From Protocol to Practice: Airborne Pathogen Surveillance in Agricultural Settings

Mia Fay Gee Berelson

A thesis presented for the degree of Doctor of Philosophy

University of East Anglia
School of Biological Sciences
&
Earlham Institute

September 2025

Abstract

Crop pathogens present a persistent threat to yields and global food security, and the current heavy reliance on fungicides is unsustainable. Improved monitoring strategies are therefore essential. This thesis describes experiments aimed to test, refine, and validate AirSeq, a sequencing-based approach for the detection of airborne pathogens. AirSeq was first applied in a commercial greenhouse to monitor strawberry pathogens, where detections correlated with manual disease scores, demonstrating its potential as a surveillance tool. Field and laboratory experiments were then conducted to refine the protocol, identifying the most effective sampler and extraction methods, and highlighting the importance of experimental controls. Using the optimised method, seasonal monitoring in wheat fields revealed community-level dynamics of fungi and oomycetes, with pathogen detections often coinciding with favourable environmental conditions and, in some cases, preceding visible symptoms. Short-term sampling further showed that airborne microbial communities fluctuate substantially over diurnal and seasonal timescales. To enhance interpretation of these datasets, a custom bioinformatic pipeline, MARMoT (Metagenomic Alignment and Reporting for Monitoring of Threats), was developed, enabling detection of high-risk pathogens from airborne samples. While some detections, were robust others may have represented false positives, emphasising the need for caution in species-level assignments.

Overall, the work in this thesis demonstrates that AirSeq can capture airborne microbial diversity and track pathogen dynamics across multiple temporal and spatial scales. The method shows clear promise as an early-warning system for crop pathogens and could be integrated into future disease surveillance frameworks. However, further validation and refinements are required before outputs can be routinely translated into actionable crop protection strategies.

Contents

Acronyms

AMR antimicrobial resistance.

ASVs amplicon sequence variants.

BOP Breeder Observation Panel.

CPM Cycles per minute.

ddNTPs dideoxynucleotides.

DEFRA Department for Environment, Food and Rural Affairs.

eDNA environmental DNA.

ELISA enzyme-linked immunosorbent assay.

FRAC Fungicide Resistance Action Committee.

G1 Greenhouse 1.

G2 Greenhouse 2.

HAC high accuracy.

HPC High-performance Computer.

HPM Hits per Million.

HYSPLIT Hybrid Single Particle Lagrangian Integrated Trajectory.

L/min Litres per Minute.

LAMP loop-mediated isothermal amplification.

LCA lowest common ancestor.

MAGs metagenome-assembled genomes.

MARMoT Metagenomic Alignment and Reporting for Monitoring of Threats.

MQ mapping quality.

NGS next-generation sequencing.

NHM Natural History Museum.

ONT Oxford Nanopore Technologies.

OTUs operational taxonomic units.

PacBio Pacific Biosciences.

PAF pairwise alignment format.

PCoA Principal Coordinate Analysis.

PCR polymerase chain reaction.

PhCR Phytophthora crown rot.

PHI-base Pathogen–Host Interaction database.

PhLR Phytophthora leather rot.

qcov query coverage.

qPCR quantitative PCR.

RH relative humidity.

rRNA Ribosomal RNA.

SNP single nucleotide polymorphism.

SUP super accuracy.

t-SNE t-distributed Stochastic Neighbor Embedding.

WGA whole-genome amplification.

WGS whole-genome sequencing.

List of Figures

List of Tables

Acknowledgements

Firstly, I would like to thank my supervisors Richard Leggett, Matt Clark, and Paul Nicholson for their guidance throughout this PhD and for ensuring the work added value to the scientific literature. I am also grateful to the Biotechnology and Biological Sciences Research Council and the UKRI-BBSRC Norwich Research Park Biosciences Doctoral Training Partnership for their support.

I owe thanks to all members of the Leggett group, past and present, for their help with ideas, analysis, writing, and many rounds of sample collection. In particular, I would like to thank Darren for his guidance in both the field and lab; Sam for teaching me how to use the command line and generously sharing his scripts; Ned for patiently helping me debug MARTi and Jade for her constant enthusiasm and continuing the AirSeq work.

I would also like to thank Darryl Playford and the team at JIC's Experimental Field Station, as well as Andrey Ivanov and colleagues at Wilkin & Sons, for providing access to sampling sites and supplying disease score data. I am also grateful to Grant Bexson for his help with the EI labs.

My gratitude goes to all of the PhD students at EI for creating such a supportive community, with a special appreciation for everyone who became involved with the Earlham Student Body. A big thank you to Becky, Insect George, Jess, Kate, and Sofia for the many Centrum lunches, cow walks, and pub quizzes that made this journey far more enjoyable.

Also to Dan, Lu and Phoebe, for being such amazing housemates with endless cups of tea, soft sits and long chats. A massive thank you to Lorcán for being incredibly supportive and patient. I am also thankful to my parents, siblings and friends who do not live nearby but were still able to support me by always being on the end of the phone when I needed it most.

I of course have to acknowledge Nick, my furry companion, whose gentle morning wake-ups, comforting purrs and occasional walks across my keyboard have been much appreciated.

Finally, this thesis is dedicated to the memory of my wonderful friends, Hannah Emily Bainbridge and Aaron Kyle Williams, I wish they could be here to celebrate this achievement with me.

Chapter 1

Introduction

1.1 Problem Statement

1.1.1 Food security in a changing world

Crop pathogens cause significant global economic losses, estimated at $220 billion per year by the UN Food and Agriculture Organization (FAO) [371]. These losses are expected to rise due to globalisation and climate change, which facilitate the spread of pathogens into new regions and create favourable conditions for their growth [85, 265]. In addition to their economic impact, plant diseases threaten global food security by reducing yields at a time when demand for food continues to increase.

Plant pathogens are commonly managed through chemical control, and in England and Wales fungicide applications have increased over the past two decades [369]. Heavy reliance on fungicides has accelerated resistance in many pathogens [82, 283, 360, 405], while excessive application poses environmental and health risks [367]. These concerns have led to the withdrawal of certain active ingredients and growing consumer pressure to reduce chemical inputs, ultimately constraining growers' options for effective disease management.

Alongside chemical control, cultural practices such as the use of sterilised seed, resistant varieties, debris reduction, and crop rotation contribute to disease management [342, 376, 383]. Adoption has increased in recent years, particularly through reductions in susceptible wheat varieties [369]. However, the development of new resistant cultivars remains slow, and cultural measures alone cannot eradicate disease.

Despite the range of chemical and cultural strategies available, effective management is often hindered by limitations in pathogen surveillance. Early and accurate detection could provide growers with reliable information on pathogen presence, