls but also disrupted the cell contents, consequently affecting the DNA integrity.

As extending the durations of the first two steps, namely homogenisation and mechanical lysis, had adverse effects on the yield, these conditions were reverted to the initial periods, where yields had improved by the additional 5 minutes from the original reactions. The focus then shifted to evaluating the extension of the binding and elution periods in reactions 12 to 15 (as outlined in Table 3.1). Each reaction with prolonged binding and elution periods demonstrated a subsequent increase in DNA yield. When reaction 15 yielded a substantial 20 ng/µL from a sample with 10 grass pollen grains per cubic meters, the optimisation process was concluded. Using the conditions of reaction 10, the Standard Operating Procedure (SOP) was subsequently formulated and used for the extraction of 64 pollen DNA from samples.

3.8.2. Sample purity optimizations

Another important factor that significantly influenced the reactions was the acquisition of samples with the desired purity profile, a prerequisite for successful PCR reactions. Following the initial sample extraction under recommended standard conditions, the purity appeared promising, with ratios falling within the desired range. However, as subsequent reactions were conducted under altered conditions, the purity ratios declined. This indicated that prolonged exposure of the pollen samples to these modified chemical environments resulted in an increased presence of contaminants within the DNA. The most notable condition change that had a substantial impact was the extended homogenisation period, which caused the A260/280 ratio to double beyond the expected range (i.e., A260/280 > 3.8). The higher concentration of proteins relative to nucleic acids possibly from the detection of phenols and residual reagents used during the extraction process suggested protein contamination.

Consequently, as of reaction 10 of the optimisation reactions in Table 3.1, enzyme treatment became a necessary addition. This involved the incorporation of Proteinase K and RNase A during the sample homogenisation step, following the recommendations of Hawkins *et al*. (2015) and Lowe *et al*. (2022) [51], [90]. The enzymatic digestion improved the ratios in the subsequent 68 reactions, resulting in an average A260/280 ratio of 1.49 and an A260/230 ratio of 1.54. These lower values indicated the presence of contaminants such as guanidine, EDTA, carbohydrates, lipids, salts, or phenols [92]. Although the improvement was slight it was not

negligible, therefore enzyme treatment was subsequently integrated into the extraction protocol.

The notable outliers within the purity ratios for A260/280 ratios in Figure 3.4 post extraction, originated from samples of the optimisation trials, whereas for the A260/230 ratio the outliers were a combination of both the samples extracted during the optimisation trials and samples extracted with optimised protocol. Despite this, the post extraction purification process successfully eliminated protein contaminants as evident in A260/280 purity. However, this concurrently introduced of additional salts and chelating agents, such as EDTA from the reagents that negatively impacted the A260/230 purity which was discernible through the decline in ratios for all the samples.

The additional purification steps coupled with the extension of DNA extraction durations extended the overall process of obtaining cleaner DNA samples from environmental pollen sources. This added time and effort proved to be a valuable investment when considering the improved yield and improved DNA quality obtained at the end of the overall process.

Amplification of taxonomic barcodes

4.1. Introduction

The second objective of this study was to selectively amplify taxonomic DNA regions (barcodes) from the extracted environmental DNA of pollen. This focused on developing an effective PCR method to amplify target regions in samples with low DNA yields and aimed to evaluate the applicability of universal primer sequences for amplifying *rbcL* and *ITS2* barcodes, specifically for differentiating Poaceae and *Ambrosia* taxa, to confirm whether the amplified products correspond to the fragment sizes reported in the literature.

4.2. Polymerase Chain Reactions (PCR)

PCR works by cycling through thermo-specific reactions to amplify a specific DNA region. It begins with denaturation, where the double-stranded DNA is separated into single strands at high temperatures around 90°C. The temperature is then lowered to allow primers to anneal to their complementary sequences on the template. Subsequently, the temperature is raised to activate DNA polymerase, which synthesizes new DNA strands by incorporating nucleotide bases (dNTPs). This denaturation, annealing, and extension cycle is repeated 25 - 35 times, exponentially increasing the target DNA with each cycle. The success of PCR relies on precise reaction conditions, including optimal primer binding, DNA polymerase activity, and accurate thermal cycling [7].

Each component has an important role in the reaction; DNA template provides the target sequence, while dNTPs serve as the building blocks for new strands, primers define the region to be amplified, and the thermostable enzyme DNA polymerase facilitates the strand synthesis. A buffer containing essential ions, such as $MgCl_2$ or KCl, ensures enzyme activity and promotes primer annealing. It is important to ensure that there is balance among these components as improper concentrations can compromise efficiency or amplification completion. While template DNA and primer concentrations can be adjusted during reaction setup following standard operating procedures, the choice of DNA polymerase is influenced by the template concentration and commercial kits available [93].

4.3. Selected taxonomic barcodes strengths and limitations

4.3.1. The *rbcL* barcode

The *rbcL* barcode has played an important role in barcoding as it is one of the early taxonomic markers to be established and used for phylogenetic construction studies since the early 1990s [94]. Although the *rbcL* barcode has been widely used, it has been noted for its lower resolution in discriminating at the species level, and sometimes even at the genus level [55]. This limitation is attributed to its selection from a highly conserved region within the gene. The high conservation makes it difficult to amplify regions that reflect the variability needed to differentiate closely related species [63]. Therefore, in species identification, especially within genera with greater variability *rbcL* may show lower discriminatory power compared to other molecular markers or barcodes [55]. Despite that, *rbcL* maintains its status as a core barcode owing to its widespread presence in the plant kingdom and offers high primer universality and sequence quality [95], [96]. These advantageous characteristics make it a preferred marker for metabarcoding analyses over the other chloroplast barcodes.

The size of the *rbcL* DNA sequence barcode may vary from species to species. For instance, Zahra *et al*. (2016) reported that the *rbcL* fragment size ranged between 576-1445 bp while Brennan *et al*. (2019) reported the sequences to vary between 702-883 bp [86], [97]. Although the obtained sequence length may vary across different plant groups, the CBOL Plant Working Group has advocated for a standardized region known as *rbcLa*. This region is 650 bp long and represents a fragment from the 5' end of the *rbcL* gene, serving as the recommended segment for consistent use in barcode analysis [45].

4.3.2. The ITS2 barcode

The ITS region, consisting of ITS1 and ITS2 subunits within the nuclear genome, was initially designated as a core barcode for plants due to its acknowledged high-resolution capabilities in both inter- and intraspecific discrimination. However, challenges arose, particularly related to intragenomic variation in ribosomal DNA and technical issues, compromising the reliability of the ITS barcode. The widely used primer pair, ITS1 and ITS4, originally designed for fungal amplification by White *et al*. in 1990, proved non-specific to plants, resulting in non-targeted amplicons and diminishing the utility of ITS in-plant community barcoding [98].

Recognising the limitations of the standard ITS primers, Chen *et al*. (2010) proposed plant specific ITS primers designed to amplify the ITS2 region [46]. This adjustment aimed at improving plant coverage and reducing co-amplification of fungal material. The ITS2 barcode, with its high species resolution and discriminatory capabilities, emerged as the equivalent of

the animal Cytochrome c Oxidase subunit I (COI) barcode for plant barcoding. As a result, several ITS2 primer sequences have been developed. Table 4.1 looks into primers that have been used for metabarcoding studies using pollen for diversity investigations including studies on plant-pollinator interactions, palynology, plant surveys and honey analyses.

Table 4.1. ITS2 primers that have been widely used in plant biodiversity studies.

Primer	Orientation	Sequence (5' – 3')	Size (bp)	Papers that have used the primers
ITS-S2F	Forward	ATGCGATACTTGGTGTGAAT	350-460	Chen (2010), Sickel (2015), Baensch (2020), Bell (2017), Bell (2019), Biella (2019), Fahner (2016), Froslev (2017), Khansaritoreh (2020), Kuzmina (2018), Lim (2018), Nürnberger (2019), Ortega (2020), Prosser (2017), Richardson (2015), Richardson (2015), Richardson (2019), Sickel(2015), Swenson and Gemeinholzer (2021), Voulgari-Kokota (2019), Wilson (2021)
ITS4	Reverse	TCCTCCGCTTATTGATATGC		White (1990), Baensch (2020), Bell (2017), Fahner (2016), Froslev (2017), Gous (2019), Gous (2021), Khansaritoreh (2020), Kuzmina (2018), Lim (2018), Nürnberger (2019), Prosser (2017), Richardson (2019), Sickel (2015), Swenson and Gemeinholzer (2021), Voulgari-Kokota (2019), Wilson (2021)
BEL-3	Reverse	GACGCTTCTCCAGACTACAAT	350-490	Chiou (2007), Chen (2010), Bell (2017), Bell (2019), Kuzmina (2018), Ortega (2020), Richardson (2015), Richardson (2015)
ITS-u2_F	Forward	GAAYCATCGARTCTTTGAACGC	265	Banchi (2020), Cheng (2016)
ITS-p4	Reverse	CCGCTTAKTGATATGCTTAAA		
ITS2F	Forward	AYGACTCTCGGCAACGGATATCTTGG	370	Coghlan (2020)
ITS2R	Reverse	CCCAVGCAGRCDTGCCC		
ITS3	Forward	GCATCGATGAAGAACGCAGC		White (1990), Gous (2019), Gous (2021)

Despite its advantages, ITS2 showed difficulties in identifying species across all plant taxa, which led to the realisation that its combination with *rbcL* could offer a more comprehensive solution [99]. The pairing of *rbcL* and ITS2 was identified as an excellent combination and is now acknowledged as the preferred standard barcodes for the precise identification of pollen species [53], [100].

4.4. PCR method

Two high fidelity polymerases were explored for amplification reactions using the primers detailed in Table 4.2. In the initial attempts at PCR of samples extracted, a Phusion Hot Start II High Fidelity Polymerase Master Mix (Thermo Scientific) was used. However, this was later changed to Q5 Hot Start High Fidelity 2× Master Mix (New England Biolabs) to improve amplification congruent with the sample concentration requirements.

4.4.1. Amplification with Phusion Hot Start II High Fidelity Polymerase

For each 50 µL reaction mixture, the following components were combined: 25 µL of 2× Phusion Master Mix (containing 1 unit of Phusion DNA Polymerase, 1.5 mM $MgCl_2$, and 400 µM of each dNTP), 0.5 µM of both forward and reverse primers, a minimum of 10 ng/µL of DNA template, and nuclease-free water up to the final volume. PCR thermocycling was carried using the manufacturer's recommended thermal and cycle conditions and its success evaluated with agarose gel electrophoresis.

4.4.2. Amplification with Q5 Hot Start High Fidelity 2× Master Mix

The composition of the Q5-based PCR was carried with 1× concentration of the Q5 Master Mix (including 1 unit of Q5 DNA polymerase, 2.0 mM Magnesium ions (Mg_2^+), and 400 µM of each dNTP), 0.5 µM of each primer, nuclease-free water, and at least 1ng/µL of the dsDNA template in a final reaction volume of 50 µL.

In the preparation of the PCR, all the components were thawed and combined in ice to maintain the stability of the polymerase, while working under UV light that ensured a sterile environment to prevent potential contaminants. For each reaction that was run, a negative control reaction containing all the components except for template DNA was included to ensure the components were contaminant free. PCR thermocycling was carried using Q5 DNA polymerase manufacturer's recommended thermal and cycle conditions and its success evaluated with agarose gel electrophoresis. The resulting amplicon sizes of the barcodes were compared against the sizes of the barcodes reported in literature.

In silico simulations were performed to predict if the amplified sequences correspond with the expected product sizes for Poaceae grasses, thereby confirming the specificity and effectiveness of the primer sequences. The primers' annealing temperatures were optimised to obtain the amplicons within the specified ranges and later the ITS2 barcode primers changed to use the primer pair sequences recommended by Prosser and Herbert [96].

Table 4.2. Primer sequences used in PCR 1 for amplification include Illumina adapters (underlined sequence) and forward sequences include 6N (highlighted bold).

Barcode	Primer name	Sequence (5' - 3')	T_a(°C)	Amplicon size (bp)
rbcL	*rbcL*aF	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**NNNNNN**ATGTCACCACAAACAGAGACTAAAGC	55	702 - 883 [8], [9]
	*rbcL*506R	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGGGACGACCATACTTGTTCA		
ITS2	ITS-S2F	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**NNNNNN**ATGCGATACTTGGTGTGAAT	50	163 - 311 [8]
	ITS3R	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGACGCTTCTCCAGACTACAAT		
ITS2	ITS-S2F	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**NNNNNN**ATGCGATACTTGGTGTGAAT	53	350 [10], [11]
	ITS4(R)	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTCCTCCGCTTATTGATATGC		

4.3.3. Post-PCR purification and quantification

The amplified PCR products were purified utilising Agencourt AMPure XP beads (BECKMAN COULTER, Life Sciences). This process involved a bead solution ratio of 64.3% v/v for optimal binding of the dsDNA. Post-purification, the concentration and quality of the amplified dsDNA were quantitatively assessed using a Qubit High Sensitivity (HS) DNA Assay and further analysed for size distribution by running a High Sensitivity D1000 Tape Station Assay. These methods provided precise quantification and quality assurance of the purified PCR products before subsequent processing.

4.5. Results

4.5.1. Comparative results of PCR amplification using two polymerase enzymes

Amplification of the targeted regions was run over a Gradient PCR producing amplicons (PCR products) at varying sizes visualised with agarose gel electrophoresis measuring the band sizes against the DNA ladder with sizes varying from 100 - 1000 bp. The intensity of the band visualised in the gel represent the concentration of the amplicons (amplified product).

PCR was run with both polymerase for both barcodes and repeated with two separate samples as shown in Figure 4.1. We compared the intensity of the amplicons on agarose gel electrophoresis to evaluate the efficacy of the polymerases for the amplification of samples of our yield.

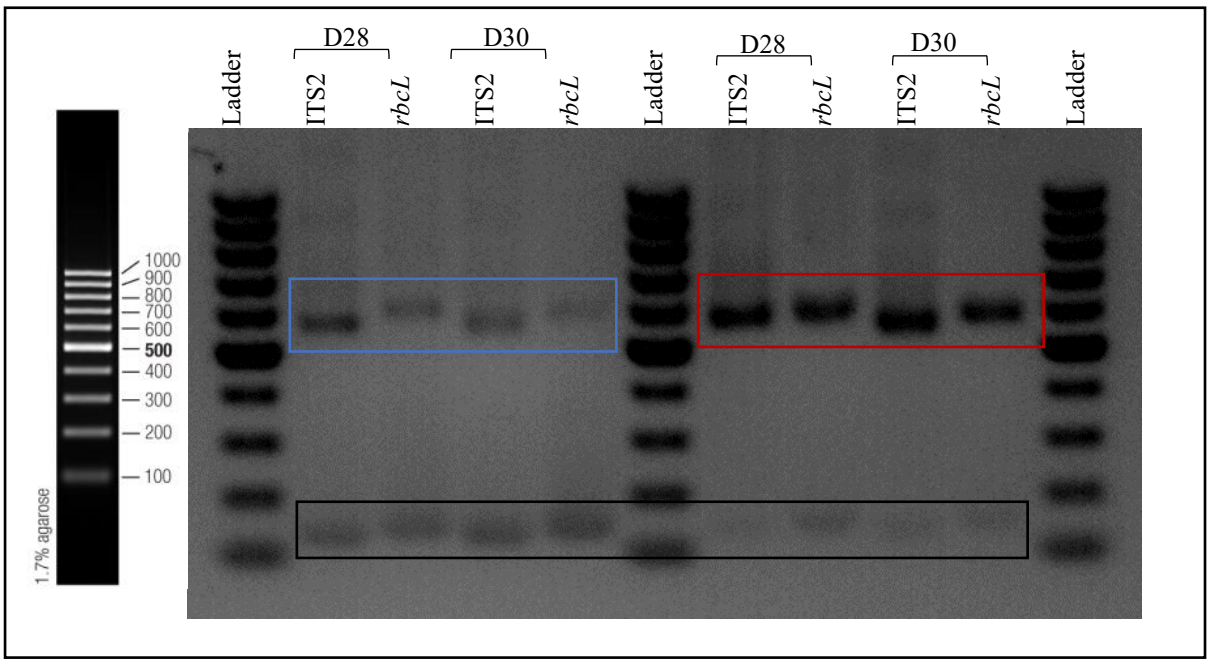


Figure 4.1. PCR reactions performed with Phusion Hot Start II High Fidelity Polymerase and Q5 Hot Start High Fidelity Polymerase on samples D28 and D30. Amplicons of the Phusion polymerase are grouped with a blue outline and Q5 polymerase in red. Primer dimers observed from these PCR were marked with the black outline.

Although both samples used for this amplification run had concentrations above the minimum required for amplification with Phusion polymerase greater than 10 ng/µL as shown in Table 3.2 in Chapter 3, the Phusion polymerase had difficulties in amplifying the barcodes as indicated by the low intensified amplicons visualised in Figure 4.1 (bands outlined in blue). Amplification of the samples with the Q5 polymerase that required at least 1 ng/µL concentration of the sample showed greater improvement producing amplicons of greater concentration as indicated by the intensity of their amplicons in Figure 4.1 (bands outlined in red) with minimal dimers in the reaction products.

Following the optimisation of the polymerases, PCR with Q5 polymerase showed to be highly efficient and was used for the subsequent reaction for the amplification of *rbcL* and ITS2 from the sample extracts.

4.5.2. Optimisation of annealing temperatures of ITS2 primers

The first few runs of PCR for the amplification of ITS2 barcode yielded PCR products that were outside the expected length range (i.e., 163 – 311 bp), too high between 500 – 600 bp.

Because the reactions for the amplification of both barcodes were run with a Gradient PCR (Figure 4.2), it was suspected that the conditions of *rbcL* amplification were influencing the ITS2 barcode amplifications. Therefore, the annealing conditions of the ITS2 were evaluated to determine the best annealing temperatures of ITS2 primers. A PCR was run with the same sample and reaction components' concentration under varying temperature conditions with annealing temperatures.

The purified samples were quantitatively assessed to determine the concentration of amplified DNA (Table 4.3). This measurement facilitated the identification of reaction conditions that maximised the yield of the desired amplicons with minimal by-product formation. The products of this run were visualised through an agarose gel electrophoresis in Figure 4.2.

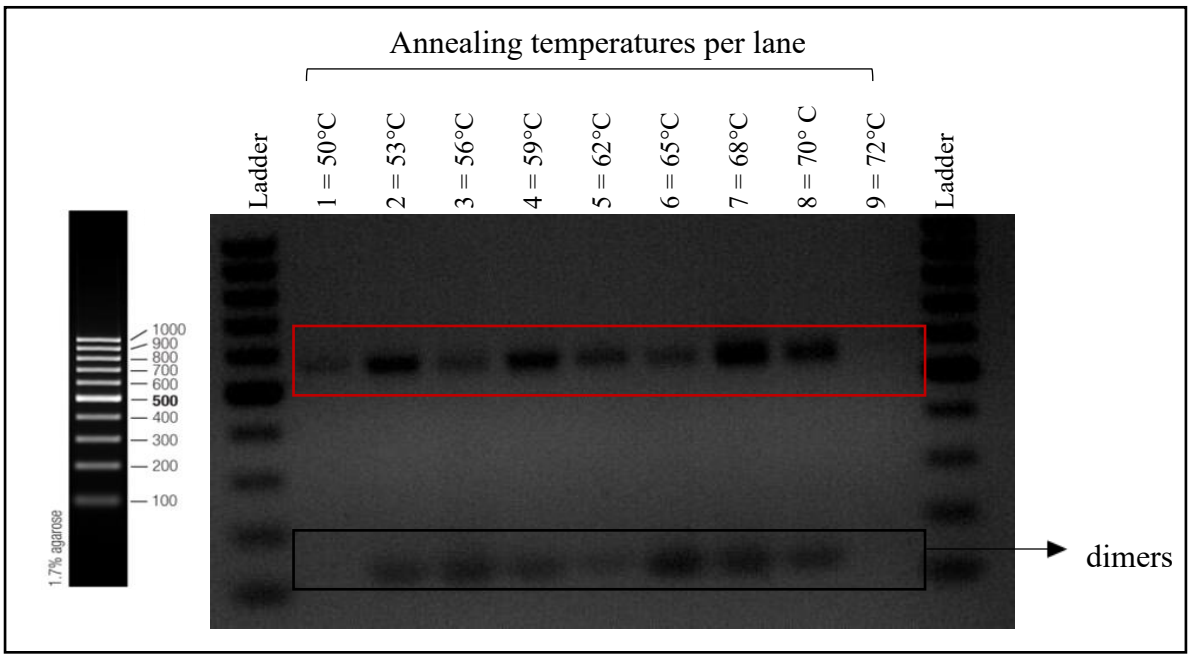


Figure 4.2. Agarose gel electrophoresis confirmation of ITS2 barcode amplification from a sample from Durban monitoring site (D28). This gel image shows the amplified ITS2 barcode sequences with PCR at different annealing temperatures. The lanes were loaded according to the PCR products from 1-9 according to the order of increasing annealing temperatures.

Therefore, preferred conditions of annealing had an intensified visually darker band which from assessing Figure 4.2 are the amplicons of temperatures 53°C, 59°C and 68°C in the order of increasing intensity.

The evaluation of quantity of amplicons' dsDNA assessed with Qubit HS Assay shown in Table 4.3 showed that the best annealing temperature was 59°C, and this reaction contained less dimers in comparison to the amplicons at 53°C and 68°C.

Table 4.3. PCR products of the varying annealing temperature evaluated for the amplification of sample D28 with ITS2.

Agarose gel lane	Primer annealing temperature (Ta = °C)	dsDNA concentration (ng/µL)
DNA ladder	-	-
1	50	10,34
2	53	51,28
3	56	12,16
4	59	73,55
5	62	20,21
6	65	18,47
7	68	22,10
8	70	19,38
9	72	<0.005
DNA ladder	-	-

Following this PCR run, looking at the size of amplicons in Figure 4.3 and their corresponding concentrations, it was established that the best annealing temperature for amplification of ITS2 region was 59°C.

4.5.3. In-silico simulation of barcode primer sequences

Based on the reactions that were performed (Figure 4.1 and Figure 4.2), the amplicons had sizes that varied from the sizes reported in literature in Table 4.2. The efficacy of the *rbcL* and ITS2 primer sequences for amplification of targeted regions was assessed by performing an in-silico simulation to determine the expected product lengths when amplifying species of Poaceae. Because there was a high deviation in sizes of ITS2 amplicons (i.e., between 500 - 600 bp) outside the expected range, another set of primers with a different reverse primer sequence was considered. The computational analysis used the complete chloroplast genome sequence of Bermuda grass, *Cynodon dactylon* (1431 bp), available under Sequence ID NC_034680.1, to predict the binding and amplification potential of the *rbcL* primers and the ITS2 primers were tested using the sequence of Timothy grass, *Phleum pratense* (556 bp).

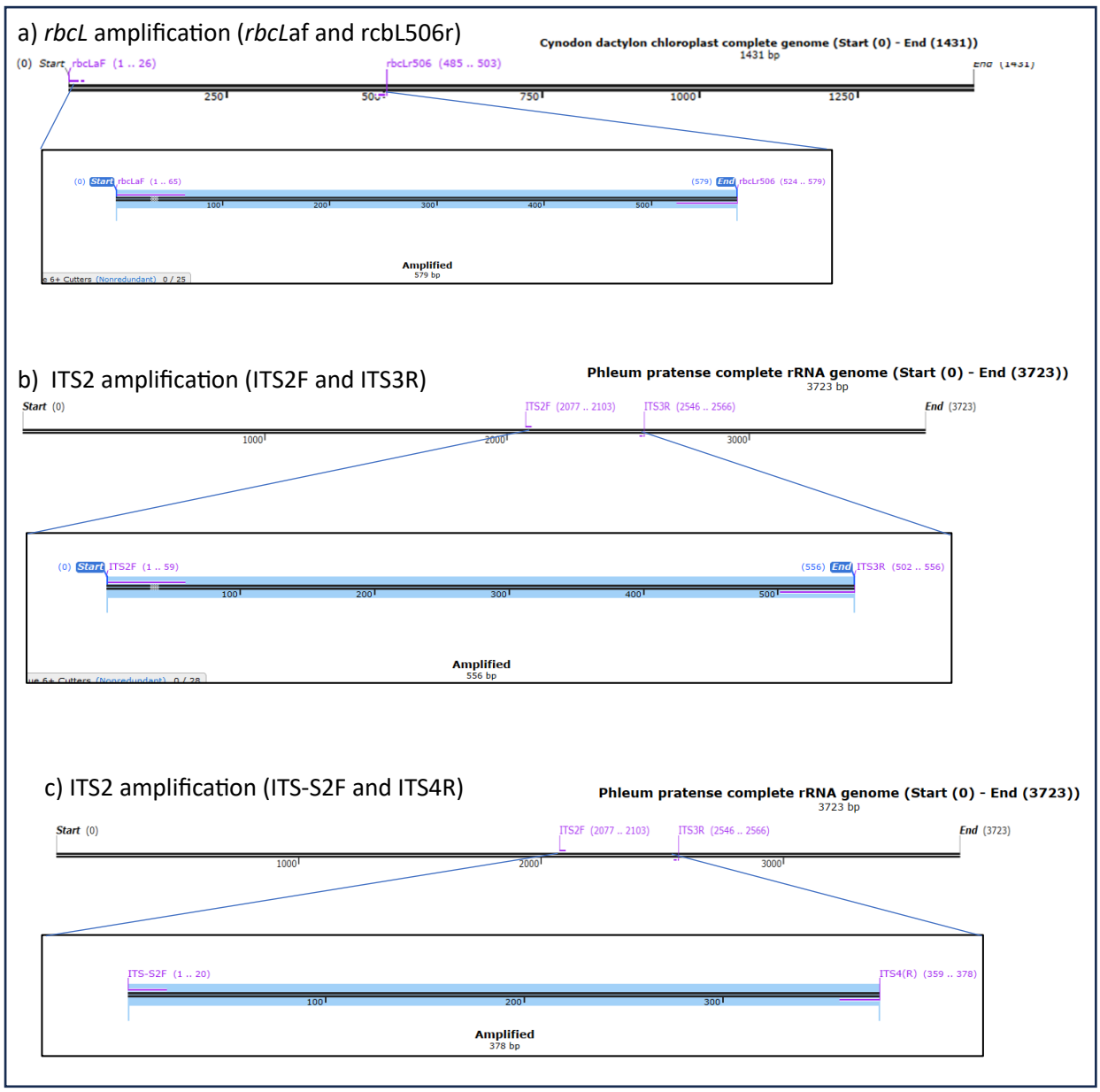
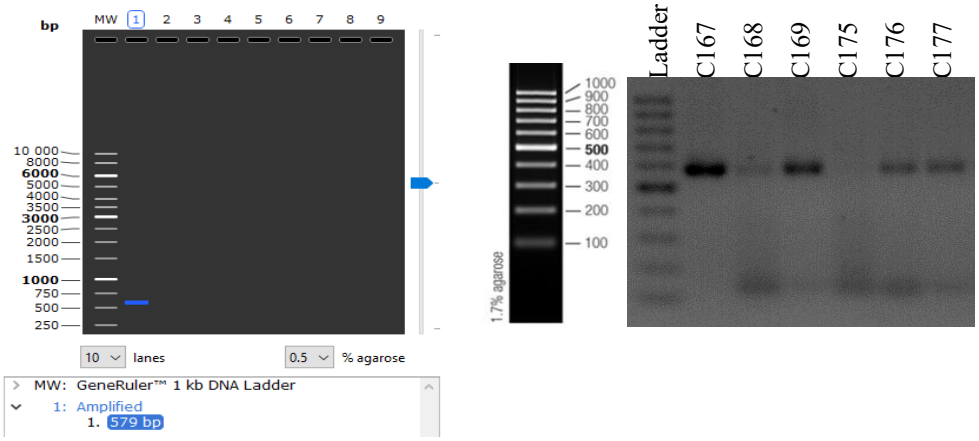


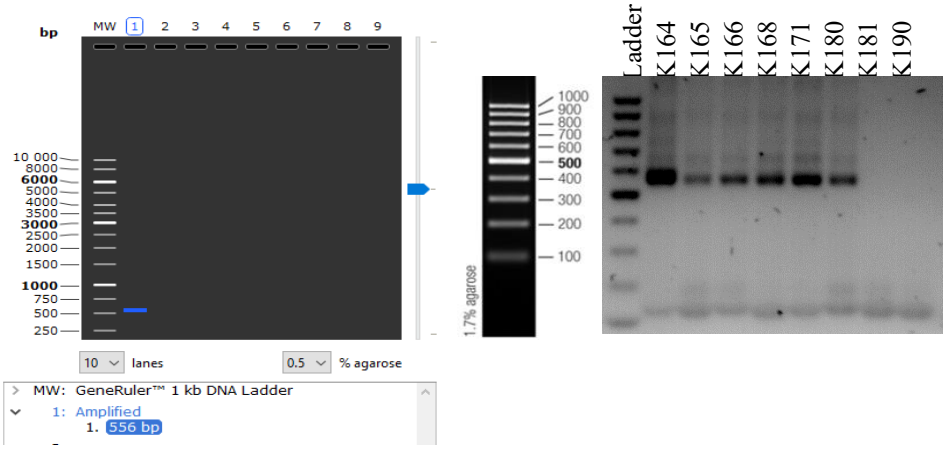
Figure 4.3. Computational simulation showing primer binding sites and amplification outcomes. The figure illustrates the specific binding positions of primers, assessed through simulation, to evaluate the amplification specificity and efficiency of primer sequences. The simulation predicts the expected fragment sizes of the PCR products for the amplification with each primer pairs. Created with SnapGene software (www.snapgene.com).

The computational analysis revealed the positions from which the *rbcL* and ITS2 primers bind on the provide species sequences in Figure 4.3, and generating a virtual agarose gel electrophoresis to show the size of the expected amplicon in Figure 4.4, predicted the primers would yield the anticipated product length in actual PCR conditions.

a) *rbcL* amplified PCR product (*rbcL*af and rcbL506r)



b) ITS2 amplified PCR product (ITS2F and ITS3R)



c) ITS2 amplified PCR product (ITS-S2F and ITS3R)

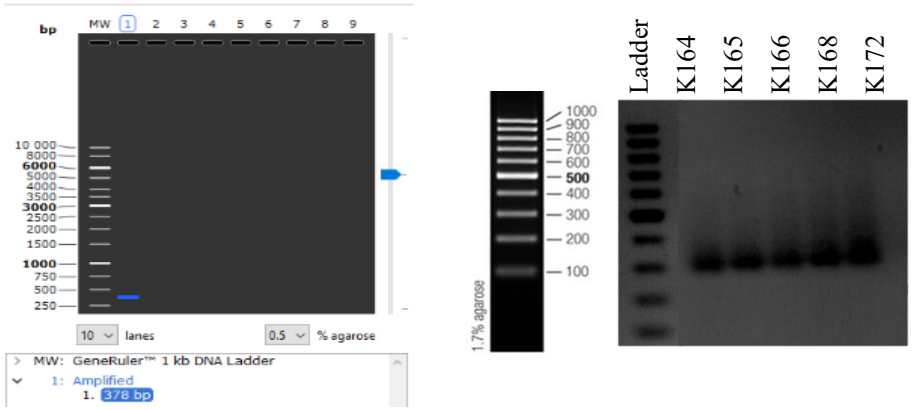


Figure 4.4. Comparison of fragment sizes resulting from in-silico (computational simulations) and in-vitro (laboratory experimental work) analyses using 0.5% and 1% agarose gel electrophoresis, respectively. The analyses employed primer pairs for amplifying barcodes, initially sourced from literature as suggested by Brennan and listed in Table 4.2 [8]. Created with SnapGene software (www.snapgene.com).

When comparing the findings of the computational analysis looking at the predicted product size length, with experimental amplicons it was apparent that the initial ITS2 primer pairs produced amplicons around 550 bp correlating to the products of our PCR. The second set of primers considered for ITS2 produced amplicons around 370 bp, and therefore since this was the preferred size length aimed for initially, subsequent ITS2 amplification used the ITS-S2F and ITS3R primer pair.

Because the ITS2 subunit sequence is a part of the nuclear ribosomal DNA that is located between the 5.8S and 28S rRNA genes with a length that often varies between 200 and 300 bp among different plant groups, the larger fragment amplified was suggestive of that the ITS2 primer pairs amplification was not specific to only amplify the ITS2 subunit. Yao *et al*. 2010 revealed that the average length of ITS2 in plants is 256 bp with the shortest going up to 167 bp and the longest up to 367 bp [101], which led to exploring other primer options with a close size range for the amplification of the targeted species in this study. Given the larger-than-expected fragment sizes and the practical consideration that sequencing plans for ITS2 barcode were designed for amplicons below 500 bp, alternative primer options were explored. This aimed at identifying primer pairs with a size range closely matching the targeted species in this study from those shown in Table 4.1. Furthermore, the decision to pursue smaller amplicon sizes was influenced by budget constraints, as sequencing costs are contingent on amplicon size, and larger fragments would incur higher expenses.

4.5.4. Summary of PCR amplification success for *rbcL* and ITS2

Following PCR, a total of 39 samples were successfully amplified for the *rbcL* region from the targeted monitoring sites. However, significant sample loss occurred during the optimisation experiments for ITS2 analysis resulting in exclusion of 12 samples which included 47% of samples from Cape Town, 29% from Kimberley and 40% from Durban. This exclusion is detailed in the flow diagram in Figure 4.5.

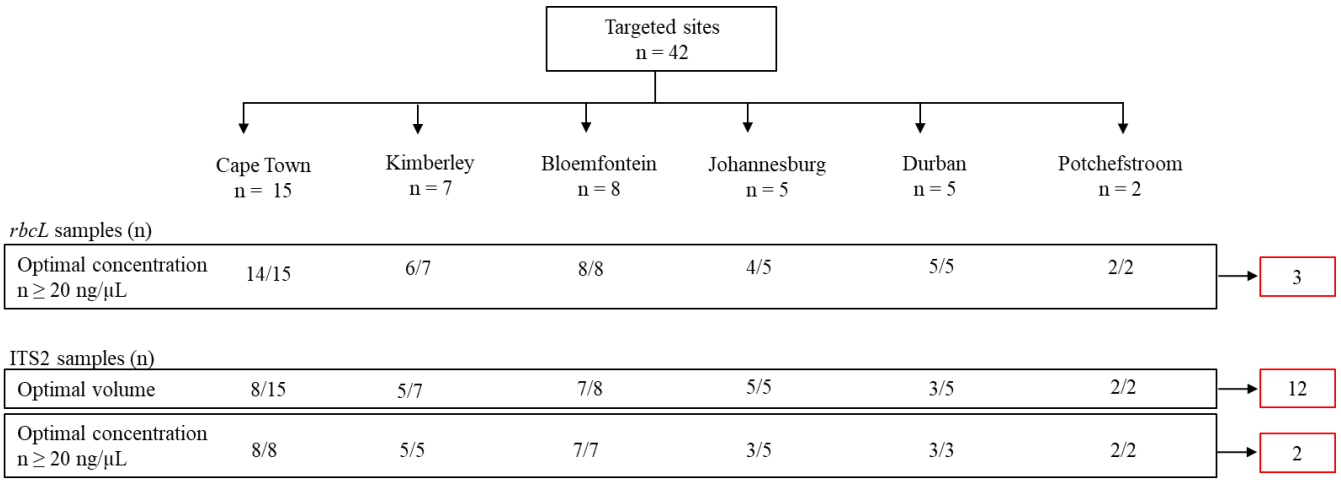


Figure 4.5. Sequential exclusion and amplification success of *rbcL* and ITS2 samples from each monitoring site.

Samples successfully amplified with at least 20 ng/µL dsDNA were a total of 39 for *rbcL* and 28 for ITS2. These were not necessarily complementary pairs of *rbcL* and ITS2 for each sample, which then influenced the final selection of samples for DNA library preparations.

4.5.5. Samples selected for subsequent processing and sequencing

The initial plan of selection of samples was to obtain a pair of both *rbcL* and ITS2 from each sample. However, from the total samples successfully amplified with optimal concentration, only 7 were complementary pairs of both barcodes from the same sample from Cape Town, Kimberley and Durban. To ensure that all the monitoring sites were included, additional samples with exceptional concentration were selected. The samples selected for each barcode are listed in Table 4.5 with the corresponding sampling date indicated in Figure 3.2. The concentration varied between 23-289 ng/µL and amplicon sizes between 169 - 686 bp.

Table 4.5. The final 25 samples selected for sequencing, consisted of the following number of samples of amplified *rbcL* and ITS2 barcodes with their corresponding amplicon size and concentration.

Sample	*rbcL*		ITS2	
	Average fragment size (bp)	dsDNA concentration (ng/µL)	Average fragment size (bp)	dsDNA concentration (ng/µL)
C169	630	121		
C176	634	90	455	265
C177	631	173	463	289
C184	634	112	456	214
D165	612	102		

D166	616	103	462	179
K164	639	87		
K165	655	68,6	468	173
K166	686	23	469	216
K172	676	32,4	461	191
B186			286	135
B189			526	278
B190			276	148
J179			249	97,9
J181			235	89,6
J190			503	205
P191			169	97,8
P192			197	86,2

The 25 samples selected included 10 samples for *rbcL* and 15 samples for ITS2, and only samples from Cape Town, Kimberley and Durban monitoring sites had both *rbcL* and ITS2. The number of distributions of selected samples for each monitoring site per barcode is shown in Table 4.6.

Table 4.6. The number of total samples per barcode for the subsequent analysis and preparation into DNA libraries for sequencing.

Site	***rbcL***	**ITS2**	**Total samples**
Cape Town	4	3	7
Kimberley	4	3	7
Johannesburg	-	3	3
Bloemfontein	-	3	3
Potchefstroom	-	2	2
Durban	2	1	3
Total samples	10	15	25

4.6. Discussion

The amplification of taxonomic barcodes is recognised as the most important stage in the effective implementation of metabarcoding. This is because the taxonomic barcode amplification generates copies of the DNA sequences containing genetic information from the DNA extracted which ensures for precise identification of the origin of pollen. Achieving the optimal method for retrieving these sequences from diverse pollen species' DNA within a single sample extract involves evaluating key steps that significantly impact and ensure the success of the amplification process. However, because of the complexity of environmental

DNA, it is impossible to definitively confirm that the amplified DNA corresponds to the targeted species.

The information that is normally relied on for guidance of the success is the size of the amplicons which is often determined through agarose gel electrophoresis which then makes accurate identification to rely heavily on the quality and quantity of taxonomic DNA that is successfully amplified. Several factors influence the amplification process, with notable contributors being the polymerase's amplification efficiency and the specificity of the PCR primers. As it was observed from our experiments the Q5 Hot Start High Fidelity polymerase proved to be more effective in these PCR, amplifying DNA from sample extracts with as little as 1 ng/µL DNA concentration in the reaction. This further showed the importance of using a highly sensitive polymerase for the amplification of DNA sequences from samples with low dsDNA concentrations. Selecting a polymerase with higher sensitivity also helped to minimise non-specific amplification (i.e., formation of dimers), thus improving overall reaction specificity.

The specificity of the primers, in turn, depends on carefully curated conditions, and the base pair information guides the extent of their amplification across a range of sequences from various species in each reaction. Prior to carrying out the PCR in the laboratory, there is a need to carry a computational analysis of the primer sequences to evaluate the properties and specificity of the primer sequences. Because for studies such as pollen metabarcoding we often rely on the literature from was done before and the previous findings to navigate the way [53], [95]. One of the things that costed us our precious samples in this study was jumping straight into the application of the methodology using primer sequences reported before evaluating their properties [86]. Initially, the first few reactions followed the suggested annealing temperatures based on the primer sequences' used by Brennan *et al* (2019) [86]. The PCR results suggested that the products of ITS2 amplification the fell outside the size range reported.

As we looked into optimising the reaction conditions for the ITS2 primers, it became clear that the primer pairs did not show optimal annealing at the suggested temperature at 50°C that was observed in the experiments by Lowe *et al*. (2022) [90]. Instead, a higher yield of ITS2 amplicons with fewer dimers was obtained at 59°C. While this optimisation was aimed at improving the efficiency of ITS2 amplification, it also led to an improvement in sensitivity as it was clearly indicated by the more intense agarose gel band at 59°C with minimal dimers.

Because the size length of the ITS2 products of PCR was longer than the amplicon sizes reported length by Lowe *et al*. (2022) [90], it was clear that the ITS2 primers were not specific to the ITS2 region. This was further confirmed by the computational analysis of the primers to predict the size length of their products when amplifying species of Poaceae. Furthermore, the ITS2 barcode primers were changed into ITS-S2F/ITS3R to produce amplicons within the initial size range, as the original plan of sequencing mapped for the DNA library preparation was based on obtaining amplicons less than 400 bp from the first PCR.

Despite samples lost during the ITS2 size length optimisations, we still fell short at successfully producing enough samples of ITS2 within the desired size range. From the 15 ITS2 samples' amplicons with sizes that varied between 169 – 686 bp, only 40% had amplicons that were less than 400 bp. Although all the *rbcL* samples' amplicons were also less in size in comparison to the size range initially targeted, their reduced length was advantageous. They were within an acceptable size range to enable for species identification as the CBOL Plant Working Group recommends *rbcL* fragments between 400 – 600 bp [45].

At the conclusion of the PCR experiments, the importance of conducting in silico analyses to assess primer specificity before conducting PCR in the laboratory was clearly observed. Such analyses can help in preventing reagent wastage and improve the SOPs for future application in pollen metabarcoding experiments. Additionally, given the complexity and low concentration of DNA in environmental pollen samples, there is still a lot of work to be done to help better the method. However, of all that is to be considered, selecting a highly sensitive polymerase capable of efficiently amplifying low-template DNA and highly specific barcode primers are of utmost importance in the first PCR step of preparing DNA samples for pollen metabarcoding.

DNA Library sequencing and bioinformatic analysis

5.1. Introduction

This chapter focuses on the study's third and overall objective that is, identifying pollen species from DNA sequence data generated by the MiSeq Illumina system. It outlines the preparation of a pollen DNA library from *rbcL* and ITS2 products of the first PCR, sequencing with the MiSeq system, bioinformatic processing of the sequence data, and taxonomic assignment of the detected pollen species.

The *rbcL* and ITS2 selected pollen samples that are products of the first PCR initially went through comprehensive quality assessment to ensure they are well aligned with the quality requirements of the MiSeq (300 cycles) V2 sequencing instrument with specifications of amplicon size greater than 150 bp but less than 500 bp (i.e., 150 bp <amplicon> 500 bp). DNA library preparation was performed using the Illumina DNA library kit that uses enzymatic fragmentation to ensure that all DNA products generated for the library conform to the specified fragment-length criteria.

5.2. Enzymatic fragmentation and DNA library preparation

DNA library preparation was conducted with Illumina DNA Prep, and unique dual indexing was achieved with the IDT for the Illumina DNA UD Indexes Set A kit (Illumina Technologies). These regents fragment and tag the amplicon sequences with Illumina adapters and index sequences in a single reaction, ensuring uniformity and precision in the resulting library. This enzymatically controlled reaction occurs through a series of sequential steps, with the key steps being: 1) dsDNA denaturation, 2) enzymatic cleavage, 3) end repair, 4) adapter ligation, 5) size selection, and 6) amplification [102].

The amplicon’s dsDNA undergoes an initial denaturation process, resulting in the creation of single-stranded DNA molecules achieved by subjecting the DNA to elevated temperatures. Subsequently, specific nucleases, recognising predetermined sites on the DNA, facilitate cleavage of the single-stranded DNA. These enzymes precisely cut DNA at defined locations, producing fragments of varying lengths. Following enzymatic cleavage, the DNA fragments undergo end repair to mend any damaged or uneven DNA ends. After this repair, adapter sequences are ligated to the cleaved DNA fragments. The adapters function as primers for

subsequent polymerase chain reaction (PCR) amplification and serve as recognition molecules for the sequencer's flow cell, initiating the sequencing reaction. To ensure the library consists of fragments within a specific length range, the DNA fragments with ligated adapters undergo size selection. This step is crucial to meet the sequencing system's requirements and plays an important role to guarantee the uniformity and precision of the library. The size-selected fragments are then subjected to amplification to generate multiple copies, ensuring an ample supply of sequences for the subsequent sequencing process [102].

5.3. Sequencing by synthesis with the MiSeq V2 Illumina instrument

The generation of sequences using the MiSeq (300 cycles) V2 sequencing instrument follows a systematic workflow. The underlying principle employed by this system is the sequencing by synthesis approach which allows for the determination of the sequence of DNA fragments by synthesizing complementary strands in a stepwise fashion and detecting each incorporated nucleotide. The process of generating sequence reads occurs through a series of sequential steps that includes: 1) library denaturation and immobilisation, 2) primer hybridisation, 3) bridge amplification, 4) sequencing by synthesis, 5) detection and imaging, 6) deblocking and 300 cycle repetition and 7) sequence data generation [103], [104].

The first phase of the reaction is set by initially denaturing the DNA library into single-stranded DNA fragments for immobilisation into the surface of the flow cell. The flow cell contains short DNA sequences or primers that are complementary to the adapter sequences, get hybridised to the single-stranded DNA sequences loaded on the flow cell. The flow cell then undergoes a bridge amplification process where the single-stranded DNA fragments are converted into double-stranded clusters with each cluster representing a localised and amplified section of the barcode DNA fragment. The Illumina system then employs reversible terminator nucleotides which are fluorescently labelled and contain a removable blocking group. With each cycle, a single nucleotide is introduced to which its signal gets recorded when incorporated into the growing DNA strand. At the end of each cycle the, the flow cell is imaged to capture the fluorescence signal generated by the incorporated nucleotide. The identity of each base incorporated is determined by the signal intensity at that position. Following imaging, the removable blocking group is cleaved allowing for the next cycle to commence and the process is repeated for a total of 300 cycles on the MiSeq V2 progressively building DNA sequences. The fluorescent signals are then translated into base calls, generating a sequence of nucleotides for each cluster. The raw data generated from the sequencing

instrument is reported in binary base call format (BCL or .bcl file) which is then subjected to bioinformatics analysis [104].

5.4. Methodology (I) – DNA library preparation and sequencing

The preparation of the library was completed using the Illumina DNA Prep (Illumina Technologies) and indexed with the IDT for Illumina DNA UD Indexes Set A kit. The libraries were equimolarly pooled with concentration of 4 nM for each library sample. Thereafter, the fragment size distribution was evaluated using the D1000 High Sensitivity Screen Tape Assay (Agilent Technologies), and concentration of the final pool determined using 1x Qubit dsDNA High Sensitivity Assay (Thermo Fisher Scientific). The values determined were used to calculate the final molarity concentration of the pooled library. Results were verified by qPCR library quantification using the NEB Next Library Quant kit for Illumina (Illumina Technologies) according to the manufacturer's instructions.

The final pooled library was denatured using 0.2 N NaOH and further diluted to a final loading concentration of 10pM. The denatured final library pool was spiked with 10% PhiX, which is the recommendation made by Illumina. Sequencing was completed on the MiSeq sequencing system using a MiSeq (300 Cycle) V2 reagent sequencing kit (Illumina Technologies) and the sequence reads data generated stored in the base call per cycle binary file (BCL file).

5.5. Methodology (II) – Bioinformatics data analysis

5.5.1. Quality assessment of sequence reads

Raw sequence data was demultiplexed using bcl2fastq software separating ii into individual samples' sequence data according to their unique indices producing files in FASQ format. The FASTQ files contain the sequence information associated with the quality scores of sequences reads. The raw sequence data was assessed for quality using the FastQC v0.12.1, and the reports generated combined into one with the MultiQC v1.18 [105], [106]. The sequence data was pre-processed to remove adapter sequences, merge forward and reverse sequence reads per sample, filter out low-quality bases, and allow the sequence reads to be classified for taxonomic assignment.

5.5.2. Adapter trimming and read pairing

Each of the DNA libraries generated two sequence reads from the sequencing of DNA fragments from both ends. The raw data sequence file when looking at each sample was in the format sample_L001_R1_001.fastq and sample_L001_R2_001.fastq, for R1 was the file that contained forward reads and R2 reverse reads. Qiime2 version 2023.9 employing the

join_paired_ends.py script was used to join the sequence reads within each sample. After the sequence reads were joined, quality control was performed through trimming and filtering.

Trimmomatic version 0.39 was used to remove adapter sequences. Since taxonomic barcodes were integrated into adapter sequences at both ends, a script for paired-end reads was executed with the following command line to trim the adapters:

```
java -jar trimmomatic-0.39.jar PE input_forward.fq.gz
input_reverse.fq.gz output_forward_paired.fq.gz
output_forward_unpaired.fq.gz output_reverse_paired.fq.gz
output_reverse_unpaired.fq.gz ILLUMINACLIP: TruSeq3-
PE.fa:2:30:10:2:True LEADING:3 TRAILING:3 MINLEN:36
```

Where input_forward.fg.gz and input_reverse.fq.gz were the forward and reverse sequence files for a specific sample.

The key steps covered by the instructions were the removal of adapters using the preloaded adapter sequences within the package, removal of low-quality bases that are below threshold quality from the start and end of the reads (leading and trailing), and the removal of the reads that fall below 36 bp in length [107].

5.5.3. Sequence filtering

The reads were filtered to eliminate low quality bases contained within the sequences looking at the Q20 quality score, that is, a probability of 1% incorrect bases call. Q20 filtering was performed in Qiime2 using usearch8.0 software for executing the filtering command line:

```
usearch8.0 -fastq _filter input.fastq -fastq_ maxee 1.0 -fastq_minlen
100 -fastqout output.q20.fq
```

Where the input file is the joined reads for a specific sample.

The command line carries key instructions for performing quality filtering on the fastq files, ensuring that reads with more than 1 error are filtered, discarding sequences shorter than 100 bp in length [108]. Reads were counted directly from the fastq files to determine the sum, mean, median and standard deviation. The product sequences pre-processed were used to classify and assign pollen taxa.

5.5.4. Universal Taxonomic Assignment eXpert (UTAX) classifications

The UTAX algorithm assigned taxa based on sequence similarity, calculating a score that reflects K-mer distances to both the top hit and the nearest neighbour at each phylogenetic level using the 2014 UTAX trained Viridiplantae database and a raw score cut-off of 20 to determine the confidence of the taxonomic assignment.

The algorithm used the following usearch8.0 command line to execute the classifications:

```
usearch8.0 -utax filteredsamplefile. q20.fq -db
viridiplantae_all_2014.utax.udb \ utax_rawscore \ -tt
viridiplantae_all.utax.tax \ -utaxout filename.utax
```

Aggregate counts of all the reads per taxon were generated to determine the sum, mean, median and standard deviation of the assigned taxa. The downside of UTAX based classifications is that the algorithm provides raw scores rather than the bootstrap equivalent confident values [78].

5.5.5. Ribosomal Database Project (RDP) classifications

The RDP classifier is a naïve Bayesian classification algorithm that uses a comparison of K-mer frequencies to an RDP trained reference database. The RDP algorithm applied a bootstrap cut-off at 85% as classification threshold [82]. The classifications were executed using a perl (version 5.36.3) script with the following command line:

```
perl /code/classify_reads.pl --out results <path_to_reads>/*.fastq \
--utax-db utax_trained/viridiplantae_all_2014.utax.udb \ --utax-
taxtree utax_trained/viridiplantae_all_2014.utax.tax \ --rdp --rdp-
jar <path_to_RDPTools>/classifier.jar\ --rdp-train-propfile
rdp_trained/rRNAClassifier.properties
```

Where the 'path_to_reads' is the location of the folder containing the fastq sample files for classification. The RDP classifier classified sequences based on a 2019 trained Viridiplantae ITS reference database, providing bootstrap confidence values greater than 0.8 for positive taxonomic assignments and disqualified taxa assignment with a confidence score lower than 0.8 [82], [109].

5.5.6. Basic Local Alignment Search Tool (BLAST) classifications

Adopting the method of classification from the National Botanic Garden of Wales (NBGW) plant Illumina Plant Pipeline, the sequences were classified using blast (Nucleotide-Nucleotide BLAST version 2.12.0+) [110]. The classifications were executed using the following command line:

```
blastn -query '/clustered.fasta' -db 'reference_database.fasta' -out
'blast-clustered.csv' – outfmt '10 std score qcovs stitle' -
max_target_seqs 20 -num_threads 16 >& /blast-log-clustered-rbcL-
multis.txt
```

To prevent spurious BLAST hits, *rbcL* and ITS2 blast formatted reference databases were used [80]. This classification looks at the top 20 hits for each sequence query and runs the search

with 16 threads for parallel processing providing the Expectation value (E-value), percentage identity and bit score associated with each classification.

Although there are no standard thresholds for BLAST classification values, the common specifications associated with positive classifications is the combination of E-value less than 0.00001, percentage identity greater than 90% and query coverage greater than 80% [110].

5.6. Verification of identified taxa

To verify the accuracy of the results, a comparison was drawn by comparing the species identified with UTAX, RDP, and BLAST, looking at the positive identifications as specified by the threshold confidence of identifications set for each algorithm. Local botanical data was retrieved from the South African National Biodiversity Institute (SANBI) to compare the species classified with the local Poaceae species in South Africa [111].

5.7. DNA libraries assessment results

This section details the quality evaluations conducted on the samples post PCR 1 and purification, preceding the DNA library preparation and pooling steps in PCR 2.

The success of library preparation and indexing reactions was confirmed through the assessment of individual library sizes using the Agilent D1000 HS Tape Station Assay. The average fragment size of *rbcL* amplicon samples ranged from 452 bp to 498 bp, while that of ITS2 amplicon samples varied from 212 bp to 421 bp. The sizes of *rbcL* and ITS2 libraries visualised are shown in Figure 5.1 and Figure 5.2 respectively.

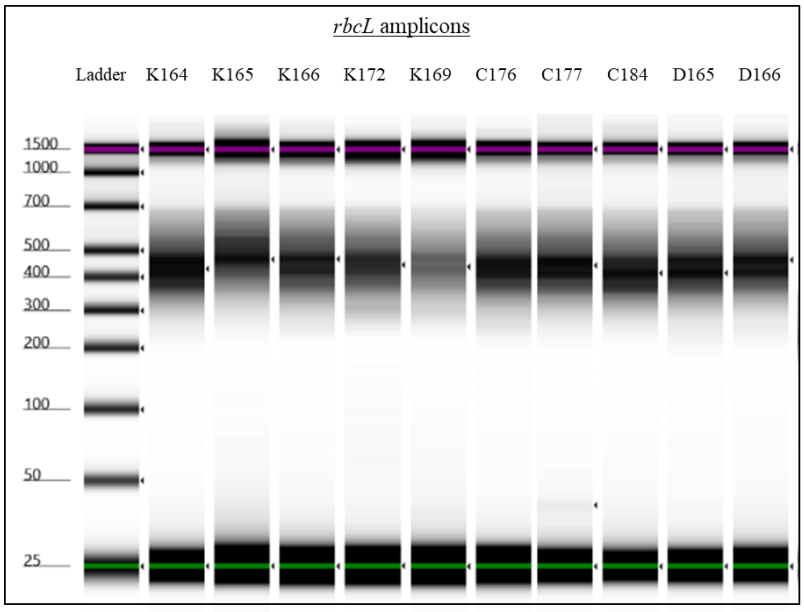


Figure 5.1. DNA libraries of *rbcL* amplicons sizes visualised on D1000 Screen Tape. The green line indicates the lowest molecular weight of the ladder while the pink is the maximum size.

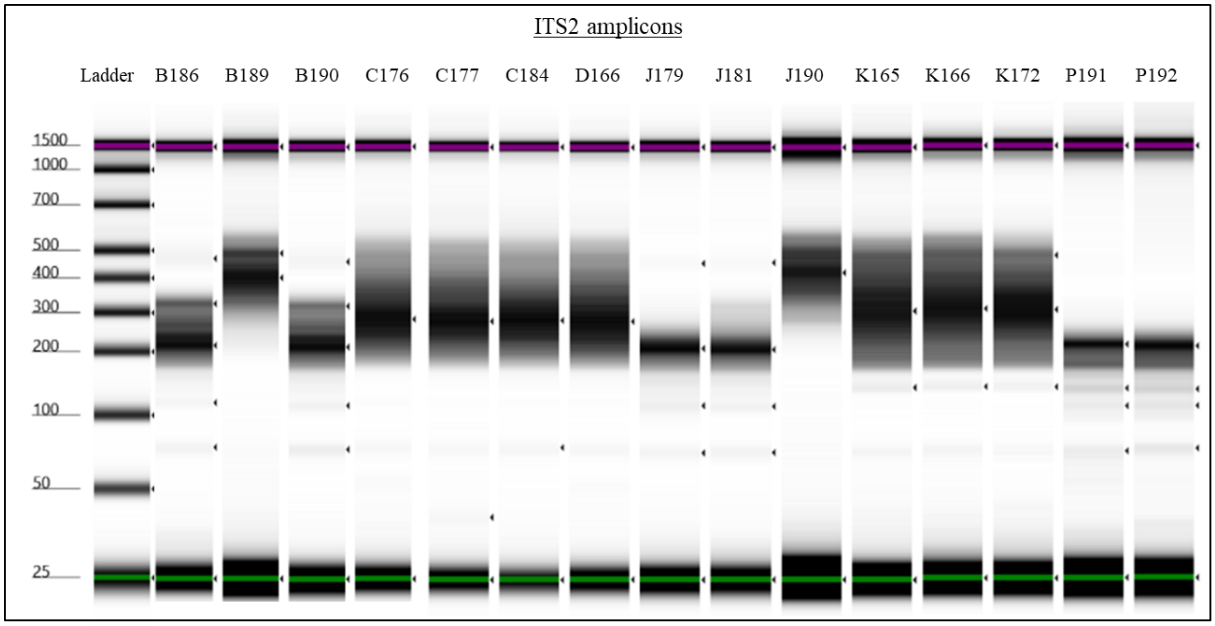


Figure 5.2. ITS2 DNA libraries prepared for sequencing visualised on D1000 Screen Tape.

The average fragment length calculated based on the calibrated concentrations of the DNA ladder (the lower marker base pairs (green line) and upper marker base pairs (pink line)) are listed in Table 5.1.

Table 5.1. DNA libraries of *rbcL* and ITS2 concentrations and average amplicon sizes.

Sample ID	Amplicon	Library concentration (ng/µl)	Average fragment size (bp)
K164	*rbcL*	9,88	460
K165	*rbcL*	7,77	498
K166	*rbcL*	6,28	467
K172	*rbcL*	5,82	457
C169	*rbcL*	3,57	456
C176	*rbcL*	7,7	452
C177	*rbcL*	7,75	466
C184	*rbcL*	10,4	453
D165	*rbcL*	11,3	463
D166	*rbcL*	7,94	475
B186	ITS2	16,8	257
B189	ITS2	5,56	421
B190	ITS2	11,9	249
C176	ITS2	26,2	334

C177	ITS2	29	334
C184	ITS2	30,6	331
D166	ITS2	31	330
J179	ITS2	7,49	216
J181	ITS2	9,93	227
J190	ITS2	4	437
K165	ITS2	12,6	346
K166	ITS2	12,5	351
K172	ITS2	14,3	343
P191	ITS2	3,09	213
P192	ITS2	6,02	212

The DNA libraries from 25 samples revealed the highest concentration of *rbcL* amplicons (11.3 ng/µl) in sample D165 from Durban, while the highest concentration of ITS2 amplicons (30.6 ng/µl) was found in sample C184 from Cape Town. The *rbcL* amplicons were shorter than the expected size range of 702 - 883 bp, whereas the ITS2 amplicons fell within the anticipated size of 311 bp ± 109 bp.

5.8. DNA sequencing results

The bioinformatic analysis of the sequence data was conducted using methods adapted from two pipelines. UTAX and RDP classifications followed the pipeline by Sickel *et al*. 2015, while BLAST classifications were based on the NBGW pipeline [78], [112].

5.8.1. Sequence read quality assessments

Quality control of raw sequence reads was based on quality reports obtained from the assessments using FastQC and the summary of the results of all samples generated with MultiQC. The summary highlights the key important findings such as the reported total sequences in each sample, their average length and percentage of duplicated reads as shown in Figure 5.3.

Upon quality assessments of the read lengths, it became evident that a miscommunication with the sequencing provider had occurred. Although only amplicons exceeding 600 bp were supposed to be fragmented, all samples were subjected to fragmentation, despite being below the required read length. Consequently, the resulting reads were shorter, with the longest average read length being only 147 bp. In 12/25 sample read pairs, the average read length was 147 bp, while 7/25 had an average read length of 118 bp. Six samples (B189, B190, J179, J181,

P191 and P192) had average read lengths below 100 bp as shown in Figure 5.3. This affected subsequent analyses, as reads shorter than 100 bp did not meet the filtering criteria, leaving very few usable sequences from samples with a low average read length.

The quality check detailed in Figure 5.4 shows that the quality per base was consistently good in most sequences, even with ambiguous bases (Per Base N content). However, the GC Content per sequence in several samples suggested issues with base composition and GC balance. Additionally, sequence duplication and overrepresentation were the most concerning, suggesting anomalies in specific samples. The overall metrics met quality thresholds while several samples such as J179, P191 and P192 showed deviations, necessitating sequence filtering.

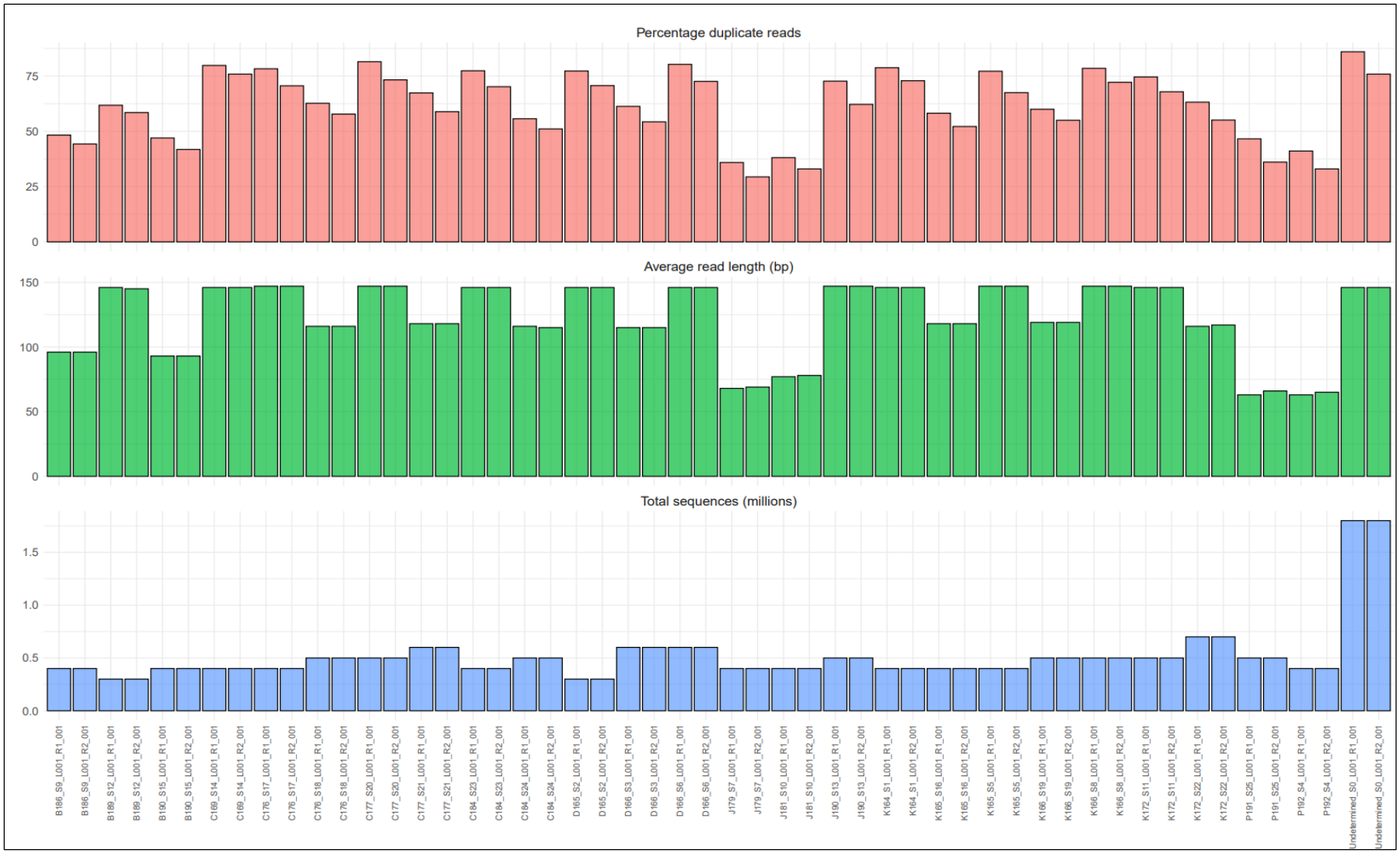


Figure 5.3. Quality assessment report showing the total sequences (millions), the average read length and duplication percentages of the reads for complimentary reads (R1 and R2) of the samples.

Figure 5.4. Status for each FastQC section showing whether results seem entirely normal (green), slightly abnormal (orange) or very unusual (red) in each of the samples.

5.8.2. Sequence reads filtering

Looking at the Phred scores of the sample sequences and their average read positions, a noticeable decline in sequencing quality was observed, particularly beyond 100 bp, as shown in Figure 5.5. Initially, the sequences maintained high-quality scores (above 30, green zone), but as sequencing progressed, a number of reads fell below the quality threshold of 20. This decline indicated an increase in sequencing errors over time. Therefore, applying quality control measures, such as filtering, was crucial to ensure reliable downstream analysis by retaining only high-quality reads. Post quality control, the sequencing reads demonstrated significant improvement, as low-quality bases were removed, minimising erroneous base calls and enhancing the overall sequence quality for most samples.

Figure 5.5. Comparison of the average quality of sequence reads per sample (represented by lines) before and after quality control implementation. The graph shows the decline in sequencing quality across sequence length and the improvement in quality after filtering low-quality bases.

Upon assessing the filtering efficiency as visualised in the PCA plot shown in Figure 5.6, which highlighted the distribution of samples based on read quality before and after, it was clear that out of the 25 samples, 15 showed a good retention rate, maintaining a substantial proportion of high-quality reads. However, 8 samples experienced significant read loss, with most of their sequences being filtered out. This reduction in retained reads was consistent with the quality report, where these samples showed numerous sequences flagged with warnings post-sequencing. Out of the six sites that were investigated, only Potchefstroom samples did not contain sequences that could be used in classification for species assignment. The ITS2 two samples from Potchefstroom were eliminated from further analysis post quality assessment and filtering. Following this, the sequence reads size variation between *rbcL* and ITS2 was assessed to understand the distribution of sequence reads generated [113].

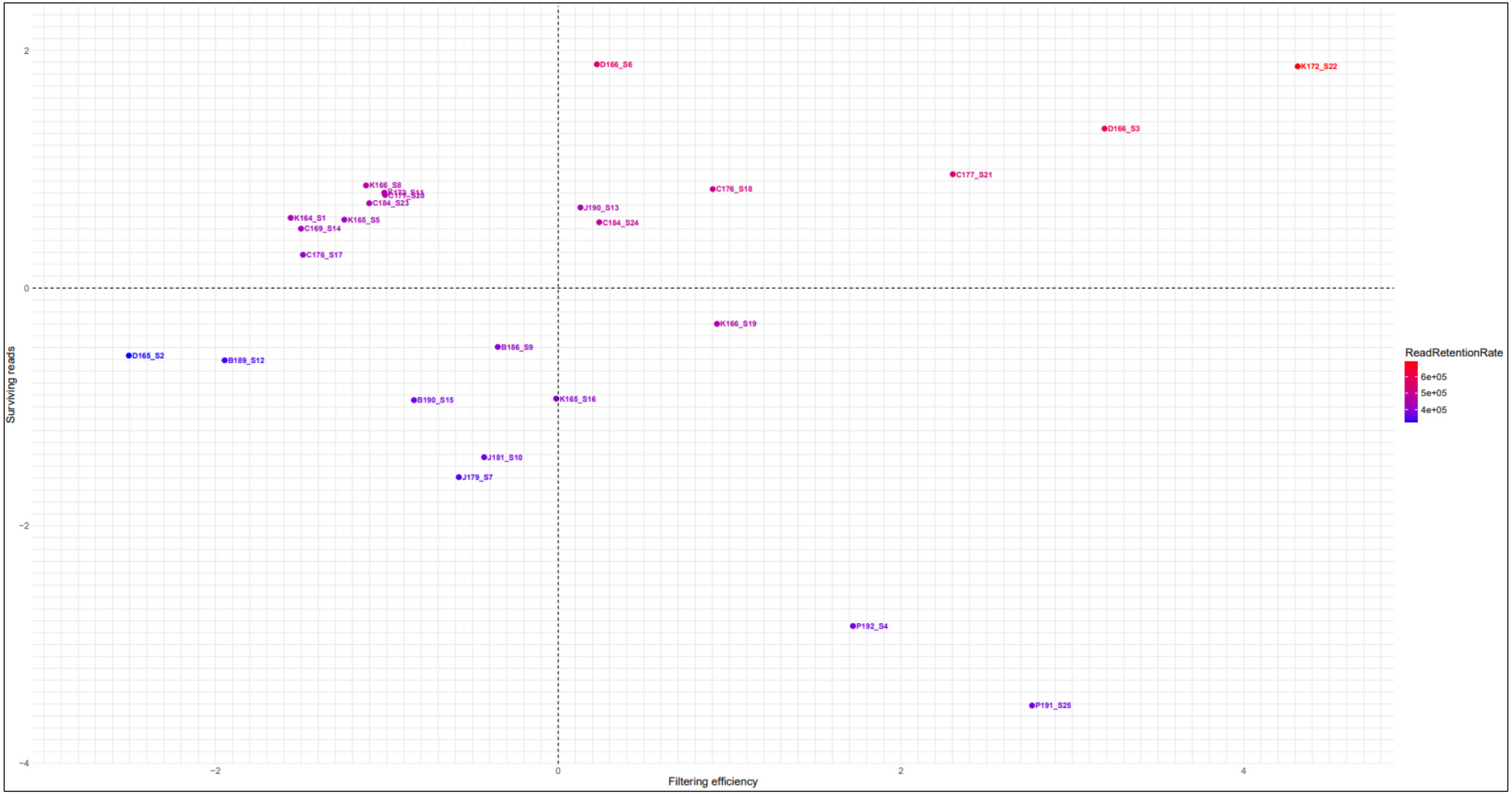


Figure 5.6. Principal Component Analysis (PCA) plot for the read survival rate and filtering efficiency across multiple samples. PC1 and PC2 capture the variation in read retention and filtering efficiency. Data points are color-coded according to the Read Retention Rate, representing the proportion of reads that passed the filtering threshold. The position of each sample on the plot reflects the balance between input reads, dropped reads, and surviving read. Created with RStudio [88].

5.8.3. Merge output - Amplicon size distribution

Amplicon size distribution curves were used to compare the distribution and abundance of filtered ITS2 and *rbcL* amplicons using the amplicon size distribution plot (Figure 5.7).

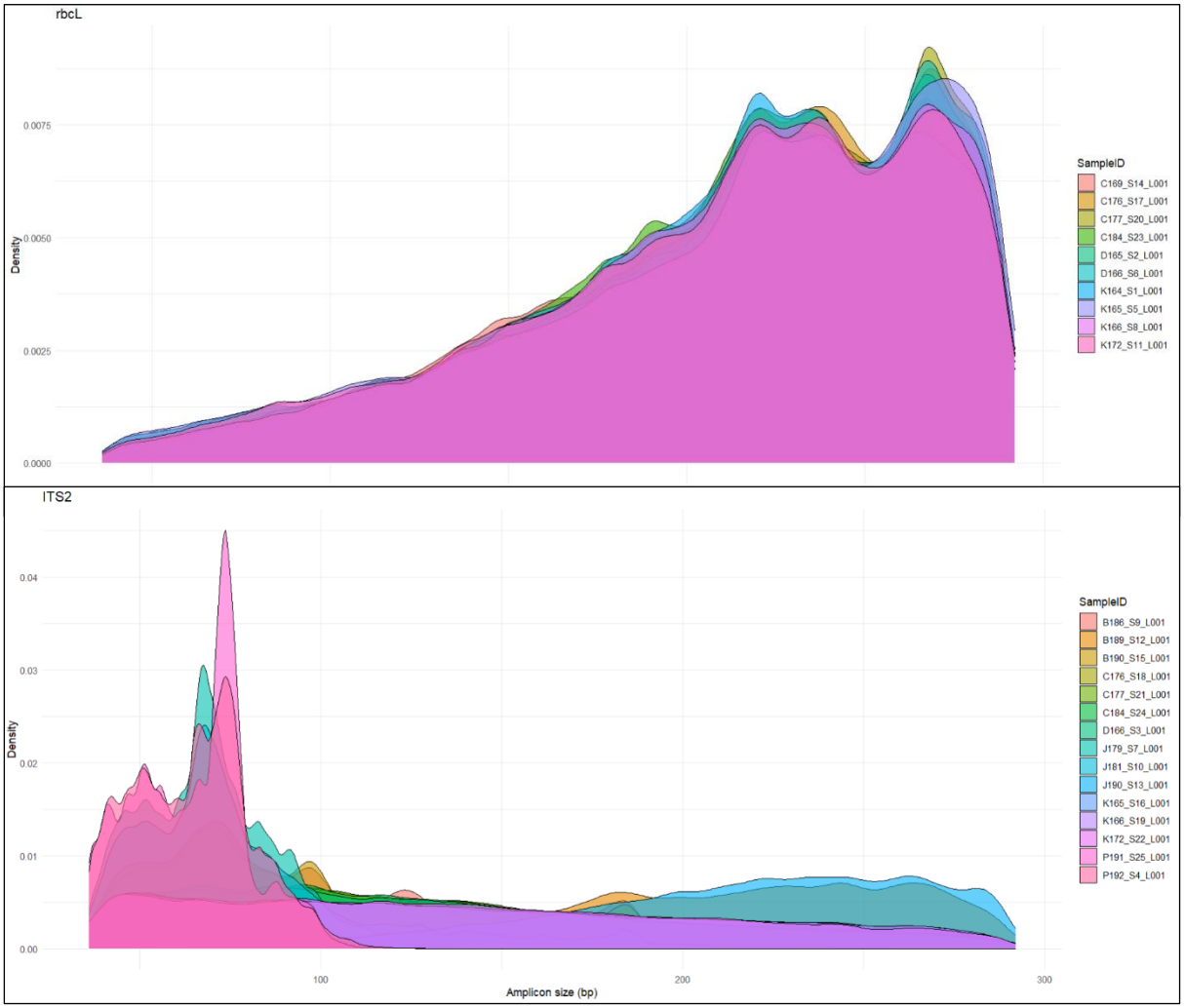


Figure 5.7. Amplicon size distribution curve showing the variability in sequences of *rbcL* and ITS2. The density indicates the relative frequency corresponding to the amplicon size. Created with RStudio [88].

The *rbcL* amplicons showed lower abundance and a broader size distribution across the entire sequences read length, indicating greater sequence diversity. ITS2 amplicons were significantly more abundant, with peaks up to four times higher than the highest *rbcL*. The ITS2 amplicons showed a concentrated distribution of sharp peaks suggesting more uniform amplification and reflecting the conserved nature of the ITS2 region in certain taxa. The concentration of amplicon sizes around specific regions supports the hypothesis that the ITS2 barcode exhibits distinct patterns of conservation and variation among taxa [101], [114].

5.8.4. UTAX Classifications – Poaceae species

Rarefaction curves were used to assess the biodiversity in each sample from taxa obtained with UTAX classifications of ITS2 and *rbcL* sequences. The curves represent the sequencing depth for each sample sequences excluding taxa assigned with 0.1% reads.

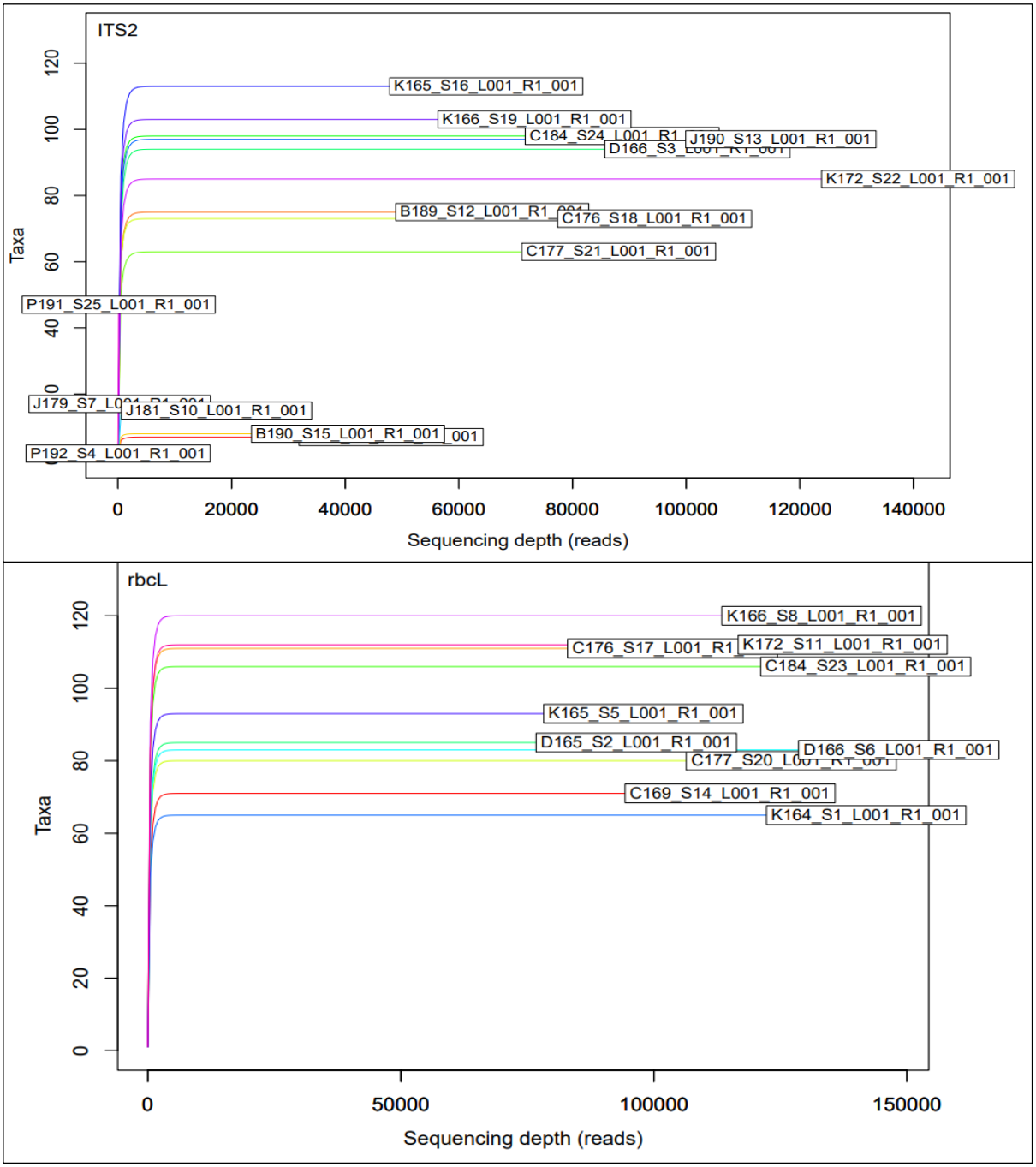


Figure 5.8. Rarefaction curves showing taxa accumulation in each sample from the ITS2 and *rbcL* sequence classifications with UTAX. The plots were generated on RStudio using vegan library package 2.6-6.1.

The operational taxonomic units (OTUs) from UTAX classifications were filtered for taxonomy of Poaceae, using a minimum cutoff of 50 sequences for taxa assignment. Six genera of interest were identified from ITS2 classifications: *Cynodon*, *Eragrostis*, *Lolium*, *Paspalum*, *Poa*, and *Stenotaphrum*. A total of 90 unique Poaceae species assigned with ITS2, 143 with *rbcL* sequences.

Figure 5.9. Stacked bar plot showing the relative abundance of Poaceae genera identified in each sample identified in the UTAX classification of ITS2 (above) and *rbcL* (below) sequences. Genera of interest are marked by the surrounding square box. Created with RStudio [88]. Created with R studio [88].

5.8.5. RDP classifications – Poaceae species

The biodiversity of the taxa obtained from the classification of sequences with RDP was also assessed with the construction of rarefactions curves of the RDP classified OUTs of ITS2 and *rbcL*.

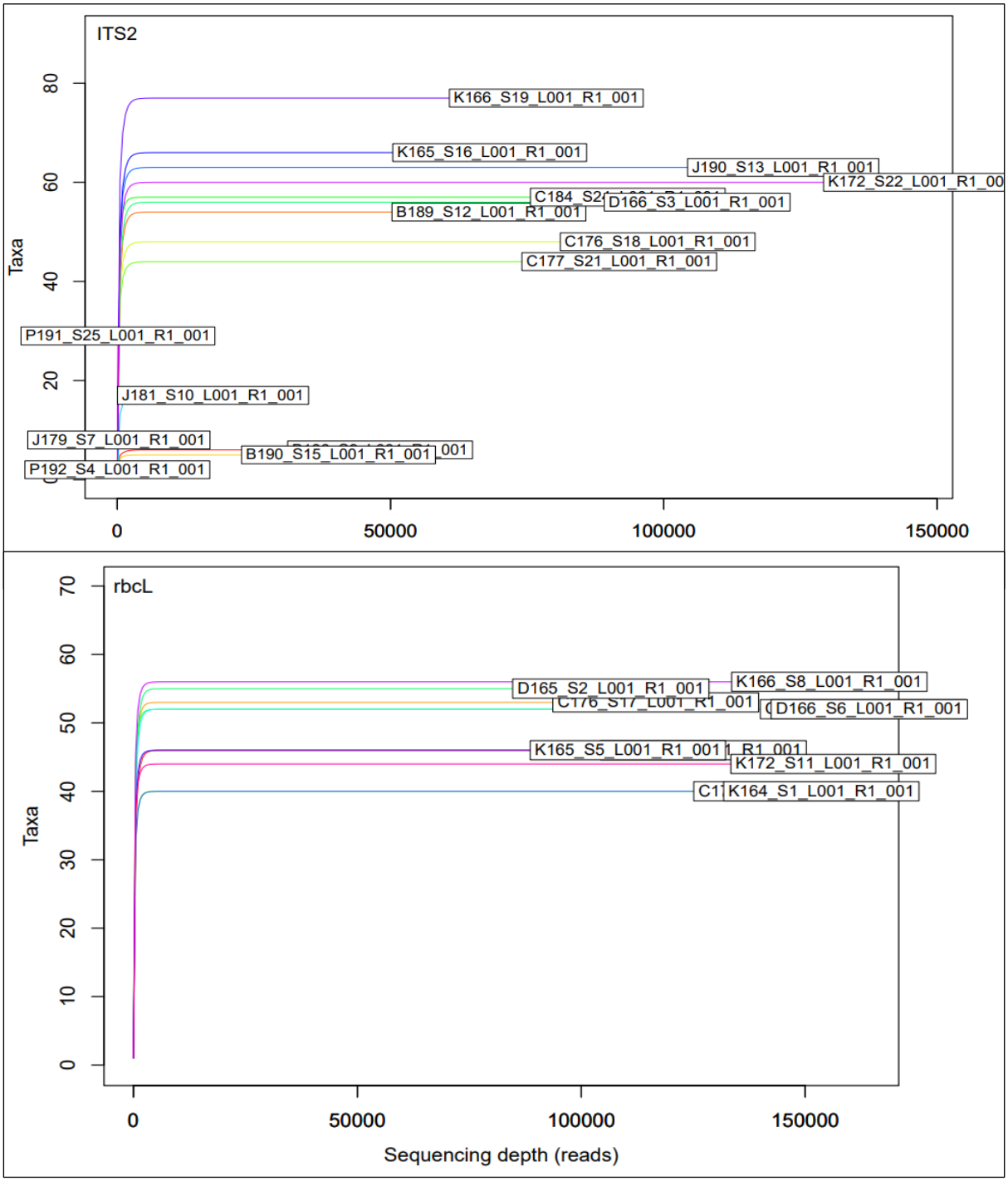


Figure 5.10. Taxa accumulated in each sample from the classification of ITS2 and *rbcL* sequences with RDP. Created with RStudio [88].

The RDP classifications of ITS2 and *rbcL* sequences showed the lowest number of unique taxa identified. For ITS2 sequences, the taxa accumulation plateaued at fewer than 80 unique taxa,

while for *rbcL* sequences, it plateaued below 60 as shown in Figure 5.10. When comparing to UTAX, the number of taxa identified with RDP was significantly lower. In ITS2 sequence classifications, *Cynodon*, *Eragrostis*, *Lolium*, *Paspalum* and *Poa* were identified and the same observed in *rbcL* sequences except for *Eragrostis* as illustrated in Figure 5.11 for genera identified in each sample. For ITS2, 37 unique Poaceae species were identified and only 11 from *rbcL* sequences.

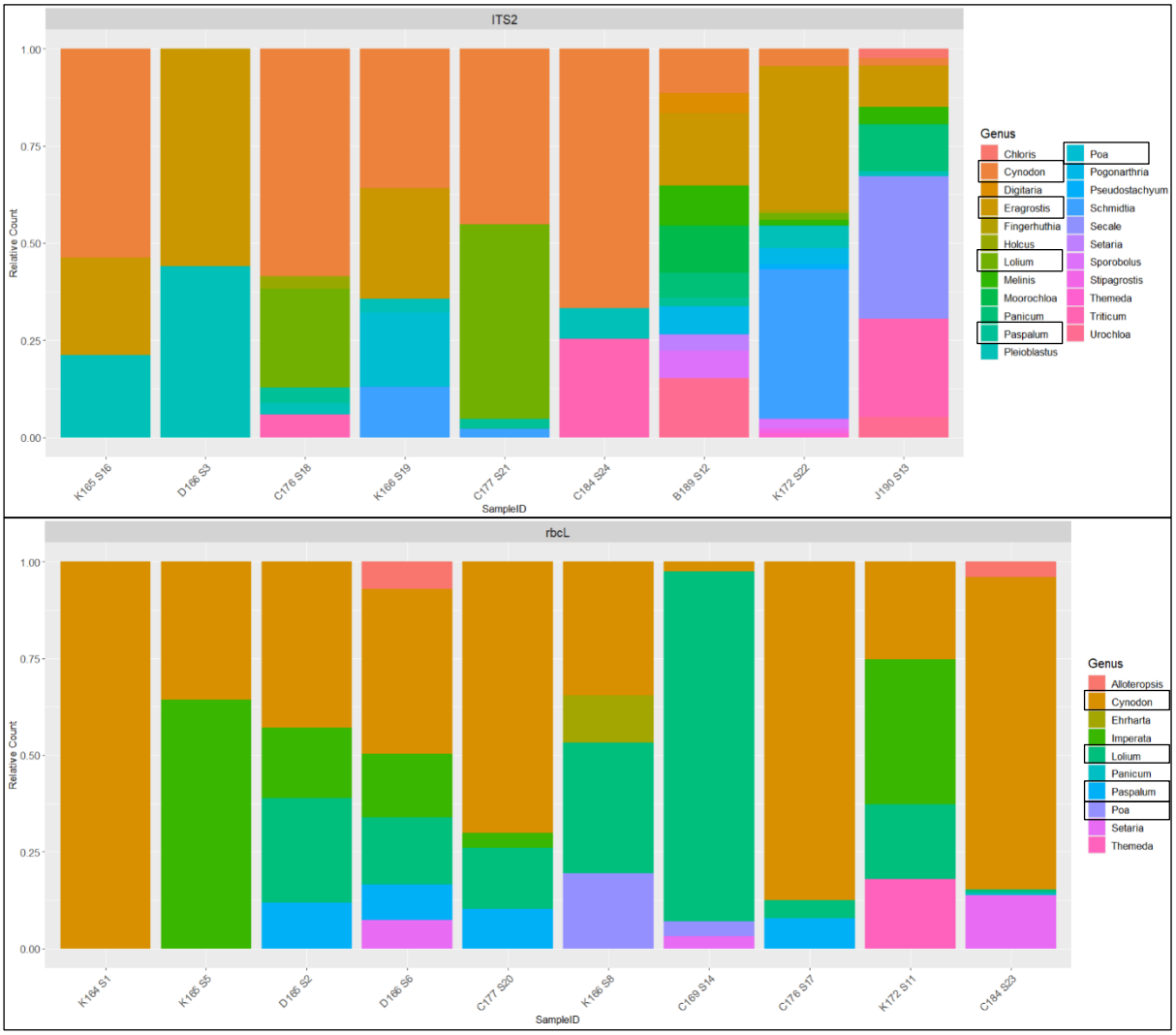


Figure 5.11. Stacked bar plot showing the relative abundance of Poaceae genera identified in each sample for both ITS2 and *rbcL* sequences classified using the RDP. Genera of interest are marked by the surrounding square box Created with R studio [88].

5.8.6. BLAST classifications – Poaceae species

The classification of sequences using BLAST followed the method described by Jones *et al.* (2021), although there were no specific guidelines for the expected output of the classification structure [112] . From the BLAST results, we extracted the aggregate counts of sequences used for taxa assignment, excluding taxa with fewer than 50 counts, and filtered for Poaceae. The genera identified in each sample are shown in Figure 5.12. Among the assigned taxa, *Cynodon*, *Eragrostis, Lolium* and *Paspalum* were identified from the ITS2 sequences. Additional to these *Hyparrhenia*, *Pennisetum* and *Stenotaphrum* were identified with *rbcL* sequences. A total of 134 unique species were identified from ITS2 sequences, and 173 with *rbcL* sequences.

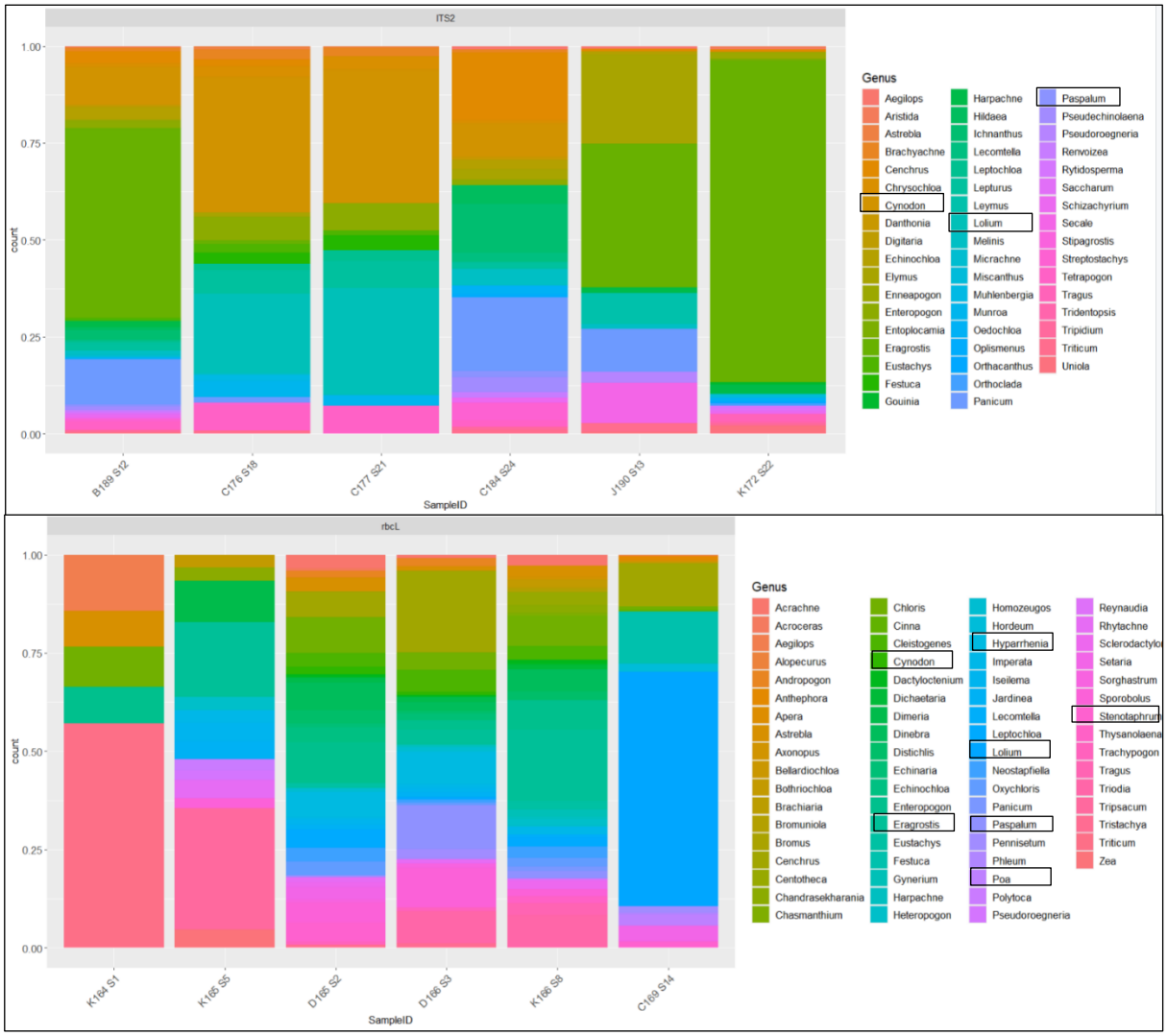


Figure 5.12. Stacked bar plot showing the relative abundance of Poaceae genera identified in each sample for ITS2 (top) and *rbcL* (bottom) marker, classified using the BLAST. Genera of interest are marked by the surrounding square box. Created with RStudio [88].

5.8.7. *Ambrosia* species classifications

When classifying the sequences of *rbcL* and ITS2 for the identification of *Ambrosia* species, it was only the *rbcL* sequences that returned positive identification across the three classification tools. Several species were identified with *Ambrosia artemisiifolia* identified with UTAX and BLAST classifications, and *Ambrosia trifida* consistently identified across the three classification tools as shown in Figure 5.13.

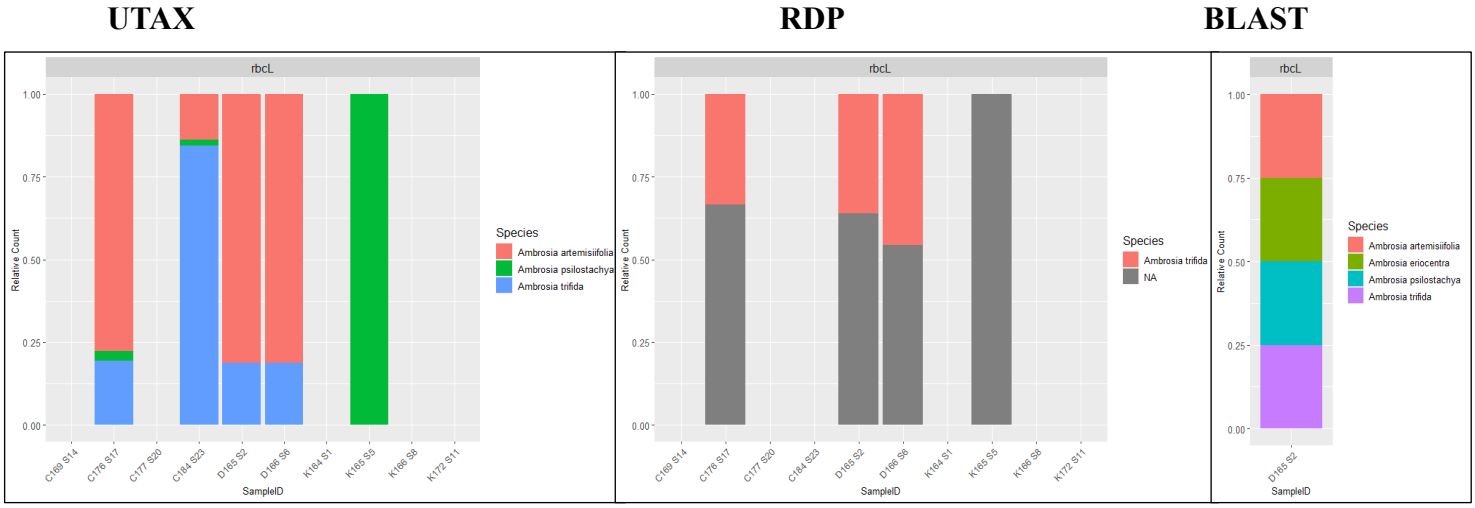


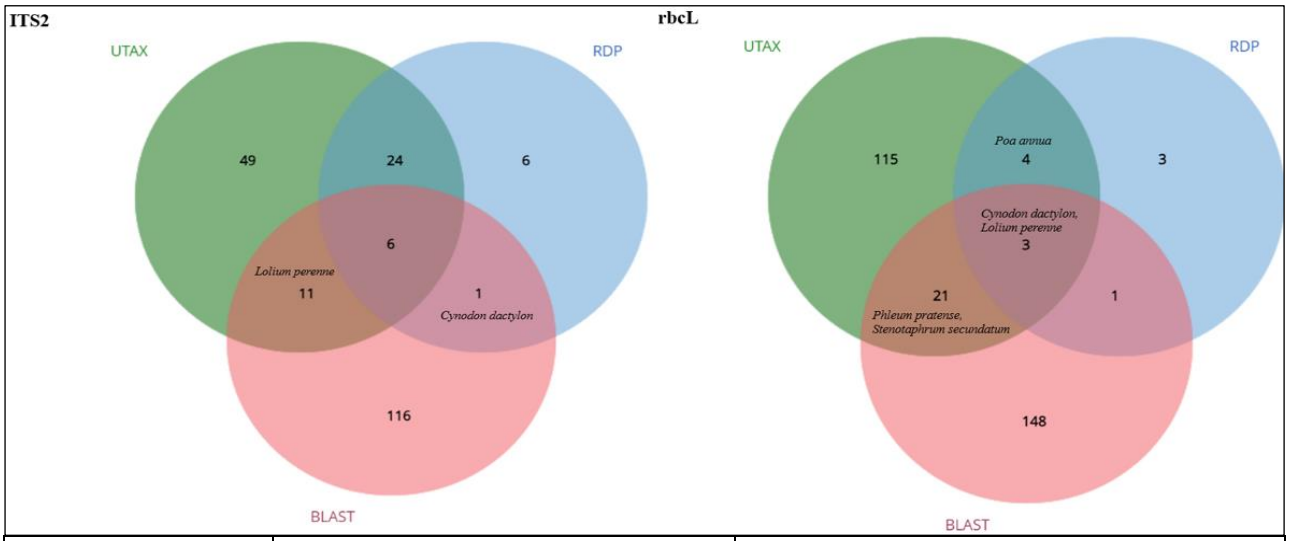
Figure 5.13. Stacked bar plots showing the relative abundance of the *Ambrosia* species identified in each of the *rbcL* samples from positive classifications with UTAX, RDP and BLAST. Created with RStudio [88].

Ambrosia species of interest, *A. artemisifolia* and *A. trifida* were identified from samples from Cape Town (C176, C184,) and Durban (D165, D166). Broader diversity was observed with UTAX classification with distinct relativity between *A. artemisifolia* and *A. trifida*. The RDP tool encountered some challenges when classifying some sequences, grouping them as unclassified 'NA'. This was a clear indication that it lacked sufficient resolution to assign some sequences confidently. The broader diversity of the *Ambrosia* species identified with BLAST reflected on the higher resolution of the tool which is likely due to its dependence on extensive reference databases. However, the increased diversity in the classifications with BLAST impacts direct comparisons across samples and therefore raises the possibility of over-classification.

The only sample that consistently showed positive identifications of the *Ambrosia* species of interest was from Durban, which corresponds with the findings of microscopically analysed pollen from the Durban monitoring site [115].

5.8.8. Summary of classification results

The names of the unique Poaceae species identified by each classification tool were extracted and matched to find the commonly identified species for *rbcL* and ITS2 classifications.



Classification/barcode	ITS2	*rbcL*
UTAX + RDP + BLAST	*Melinis repens, Triticum aestivum,* ***Eragrostis curvula****, Lolium multiflorum, Secale cereale, Triticum turgidum*	*Paspalum dilatatum,* ***Cynodon dactylon, Lolium perenne***
UTAX + RDP	*Pleioblastus chino, Lolium canariense, Setaria incrassate, Pseudostachyum polymorphum, Urochloa subulifolia, Eragrostis gummiflua, Panicum stapfianum, Themeda triandra, Chloris pectinate, Digitaria velutina, Eragrostis lehmanniana, Eragrostis superba, Fingerhuthia africana, Holcus lanatus Moorochloa eruciformis, Panicum schinzii, Panicum sumatrense, Paspalum vaginatum, Schmidtia kalahariensis, Schmidtia pappophoroides, Sporobolus africanus, Sporobolus ioclados, Triticum petropavlovskyi, Urochloa oligotricha*	*Themeda triandra,* ***Poa annua****, Ehrharta calycina, Setaria viridis*
UTAX + BLAST	*Elymus stipifolius, Paspalum dilatatum, Cenchrus americanus, Lolium rigidum,* ***Lolium perenne****, Pseudoroegneria strigose, Eragrostis pectinacean, Cenchrus purpureus, Cynodon incompletes, Eragrostis minor, Eragrostis pilosa*	*Paspalum thunbergia,* ***Phleum pratense****, Thysanolaena latifolia, Cenchrus compressus, Eragrostis tenuifolia, Echinochloa crus-galli, Lolium multiflorum, Hordeum brevisubulatum, Tripsacum dactyloides, Cenchrus clandestinus, Eragrostis ferruginea,* ***Stenotaphrum secundatum****, Zea mays, Cenchrus longisetus, Eragrostis cilianensis, Dimeria manongarivensis, Lolium temulentum, Paspalum conjugatum, Eragrostis japonica, Centotheca lappacea, Eragrostis pilosa*
RDP + BLAST	***Cynodon dactylon***	*Imperata cylindrica*

Figure 5.14. Comparison of the unique Poaceae species classified using UTAX, RDP, and BLAST to identify the taxa commonly identified by the tools in the classification of ITS2 and *rbcL* sequences. The Venn diagram shows the overlapping species identified across the tools, while the table lists the species that were commonly identified by two or more classification methods for both ITS2 and *rbcL* sequences. Venn diagram created with Venny version 1.0 [116].

Cynodon dactylon and *Lolium perenne* were commonly identified across all three classification tools using *rbcL*, while *Phleum pratense* and *Stenotaphrum secundatum* were identified with both UTAX and BLAST. Additionally, *Cynodon dactylon* was identified using UTAX and RDP. For ITS2 sequences, *Lolium perenne* was identified with both UTAX and BLAST, while *Cynodon dactylon* was identified using RDP and BLAST. It is important to note that there were more common species identified with UTAX and BLAST from *rbcL* sequences, whereas there were more commonalities observed between UTAX and RDP from ITS2 sequences, as shown in Figure 5.13.

Based on the SANBI data for Poaceae species in South Africa, the ratio of species confirmed in South Africa to the species identified with metabarcoding ratios were as follows: For *rbcL* sequences UTAX identified 52/143 species, RDP identified 11/11, and BLAST identified 61/173. For ITS2 sequences, UTAX identified 38/90 species, RDP identified 27/37, and BLAST identified 45/134. RDP was more precise in its identifications despite having fewer overall identifications, it was better at identifying South African species. UTAX, meanwhile, identified more species as compared with RDP but with less regional relevance. BLAST identified the most species overall, but many were likely not relevant to South Africa, indicating the most recently developed ITS2 and *rbcL* reference database used were broad but less targeted at our local species.

5.8.9. Validation of the accuracy of the matched and identified species

The performance of these taxonomic classification tools was evaluated across two barcode sequences of *rbcL* and ITS2. Each tool applied different criteria for positive identification, with UTAX using a raw score threshold of 80, RDP requiring a confidence score of 0.8 or higher, and BLAST relying on an E-value of ≤1.00e-5 and the value for each taxa assignment per sample recorded in Table 5.2.

Table 5.2. Classification tools performance scores as a measure of positive assignment used to assign taxa to targeted species at each monitoring site for the *rbcL* and ITS2 sequences analysed.

				UTAX		RDP		BLAST	
Classification tool Threshold for positive matches				Raw score ≥ 80		Confidence score ≥ 0.8		E-value ≤1.00e-5	
Targeted species	Monitoring Site	Period Collected	SampleID	*rbcL*	ITS2	*rbcL*	ITS2	*rbcL*	ITS2
Cynodon dactylon	Cape Town	06-Nov-22	C169 S14	50.0		1.0			
	Cape Town	25-Dec-22	C176 S17	50.4		1.0			
	Cape Town	25-Dec-22	C176 S18		56.2		1.0		5.64e-56
	Cape Town	01-Jan-23	C177 S20			1.0			
	Cape Town	01-Jan-23	C177 S21		58.3		1.0		6.03e-56
	Cape Town	19-Feb-23	C184 S24		57.4		1.0		4.02e-53
	Kimberley	02-Oct-22	K164 S1	50.0		1.0			
	Kimberley	09-Oct-22	K165 S5	50.0		1.0			
	Kimberley	09-Oct-22	K165 S16		60.3		1.0		
	Kimberley	16-Oct-22	K166 S8	50.8		1.0			
	Kimberley	16-Oct-22	K166 S19		63.6		1.0		
	Kimberley	27-Nov-22	K172 S11	50.0		1.0			
	Kimberley	27-Nov-22	K172 S22		61.0		1.0		
	Bloemfontein	26-Mar-23	B189 S12		53.2				6.13e-56
	Durban	09-Oct-22	D165 S2	50.0		1.0		2.65e-105	
	Durban	16-Oct-22	D166 S3		58.0			2.78e-105	
	Durban	16-Oct-22	D166 S6	50.0		1.0			
	Johannesburg	02-Apr-23	J190 S13		56.1				5.26e-68
Eragrostis curvula	Cape Town	01-Jan-23	C177 S21		54.7		1.0		
	Kimberley	09-Oct-22	K165 S16		62.0				
	Kimberley	16-Oct-22	K166 S19			1.0		3.18e-86	
	Kimberley	27-Nov-22	K172 S22		61.4		1.0		4.12 e-104
	Bloemfontein	26-Mar-23	B189 S12		50.0		1.0		3.12e-62
	Durban	16-Oct-22	D166 S3	50.0		1.0			
	Johannesburg	02-Apr-23	J190 S13		54.1		1.0		1.69 e-78
Lolium perenne	Cape Town	06-Nov-22	C169 S14	49.2		1.0		8.16e-12	
	Cape Town	25-Dec-22	C176 S17			1.0			
	Cape Town	25-Dec-22	C176 S18		51.9				4.35e-109
	Cape Town	01-Jan-23	C177 S20			1.0			
	Cape Town	01-Jan-23	C177 S21		52.7				1.77e-128
	Cape Town	19-Feb-23	C184 S23			1.0			
	Kimberley	16-Oct-22	K166 S8			1.0			
	Kimberley	16-Oct-22	K166 S19		51.3				
	Kimberley	27-Nov-22	K172 S11			1.0			
	Durban	09-Oct-22	D165 S2			1.0		4.77e-128	
	Durban	16-Oct-22	D166 S6			1.0			
Poa annua	Cape Town	06-Nov-22	C169 S14	57.1		0.99			

	Kimberley	09-Oct-22	K165 S16		48.2				
	Kimberley	16-Oct-22	K166 S8			1.0			
	Kimberley	16-Oct-22	K166 S19		50.3		1.0		
Stenotaphrum secundatum	Cape Town	25-Dec-22	C176 S17	44.2					
	Durban	09-Oct-22	D165 S2	44.5				2.16e-126	
	Durban	16-Oct-22	D166 S6	44.7					
Paspalum notatum	Johannesburg	02-Apr-23	J190 S13		50.0				4.93e-101
Phragmites australis	Cape Town	25-Dec-22	C176 S18		62.0				5.05e-58
	Cape Town	19-Feb-23	C184 S23		50.0				6.43e-113
Phleum pratense	Cape Town	06-Nov-22	C169 S14	50.0				4.75e-71	
Hyparrhenia hirta	Durban	09-Oct-22	D165 S2	45.3					
	Durban	16-Oct-22	D166 S6	49.0				2.72e-105	
Ambrosia artemisiifolia	Durban	09-Oct-22	D165 S2	52.1		0.99		1.08e-134	
	Durban	16-Oct-22	D166 S6	52.0		1.0			
	Cape Town	25-Dec-22	C176 S17	58.2					
	Cape Town	19-Feb-23	C184 S23	42.0					
Ambrosia trifida	Durban	09-Oct-22	D165 S2	44.5		0.97		1.08e-134	
	Durban	16-Oct-22	D166 S6	44.6		0.98			
	Cape Town	25-Dec-22	C176 S17	55.0		1.0			
	Cape Town	19-Feb-23	C184 S23	42.1					

For *Cynodon dactylon* (Bermuda grass), the *rbcL* barcode yielded consistent UTAX identifications across 15 samples which included samples from Cape Town collected between October 2022 (C169) and February 2023 (C184), Kimberley between September 2022 (K164) and November 2022 (K172), Durban in October 2022 (D165/D166), Bloemfontein (B189) and Johannesburg (J190) in March and April 2023 respectively. However, none of the raw scores met the required threshold of 80, suggesting limited accuracy for this tool. RDP produced more reliable results, with ITS2 sequences consistently identifying Bermuda grass with a confidence score of 1.0 across all samples, and BLAST further confirmed these matches with extremely low E-values (e.g., 5.64×10^{-56}). This indicated that ITS2 sequences were more effective for the identification of Bermuda grass than *rbcL* when classifying the sequences using BLAST and RDP.

UTAX also failed to reach the threshold for positive identification of *Lolium perenne* (ryegrass) when classifying *rbcL* sequences. However, with ITS2, RDP successfully identified the species in several samples with a perfect confidence score of 1.0, while BLAST produced very low E-values that further validated the identifications. Therefore, ryegrass was reliably identified with ITS2 sequences using both RDP and BLAST. The identification of *Poa annua* followed a similar trend, however, ITS2 proved more effective, as RDP identified the species with high confidence in multiple samples, while BLAST did not return significant results. This highlights the effectivity of RDP for classifying ITS2-based sequences for the identification of *Poa annua*.

With *Stenotaphrum secundatum* (buffalo grass), neither UTAX nor RDP produced positive identifications, but BLAST returned significant results for ITS2 sequences, with E-values as low as 2.16×10^{-126}. suggesting that BLAST was the most sensitive tool for detecting buffalo grass in ITS2 sequences, even when other tools fail to provide matches. The identification of *Phleum pratense* was similarly successful with BLAST, which produced highly significant matches using *rbcL* (e.g., C169 S14: 4.75×10^{-71}). However, neither UTAX nor RDP yielded positive identifications for this species, indicating that BLAST may be better suited for identifying *Phleum pratense* in these samples. *Hyparrhenia hirta* identifications also followed this pattern, with UTAX returning low scores that did not meet the threshold for positive identification, while BLAST successfully identified the species with significant E-values (D166 S6: 2.72×10^{-105}). This further showed BLAST's sensitivity in detecting species that other tools may miss. *Ambrosia artemisiifolia* and *Ambrosia trifida* species were classified with UTAX from sequences of *rbcL* but identified with raw scores that were below threshold of positive identification. However, they were successfully identified using RDP from ITS2 sequences with confidence scores above 0.95, while BLAST returned extremely low E-values, confirming they were positive identifications from *rbcL* sequences.

5.8.10. Statistical evaluation of sequence reads processing

A total of 13 389 871 sequences were generated by the sequencer, with 33% belonging to the *rbcL* samples, 53% to ITS2, and the remainder from the 'undetermined' sample which consisted of both *rbcL* or ITS2 reads.

The total number of sequences reads, their statistical measures (sum, mean, standard deviation (SD), and median), and their distribution across different barcodes (*rbcL* and ITS2) and undetermined sequences were calculated and detailed in Table 5.3. The data divided into raw sequences (generated by the sequencer), filtering output (sequences that remained after filtering) and classified sequences statistical metrics is summarised in Table 5.3.

Table 5.3. Summary of the total sequence reads number across the major stages of analysis.

Sequence reads	Sum	Mean	SD	Median
Raw				
All sequences	13389871	514995	285377	448022
rbcL	4467934 (33%)	446793	76582	438346
ITS2	7074562 (53%)	471637	97566	451520
undetermined	1847375 (14%)			
Filtering output				
All sequences	2592290	99704	54563	110146
rbcL	1378131 (53%)	137813	23549	141439
ITS2	1068609 (41%)	71241	53891	70936
undetermined	145550 (6%)			
Classified				
UTAX classified				
UTAX – *rbcL*	2253073	90123	47998	97923
UTAX – ITS2	2517597	96831	54269	105291
RDP classified				
RDP – *rbcL*	2411618	96465	53530	106596
RDP – ITS2	2558635	98409	54379	107528

A total of 2 517 597 ITS2 sequences classified with UTAX yielded 90 unique Poaceae species, while 2 253 073 *rbcL* sequences classified yielded 143 Poaceae species, along with 3 *Ambrosia* species. Despite a higher number of ITS2 sequences compared to *rbcL* sequences, the *rbcL* marker identified more Poaceae species and were the only sequences that identified *Ambrosia*, suggesting that *rbcL* sequences may provide more taxonomic resolution or broader species coverage in this context. RDP classifications identified fewer Poaceae species overall, however *Ambrosia* species could be identified from *rbcL* sequences. From 2 558 635 ITS2 sequences classified, only 37 Poaceae species were identified along and from 2 411 618 *rbcL* sequences classified only 11 Poaceae species with 1 *Ambrosia* species were classified. The lower species count in RDP may be attributed to the tool's design, as it was primarily developed for classifying rRNA sequences, which includes ITS2 [78]. This focus on rRNA sequences may introduce a bias, limiting its ability to classify non-rRNA sequences such as those from *rbcL*. From BLAST classifications, more Poaceae species including species of *Ambrosia* were identified from *rbcL* sequences compared to ITS2 sequences, supporting the earlier observations that *rbcL* sequences provide broader taxonomy in comparison to ITS2.

5.9. Discussion

Metabarcoding for species identification relies on the quality of sequencing or sequenced products which are determined by the preparation measures. Although the fragmentation of the sequence amplicons did not completely hinder the identification of species from ITS2 sequences it was clear that it limited its capabilities in detecting more diversity from the samples.

When comparing the species identified with the three classifications tools, *rbcL* sequences consistently yielded a higher diversity and consistently identified species of interest across the three classification tools. All the samples of Cape Town (7), Kimberley (7) and Durban (3) had Poaceae species identified. However, from the Bloemfontein (3) and Johannesburg (3) samples that were analysed, only one sample from each had Poaceae species identified from ITS2 sequences. Despite *rbcL* barcode's broader species detection, the ITS2 barcode consistently outperformed *rbcL* in identifying the species of interest such as species of *Cynodon*, *Lolium, Poa* and *Ambrosia*. This was a clear indication of the complementary role between the two barcodes, with *rbcL* offering broader species coverage and ITS2 excelling in finer taxonomic distinctions for certain genera.

The classification approach was a comparative analysis looking at three taxonomic classification tools: UTAX, RDP, and BLAST. Each tool relied on barcode reference databases built from NCBI-deposited Viridiplantae sequences. The UTAX and RDP reference databases were curated in 2016 for *rbcL* sequences and 2020 for ITS2 sequences [78], [79], [82], whereas the reference databases used in BLAST classifications were developed more recently, in 2021 [80]. RDP and BLAST were the most reliable of the classification tools explored in this study, with RDP providing confident identifications and BLAST correlating with the very low E-values. BLAST consistently outperformed both UTAX and RDP in species identification, particularly for *rbcL* sequences. This stresses out the need for more comprehensive and up-to-date reference databases, like the *rbcL* and ITS2 reference databases developed by Dubois *et al* 2022 [80] that were used in BLAST classifications in every metabarcoding analysis as it yielded a significantly broader biodiversity, whether this is from the power of the BLAST classification tool, it remains unclear. Additionally, because BLAST identified the most number of species but most not present in South Africa, it suggests that BLAST assignment in this database lacks specificity. Another contributing factor the inability to map the species detected with BLAST in South African biodiversity is the incomplete records of the SANBI.

The metabarcoding identification success was relatively high in identifying grasses, including *Cynodon dactylon* (Bermuda grass), *Eragrostis curvula* (weeping love grass), *Lolium perenne* (perennial ryegrass), *Poa annua* (annual bluegrass), *Stenotaphrum secundatum* (buffalo grass), and *Paspalum notatum* (Bahia grass), with multiple species detected across different regions at their flowering period. *Cynodon dactylon* was successfully identified in all five monitoring sites (Cape Town, Kimberley, Bloemfontein, Johannesburg, Durban), showcasing its wide distribution and the reliability of the metabarcoding approach for its detection. *Eragrostis curvula* was detected in Kimberley, Bloemfontein, Durban, and Johannesburg, further indicating the metabarcoding's efficiency and ability to identify widely spread species across diverse geographic areas. *Lolium perenne* was found in Cape Town, Kimberley, and Durban, regions where its presence has been confirmed. *Poa annua* was identified in Cape Town and Kimberley, reflecting its presence in these specific locations. *Stenotaphrum secundatum* was detected in Cape Town and Durban, highlighting its distribution in these regions. *Paspalum notatum* was identified exclusively in Johannesburg, demonstrating the ability of the metabarcoding method to pinpoint species at specific sites.

Additionally, the metabarcoding method proved particularly effective in identifying allergenic species, such as *Ambrosia artemisiifolia* (common ragweed) and *Ambrosia trifida* (giant ragweed), which are known contributors to seasonal pollen allergies. Both species were successfully detected in Durban, a region where its pollen has been suggested to be found since the established of SAPNET.

However, there were challenges with identifying certain species which are important contributors to grass pollen allergies. Species such as *Avena fatua, Briza maxima, Briza minor, Lagurus ovatus, Pennisetum clandestinum,* and *Sorghum halepense* were not identified. This could be attributed to limitations in current reference databases or the quality of DNA recovered from environmental samples.

Chapter 6

Conclusion, limitations and future work

6.1. Conclusion

With this pilot study, we were able to establish a method of extracting DNA of sufficient quantity and quality which consequently allowed for the amplification of barcode markers that were used to successfully identify the species of grasses and ragweed in South Africa using Next-generation sequencing. Site-specific pollen identification success varied across the five selected sites of pollen monitoring (Bloemfontein, Cape Town, Durban, Johannesburg and Kimberley) with Cape Town and Durban showing particularly high identification rates of many of the target species. While a higher DNA quantity and quality are generally expected to improve the number of taxa identified, our findings suggest that there is no correlation between a yield and purity of samples with the taxa identified since Durban did not have samples of greater yield or higher purity. Cape Town identified five of the targeted species, while Durban showed broad species diversity, with several grass species as well as both *Ambrosia* species detected.

Kimberley and Johannesburg also showed solid identification success, with Kimberley detecting four species, including *Poa annua* and *Eragrostis curvula*, and Johannesburg identifying *Cynodon dactylon* and *Paspalum notatum*. Bloemfontein presented a more limited outcome, with only *Eragrostis curvula* identified. While we saw it was necessary to optimise the DNA extraction method for a greater yield, our results indicate that aiming for a greater quantity exceeding the required concentration for PCR does not necessarily improve species detection. However, maintaining a higher initial DNA allows for aliquoting to ensure for consistency across reactions and helps to prevent sample loss. Preserving aliquots should be standard practise to avoid irreversible sample loss especially since pollen availability varies by season at each monitoring site. This approach could aid in improving the reliability and reproducibility of the sequencing results, ensuring that samples can be reanalysed should the need arise.

In addition to allergenic grass species, the study also successfully identified *Ambrosia artemisiifolia* (common ragweed) and *Ambrosia trifida* (giant ragweed), both of which are significant contributors to pollen allergies in many regions. The reliable detection of these species across different samples further highlights the potential of metabarcoding in pollen monitoring as it relates with pollen allergies, where identifying airborne allergens in each area is important for both public health awareness and medical interventions.

However, *Phleum pratense* and *Ambrosia psilostachya* a species not known to occur in our southern regions were positively identified from ITS2 and *rbcL* sequence classifications using BLAST. This raised concerns about potential sequence mismatches leading to incorrect taxa assignments, or the possibility of false identification from the ITS2 reference databases that were used. But if that were the case, how could we be certain that the positively targeted identified species were correct matches or accurately assigned taxa.

Although the study was a success, we also encountered challenges such as unidentified species and site-specific variation, which showed areas that both methodology and reference databases bioinformatical analysis could be improved. Our results confirm the value of metabarcoding as a tool for environmental monitoring, species identification, and allergy research, with promising applications for future studies in diverse regions and biomes.

6.2. Limitations

The study only covered six regions and a relatively small number of samples (n = 25). Potchefstroom was excluded from analysis due to quality filtering, further reducing the sample diversity. This limited geographic and sample coverage may affect the generalisability of the results to other regions. The samples were collected during specific periods within the Poaceae pollen season and not may not reflect the full range of species that could be identified throughout the year, potentially limiting the understanding of seasonal variation in airborne pollen diversity.

The wrong method used for the preparation of the DNA library for sequencing led to fragmentation of the amplicons, the reassembling of the reads may have influenced the results obtained. Although not yet confirmed, there is high likelihood that chimeras of the amplicons could have been created largely reducing the confidences of correct taxonomic assignment.

6.3. Future work

To address the limitations identified and improve the reliability of species identification in future studies, several points of action can be applied for the improvement of metabarcoding pollen analysis, and these include the following:

- Prioritising improved communication with sequencing providers to ensure proper handling of samples, specifically in avoiding fragmentation and producing longer reads. This would help minimise the loss of sequences due to 'shorter-than-expected' reads and enable deeper and more accurate analysis.
- Using up to date reference databases for UTAX and RDP taxonomic classifications. Updating these databases for *rbcL* and ITS2 barcodes, will be essential for better classification accuracy. Collaboration with global initiatives to incorporate newly discovered species and frequent updates to reference databases will help ensure that tools like RDP and UTAX remain relevant and reliable for species classification and applicability in South Africa.
- Expanding the geographic and seasonal scope of sampling would provide a broader understanding of Poaceae pollen diversity in South Africa. In the future we should also consider sampling throughout different seasons to improve on the generalisation of the findings and capture seasonal variations of the species.
- Future work should consider larger sampling for broader coverage to improve statistical power and reliability of results for a more accurate assessment of biodiversity in different SAPNET biomes and improve the detection of rare or region-specific species.

6.4. Literature Cited

[1] ChatGPT, "OpenAI's ChatGPT: A Revolution in Language AI."

[2] Ordman D, "Pollinosis in South Africa; a study of seasonal hay-fever," *South African Med. J.*, vol. 21, no. 2, pp. 38–48, 1947.

[3] Ordman D, "Seasonal respiratory allergy and the associated pollens in South Africa," *South African Med. J.*, vol. 37, pp. 321–325, 1963.

[4] Ordman D, "The Regional Aspects of Respiratory Allergy in South Africa," *South African Med. J.*, vol. 38, pp. 369–372, 1964.

[5] Potter C, "Common indoor and outdoor aero-allergens," *C. - South African Med. J.*, vol. 28, no. 9, Sep. 2010.

[6] Cadman A, "Incidence of atmospheric pollen in the Pretoria-Witwatersrand-

Vereeniging region during 1987/1988," *South African Med. J.*, vol. 79, no. 2, pp. 84–87, 1991.

[7] Esterhuizen N *et al.*, "The South African Pollen Monitoring Network: Insights from 2 years of national aerospora sampling (2019–2021)," *Clin. Transl. Allergy*, vol. 13, no. 11, Nov. 2023, doi: 10.1002/clt2.12304.

[8] Ajikah L, Neumann F, Berman D, and Peter J, "Aerobiology in South Africa: A new hope!," *South Africa J. Sci.*, vol. 116, no. 7, p. 4, Jul. 2020, [Online]. Available: https://doi.org/10.17159/ sajs.2020/8112

[9] Alca P and García-Mozo H, "Trends in grass pollen season in southern Spain," pp. 157–169, 2010, doi: 10.1007/s10453-009-9153-3.

[10] Van Oudtshoorn F, *Guide to Grasses of Southern Africa*. Briza Publications, 2012.

[11] Fish L, Mashau AC, Moesha MJ, and Nembudani MT, *Identification guide to southern African grasses. An identification manual with keys, descriptions and distributions.*, vol. 36. South African National Biodiversity Institute, Pretoria: Strelitzia, 2015.

[12] Bastl K, Bastl M, Berger M, Dirr L, and Berger UE, "Phenology as a tool to gain more insights into the grass pollen season," Feb. 01, 2024. doi: 10.1007/s40629-023-00264-7.

[13] Bastl M, Bastl K, Dirr L, Berger M, and Berger U, "Variability of grass pollen allergy symptoms throughout the season: Comparing symptom data profiles from the Patient's Hayfever Diary from 2014 to 2016 in Vienna (Austria)," *World Allergy Organ. J.*, vol. 14, no. 3, Mar. 2021, doi: 10.1016/j.waojou.2021.100518.

[14] Frisk CA *et al.*, "Microscale pollen release and dispersal patterns in flowering grass populations," *Sci. Total Environ.*, vol. 880, Jul. 2023, doi: 10.1016/j.scitotenv.2023.163345.

[15] García-Mozo H, "Poaceae pollen as the leading aeroallergen worldwide: A review," *Allergy*, no. 72, pp. 1849–1858, May 2017, doi: 10.1111/all.13210.

[16] Shoko C and Mutanga O, "Seasonal discrimination of C3 and C4 grasses functional types: An evaluation of the prospects of varying spectral configurations of new generation sensors," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 62, pp. 47–55, 2017, doi: 10.1016/j.jag.2017.05.015.

[17] Luschkova D, Traidl-Hoffmann C, and Ludwig A, "Climate change and allergies," Jun. 01, 2022, *Springer Medizin*. doi: 10.1007/s40629-022-00212-x.

[18] Choi YJ, Lee KS, and Oh JW, "The Impact of Climate Change on Pollen Season and Allergic Sensitization to Pollens," Feb. 01, 2021, *W.B. Saunders*. doi: 10.1016/j.iac.2020.09.004.

[19] Ziska LH and Beggs PJ, "Anthropogenic climate change and allergen exposure: The role of plant biology," *J. Allergy Clin. Immunol.*, vol. 129, no. 1, pp. 27–32, 2012, doi: 10.1016/j.jaci.2011.10.032.

[20] Emberlin J, "The effects of patterns in climate and pollen abundance on allergy," *Allergy*, vol. 49, no. s18, pp. 15–20, 1994, doi: https://doi.org/10.1111/j.1398-9995.1994.tb04233.x.

[21] Neumann FH *et al.*, "Ecological and Allergenic Significance of Atmospheric Pollen Spectra from a Grassland-Savanna Ecotone in North-West Province (South Africa)," *Palynology*, Oct. 2024, doi: 10.1080/01916122.2024.2411234.

[22] Berman D, "Regional-specific pollen and fungal spore allergens in South Africa," *Curr. Allergy Clin. Immunol.*, vol. 26, no. 4, 2013.

[23] Frisk CA, Adams-Groom B, and Smith M, "Isolating the species element in grass pollen allergy: A review," *Sci. Total Environ.*, vol. 883, no. 163661, Apr. 2023, doi: 10.1016/j.scitotenv.2023.163661.

[24] Kailaivasan T and Davies JM, "The molecular allergology of subtropical grass pollen," *Mol. Immunol.*, vol. 100, no. March, pp. 126–135, 2018, doi: 10.1016/j.molimm.2018.03.012.

[25] Pablos I., Wildner S., Asam C., Wallner M., and Gadermaie G., "Pollen Allergens for Molecular Diagnosis," *Curr. Allergy Asthma Rep.*, vol. 16, no. 4, 2016, doi: 10.1007/s11882-016-0603-z.

[26] de Weger LA, van Hal PW, Bos B, Molster F, Mostert M, and Hiemstra PS, "Personalized Pollen Monitoring and Symptom Scores: A Feasibility Study in Grass Pollen Allergic Patients," *Front. Allergy*, vol. 2, 2021, doi: 10.3389/falgy.2021.628400.

[27] Zhu X *et al.*, "Floating in the air: forecasting allergenic pollen concentration for managing urban public health," 2024, *Taylor and Francis Ltd.* doi: 10.1080/17538947.2024.2306894.

[28] Schüler L and Behling H, "Poaceae pollen grain size as a tool to distinguish past grasslands in South America: A new methodological approach," *Veg. Hist. Archaeobot.*, vol. 20, no. 2, pp. 83–96, Sep. 2011, doi: 10.1007/s00334-010-0265-z.

[29] Roubik DW and Jorge Enrique Moreno P, "Pollen and Spores of Barro Colorado Island," *Kew Bull.*, vol. 47, no. 4, p. 791, Jan. 1992, doi: 10.2307/4110734.

[30] Joly C, Barillé L, Barreau M, Mancheron A, and Visset L, "Grain and annulus diameter as criteria for distinguishing pollen grains of cereals from wild grasses," *Rev. Palaeobot. Palynol.*, vol. 146, no. 1–4, pp. 221–233, 2007, doi: 10.1016/j.revpalbo.2007.04.003.

[31] Chen K, Marusciac L, Tamas PT, Valenta R, and Panaitescu C, "Ragweed Pollen Allergy: Burden, Characteristics, and Management of an Imported Allergen Source in Europe," Jul. 01, 2018, *S. Karger AG*. doi: 10.1159/000487997.

[32] Robbins RR, Dickinson DB, and Rhodes AM, "Morphometric analysis of pollen from four species of Ambrosia (Compositae)," *Am. J. Bot.*, vol. 66, no. 5, pp. 538–545, 1979, doi: https://doi.org/10.1002/j.1537-2197.1979.tb06256.x.

[33] Albertini R, Veronesi L, Colucci ME, and Pasquarella C, "The scenario of the studies on ragweed (Ambrosia Sp.) and related issues from its beginning to today: a useful tool for future goals in a one health approach," *Acta Biomed.*, vol. 93, no. 5, 2022, doi: 10.23750/abm.v93i5.13771.

[34] Yair Y, Sibony M, Goldberg A, Confino-Cohen R, Rubin B, and Shahar E, "Ragweed species (Ambrosia spp.) in Israel: distribution and allergenicity," *Aerobiologia (Bologna).*, vol. 35, no. 1, pp. 85–95, 2019, doi: 10.1007/s10453-018-9542-6.

[35] Suanno C, Aloisi I, Fernández-González D, and Del Duca S, “Monitoring techniques for pollen allergy risk assessment,” Jun. 01, 2021, *Academic Press Inc.* doi: 10.1016/j.envres.2021.111109.

[36] Richter R *et al.*, “Spread of invasive ragweed: Climate change, management and how to reduce allergy costs,” *J. Appl. Ecol.*, vol. 50, no. 6, pp. 1422–1430, Dec. 2013, doi: 10.1111/1365-2664.12156.

[37] Oswalt ML and Marshall GD, “Ragweed as an example of worldwide allergen expansion,” Sep. 2008. doi: 10.2310/7480.2008.00016.

[38] Asero R, Weber B, Mistrello G, Amato S, Madonini E, and Cromwell O, “Giant ragweed specific immunotherapy is not effective in a proportion of patients sensitized to short ragweed: Analysis of the allergenic differences between short and giant ragweed,” *J. Allergy Clin. Immunol.*, vol. 116, no. 5, pp. 1036–1041, Nov. 2005, doi: 10.1016/j.jaci.2005.08.019.

[39] Kato Y, Akasaki S, Muto-Haenuki Y, Fujieda S, Matsushita K, and Yoshimoto T, “Nasal sensitization with ragweed pollen induces local-allergic-rhinitis- like symptoms in mice,” *PLoS One*, vol. 9, no. 8, Aug. 2014, doi: 10.1371/journal.pone.0103540.

[40] Bocsan IC *et al.*, “Characterization of patients with allergic rhinitis to ragweed pollen in two distinct regions of Romania,” *Med.*, vol. 55, no. 11, Nov. 2019, doi: 10.3390/medicina55110712.

[41] Jones NR *et al.*, “Ragweed pollen and allergic symptoms in children: Results from a three-year longitudinal study,” *Sci. Total Environ.*, vol. 683, pp. 240–248, Sep. 2019, doi: 10.1016/j.scitotenv.2019.05.284.

[42] Rowney FM *et al.*, “Environmental DNA reveals links between abundance and composition of airborne grass pollen and respiratory health,” *Curr. Biol.*, vol. 31, no. 9, pp. 1995-2003.e4, May 2021, doi: 10.1016/j.cub.2021.02.019.

[43] Hebert PDN, Cywinska A, Ball SL, and DeWaard JR, “Biological identifications through DNA barcodes,” *Proc. R. Soc. B Biol. Sci.*, vol. 270, no. 1512, pp. 313–321, 2003, doi: 10.1098/rspb.2002.2218.

[44] Taberlet P *et al.*, “Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding,” *Nucleic Acids Res.*, vol. 35, no. 3, Feb. 2007, doi: 10.1093/nar/gkl938.

[45] CBOL Plant Working Group, “A DNA barcode for land plants,” *Proc. thNational Acad. Sci. United States Am.*, vol. 106, no. 31, pp. 12794–12797, Aug. 2009, doi: 10.1073/pnas.0905845106.

[46] Chen S *et al.*, “Validation of the ITS2 Region as a Novel DNA Barcode for Identifying Medicinal Plant Species,” vol. 5, no. 1, pp. 1–8, 2010, doi: 10.1371/journal.pone.0008613.

[47] Pornon A *et al.*, “Using metabarcoding to reveal and quantify plant-pollinator interactions,” *Nat. Publ. Gr.*, no. May, pp. 1–12, 2016, doi: 10.1038/srep27282.

[48] Zimmermann J, Glöckner G, Jahn R, Enke N, and Gemeinholzer B, “Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies,” *Mol. Ecol. Resour.*, vol. 15, no. 3, pp. 526–542, 2015, doi: 10.1111/1755-0998.12336.

[49] Gous A, Swanevelder DZH, Eardley CD, and Willows-Munro S, “Plant–pollinator interactions over time: Pollen metabarcoding from bees in a historic collection,” *Evol. Appl.*, vol. 12, no. 2, pp. 187–197, 2019, doi: 10.1111/eva.12707.

[50] Leontidou K, Vernesi C, De Groeve J, Cristofolini F, Vokou D, and Cristofori A, “Taxonomic identification of airborne pollen from complex environmental samples by DNA metabarcoding: a methodological study for optimizing protocols,” *Aerobiologia (Bologna).*, p. 099481, 2017, doi: 10.1007/s10453-017-9497-z.

[51] Hawkins J *et al.*, “Using DNA metabarcoding to identify the floral composition of honey: A new tool for investigating honey bee foraging preferences,” *PLoS One*, vol. 10, no. 8, Aug. 2015, doi: 10.1371/journal.pone.0134735.

[52] Simel EJ, Saidak LR, and Tuskan GA, “Method of extracting dsDNA from non-germinated gymnosperm and angiosperm pollen,” *Biotechniques*, vol. 22, no. 3, pp. 390–394, 1997, doi: 10.2144/97223bm02.

[53] Leontidou K *et al.*, “Efficient and sensitive identification and quantification of airborne pollen using next-generation DNA sequencing,” *Sci. Rep.*, vol. 15, no. 1, pp. 8–16, 2019, doi: 10.1111/1755-0998.12288.

[54] Cheng T, Xu C, Lei L, Li C, Zhang Yu, and Zhou S, “Barcoding the kingdom Plantae: New PCR primers for ITS regions of plants with improved universality and specificity,” *Mol. Ecol. Resour.*, vol. 16, no. 1, pp. 138–149, Jan. 2016, doi: 10.1111/1755-0998.12438.

[55] Dong W *et al.*, “Discriminating plants using the DNA barcode rbcLb: An appraisal based on a large data set,” *Mol. Ecol. Resour.*, vol. 14, no. 2, pp. 336–343, 2014, doi: 10.1111/1755-0998.12185.

[56] Bell KL *et al.*, “Pollen DNA barcoding: Current applications and future prospects,” *Genome*, vol. 59, no. 9, pp. 629–640, 2016, doi: 10.1139/gen-2015-0200.

[57] Bruni I *et al.*, “A DNA barcoding approach to identify plant species in multiflower,” *FOOD Chem.*, vol. 170, pp. 308–315, 2015, doi: 10.1016/j.foodchem.2014.08.060.

[58] Hornick T *et al.*, “An integrative environmental pollen diversity assessment and its importance for the Sustainable Development Goals,” *Plants People Planet*, vol. 4, no. 2, pp. 110–121, 2022, doi: 10.1002/ppp3.10234.

[59] Kelley L, Rose E, McCullough B, Martinez M, and Baudelet M, “Non-destructive DNA analysis of single pollen grains,” *Forensic Chem.*, vol. 20, no. April, p. 100275, 2020, doi: 10.1016/j.forc.2020.100275.

[60] Bell KL *et al.*, “Applying pollen DNA metabarcoding to the study of plant–pollinator interactions,” *Appl. Plant Sci.*, vol. 5, no. 6, Jun. 2017, doi: 10.3732/apps.1600124.

[61] Moore MA, Scheible MKR, Robertson JB, and Meiklejohn KA, “Assessing the lysis of diverse pollen from bulk environmental samples for DNA metabarcoding,” *Metabarcoding and Metagenomics*, vol. 6, pp. 293–303, 2022, doi: 10.3897/mbmg.6.89753.

[62] Moeller JR, Moehn NR, Waller DM, and Givnish TJ, “Paramagnetic cellulose DNA isolation improves DNA yield and quality among diverse plant taxa,” *Appl. Plant Sci.*, vol. 2, no. 10, Oct. 2014, doi: 10.3732/apps.1400048.

[63] Newmaster SG, Fazekas AJ, and Ragupathy S, “DNA barcoding in land plants: Evaluation of rbcL in a multigene tiered approach,” Mar. 2006. doi: 10.1139/B06-047.

[64] Duan H, Wang W, Zeng Y, Guo M, and Zhou Y, “The screening and identification of DNA barcode sequences for Rehmannia,” *Sci. Rep.*, vol. 9, no. 1, pp. 1–12, 2019, doi: 10.1038/s41598-019-53752-8.

[65] Loera-Sánchez M, Studer B, and Kölliker R, “DNA barcode trnH-psbA is a promising candidate for efficient identification of forage legumes and grasses,” *BMC Res. Notes*, vol. 13, no. 1, Jan. 2020, doi: 10.1186/s13104-020-4897-5.

[66] Kress WJ and Erickson DL, “A Two-Locus Global DNA Barcode for Land Plants: The Coding rbcL Gene Complements the Non-Coding trnH-psbA Spacer Region,” *PLoS One*, vol. 2, no. 6, 2007, doi: 10.1371/journal.pone.0000508.

[67] Hollingsworth ML *et al.*, “Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants,” *Mol. Ecol. Resour.*, pp. 439–457, Sep. 2009, doi: 10.1111/j.1755-0998.2008.02439.x.

[68] Peterson PM, Romaschenko K, and Soreng RJ, “A laboratory guide for generating DNA barcodes in grasses: a case study of Leptochloa s.l. (Poaceae: Chloridoideae),” *WebbWebbia J. Plant Taxon. Geogr.*, vol. 69, no. 1, pp. 1–12, May 2014, doi: 10.1080/00837792.2014.927555.

[69] Shneyer VS and Rodionov AV, “Plant DNA Barcodes,” vol. 9, no. 4, pp. 295–300, 2019, doi: 10.1134/S207908641904008X.

[70] Ghitarrini S *et al.*, “New biomolecular tools for aerobiological monitoring: Identification of major allergenic Poaceae species through fast real-time PCR,” *Ecol. Evol.*, vol. 8, no. 8, pp. 3996–4010, 2018, doi: 10.1002/ece3.3891.

[71] Blattner FR, “Direct Amplification of the Entire ITS Region from Poorly Preserved Plant Material Using Recombinant PCR,” *Biotechniques*, vol. 27, no. 6, pp. 1180–1184, Jul. 1999.

[72] Han J *et al.*, “The Short ITS2 Sequence Serves as an Efficient Taxonomic Sequence Tag in Comparison with the Full-Length ITS,” vol. 2013, pp. 3–10, 2013.

[73] Carneiro de Melo Moura C *et al.*, “Biomonitoring via DNA metabarcoding and light microscopy of bee pollen in rainforest transformation landscapes of Sumatra,” *BMC Ecol. Evol.*, vol. 22, no. 1, Dec. 2022, doi: 10.1186/s12862-022-02004-x.

[74] Arstingstall KA *et al.*, “Investigating the use of pollen DNA metabarcoding to quantify bee foraging and effects of threshold selection,” *PLoS One*, vol. 18, no. 4 April, Apr. 2023, doi: 10.1371/journal.pone.0282715.

[75] Martin GS, Hautier L, Mingeot D, and Dubois B, “How reliable is metabarcoding for pollen identification? An evaluation of different taxonomic assignment strategies by cross-validation,” *PeerJ*, vol. 12, 2024, doi: 10.7717/peerj.16567.

[76] Karsch-Mizrachi I, Takagi T, and Cochrane G, “The international nucleotide sequence database collaboration,” *Nucleic Acids Res.*, vol. 46, no. D1, pp. D48–D51, Jan. 2018, doi: 10.1093/nar/gkx1097.

[77] Lathe W, Williams J, Mangan M, and Karolchik D, “Genomic data resources: challenges and promises,” *Nat. Educ.*, vol. 1, no. 3, p. 2, 2008.

[78] Sickel W *et al.*, "Increased efficiency in identifying mixed pollen samples by meta-barcoding with a dual-indexing approach," *BMC Ecol.*, vol. 15, no. 1, Jul. 2015, doi: 10.1186/s12898-015-0051-y.

[79] Bell KL, Loeffler ViM, and Brosi BJ, " An rbcL Reference Library to Aid in the Identification of Plant Species Mixtures by DNA Metabarcoding ," *Appl. Plant Sci.*, vol. 5, no. 3, p. 1600110, 2017, doi: 10.3732/apps.1600110.

[80] Dubois B *et al.*, "A detailed workflow to develop QIIME2-formatted reference databases for taxonomic analysis of DNA metabarcoding data," *BMC Genomic Data*, vol. 23, no. 1, Dec. 2022, doi: 10.1186/s12863-022-01067-5.

[81] ullah S, Huyop F, Wahab RA, Sujana IGA, Antara NS, and Gunam IBW, "Using pollen DNA metabarcoding to trace the geographical and botanical origin of honey from Karangasem, Indonesia," *Heliyon*, vol. 10, no. 12, Jun. 2024, doi: 10.1016/j.heliyon.2024.e33094.

[82] Banchi E, Ametrano CG, Greco S, Stanković D, Muggia L, and Pallavicini A, "PLANiTS: A curated sequence reference dataset for plant ITS DNA metabarcoding," *Database*, vol. 2020, 2020, doi: 10.1093/database/baz155.

[83] Milla L, Sniderman K, Lines R, Mousavi-Derazmahalleh M, and Encinas-Viso F, "Pollen DNA metabarcoding identifies regional provenance and high plant diversity in Australian honey," *Ecol. Evol.*, vol. 11, no. 13, pp. 8683–8698, Jul. 2021, doi: 10.1002/ece3.7679.

[84] Núñez A *et al.*, "Validation of the Hirst-Type Spore Trap for Simultaneous Monitoring of Prokaryotic and Eukaryotic Biodiversities in Urban Air Samples by Next-Generation Sequencing," 2017. [Online]. Available: https://journals.asm.org/journal/aem

[85] Campbell BC *et al.*, "Science of the Total Environment Metabarcoding airborne pollen from subtropical and temperate eastern Australia over multiple years reveals pollen aerobiome diversity and complexity," *Sci. Total Environ.*, vol. 862, no. August 2022, p. 160585, 2023, doi: 10.1016/j.scitotenv.2022.160585.

[86] Brennan GL *et al.*, "Temperate airborne grass pollen defined by spatio-temporal shifts in community composition," *Nat. Ecol. Evol.*, vol. 3, pp. 750–754, Apr. 2019, doi: 10.1038/s41559-019-0849-7.

[87] de Weger LA, Bruffaerts N, and Koenders MM, "Long-Term Pollen Monitoring in the Benelux: Evaluation of Allergenic Pollen Levels and Temporal Variations of Pollen Seasons," vol. 2, no. July, pp. 1–12, 2021, doi: 10.3389/falgy.2021.676176.

[88] Wickham H, "ggplot2: Elegant Graphics for Data Analysis." Springer-Verlag New York, 2016. [Online]. Available: https://ggplot2.tidyverse.org

[89] Mardis E and McCombie WR, "Library Quantification: Fluorometric Quantitation of Double-Stranded or Single-Stranded DNA Samples Using the Qubit System," *Cold Spring Harb. Protoc.*, vol. 2017, no. 6, p. pdb.prot094730, Jun. 2017, doi: 10.1101/pdb.prot094730.

[90] Lowe A, Jones L, Brennan G, Creer S, and De Vere N, "Seasonal progression and differences in major floral resource use by bees and hoverflies in a diverse horticultural and agricultural landscape revealed by DNA metabarcoding," *J. Appl.*

Ecol., no. January, pp. 1484–1495, 2022, doi: 10.1111/1365-2664.14144.

[91] Waters DLE and Shapter FM, *Cereal Genomics: Methods and Protocols, Methods in Molecular Biology*, vol. 1099. New York: Springer Science+Business Media, 2014. doi: 10.1007/978-1-62703-715-0_7.

[92] Lucena-Aguilar G, Sánchez-López AM, Barberán-Aceituno C, Carrillo-Ávila J, López-Guerrero J, and Aguilar-Quesada R, "DNA Source Selection for Downstream Applications Based on DNA Quality Indicators Analysis," *Biopreserv. Biobank.*, vol. 14, no. 4, pp. 264–270, Aug. 2016, doi: 10.1089/bio.2015.0064.

[93] U'Ren JM and Arnold AE, "Illumina MiSeq Dual-barcoded Two-step PCR Amplicon Sequencing Protocol v1," Washington, USA, Apr. 2017. doi: 10.17504/protocols.io.fs9bnh6.

[94] Manhart JR, "Phylogenetic Analysis of Green Plant rbcL," *Mol. Phylogenet. Evol.*, vol. 3, no. 2, pp. 114–127, 1994.

[95] Bell KL, Burgess KS, Okamoto KC, Aranda R, and Brosi BJ, "Review and future prospects for DNA barcoding methods in forensic palynology," *Forensic Sci. Int. Genet.*, vol. 21, pp. 110–116, 2016, doi: 10.1016/j.fsigen.2015.12.010.

[96] Prosser SWJ and Hebert PDN, "Rapid identification of the botanical and entomological sources of honey using DNA metabarcoding," *Food Chem.*, vol. 214, pp. 183–191, 2017, doi: 10.1016/j.foodchem.2016.07.077.

[97] Zahra NB, Khan Shinwari Z, and Qaiser M, "DNA Barcoding: A tool for standardization of Herbal Medicinal Products (HMPS) of Lamiaceae from Pakistan," *Pak. J. Bot*, vol. 48, no. 5, pp. 2167–2174, 2016.

[98] White TJ, Bruns T, Lee S, and Taylor J, *PCR Protocols: A Guide to Methods and Applications*, no. January. 1990.

[99] Ralte L and Singh YT, "Use of rbcl and its2 for dna barcoding and identification of solanaceae plants in hilly state of Mizoram, India," *Res. Crop.*, vol. 22, no. 3, pp. 616–623, Sep. 2021, doi: 10.31830/2348-7542.2021.110.

[100] Richardson RT, Lin C, Sponsler DB, Quijia JO, Goodell K, and Johnson RM, "Application of ITS2 metabarcoding to determine the provenance of pollen collected by honey bees in an agroecosystem," *Appl. Plant Sci.*, vol. 3, no. 1, pp. 1–6, 2015, doi: 10.3732/apps.1400066.

[101] Yao H *et al.*, "Use of ITS2 region as the universal DNA barcode for plants and animals," *PLoS One*, vol. 5, no. 10, Oct. 2010, doi: 10.1371/journal.pone.0013102.

[102] Kircher M, Sawyer S, and Meyer M, "Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform," *Nucleic Acids Res.*, vol. 40, no. 1, Jan. 2012, doi: 10.1093/nar/gkr771.

[103] Ravi RK, Walton K, and Khosroheidari M, *Disease Gene Identification: Methods and Protocols, Methods in Molecular Biology*, vol. 1706. in Methods in Molecular Biology, vol. 1706. New York, NY: Springer Science+Business Media, 2018. doi: 10.1007/978-1-4939-7471-9.

[104] Nelson A and Stewart CJ, *Innate Lymphoid Cells: Methods and Protocols, Methods in Molecular Biology*, vol. 2121. Springer Science+Business Media, 2020. [Online].

Available: http://www.springer.com/series/7651

[105] Smith AD and de Sena Brandine G, “Falco: High-speed FastQC emulation for quality control of sequencing data,” *F1000Research*, vol. 8, no. 1874, Jan. 2021, doi: 10.12688/f1000research.21142.2.

[106] Ewels P, Magnusson M, Lundin S, and Käller M, “MultiQC: Summarize analysis results for multiple tools and samples in a single report,” *Bioinformatics*, vol. 32, no. 19, pp. 3047–3048, Oct. 2016, doi: 10.1093/bioinformatics/btw354.

[107] Bolger AM, Lohse M, and Usadel B, “Trimmomatic: A flexible trimmer for Illumina sequence data,” *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, Aug. 2014, doi: 10.1093/bioinformatics/btu170.

[108] Edgar RC, “Search and clustering orders of magnitude faster than BLAST,” *Bioinformatics*, vol. 26, no. 19, pp. 2460–2461, Aug. 2010, doi: 10.1093/bioinformatics/btq461.

[109] Keller A *et al.*, “Evaluating multiplexed next-generation sequencing as a method in palynology for mixed pollen samples,” *Plant Biol.*, vol. 17, no. 2, pp. 558–566, Mar. 2015, doi: 10.1111/plb.12251.

[110] Barman P, Bharali R, and Dey S, “BLAST: An introductory tool for students to Bioinformatics Applications,” *Keanean J. Sci.*, pp. 67–76, 2013, [Online]. Available: https://www.researchgate.net/publication/267332265

[111] Trytsman M, Müller FL, and van Wyk AE, “Diversity of grasses (Poaceae) in southern Africa, with emphasis on the conservation of pasture genetic resources,” *Genet. Resour. Crop Evol.*, vol. 67, no. 4, pp. 875–894, Apr. 2020, doi: 10.1007/s10722-020-00886-8.

[112] Jones L, Brennan GL, Lowe A, Creer S, Ford CR, and de Vere N, “Shifts in honeybee foraging reveal historical changes in floral resources,” *Commun. Biol.*, vol. 4, no. 1, Dec. 2021, doi: 10.1038/s42003-020-01562-4.

[113] Li H, “Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM,” May 2013, [Online]. Available: http://arxiv.org/abs/1303.3997

[114] Sokołowska J, Fuchs H, and Celiński K, “Assessment of ITS2 Region Relevance for Taxa Discrimination and Phylogenetic Inference among Pinaceae,” *Plants*, vol. 11, no. 8, Apr. 2022, doi: 10.3390/plants11081078.

[115] G. D *et al.*, “Ambrosia (ragweed) pollen — A growing aeroallergen of concern in South Africa,” *World Allergy Organ. J.*, vol. 17, no. 12, p. 101011, 2024, doi: 10.1016/j.waojou.2024.101011.

[116] Oliveros JC, “Venny: An interactive tool for comparing lists with Venn’s diagrams,” 2015, 1.0. [Online]. Available: ttps://bioinfogp.cnb.csic.es/tools/venny/index.html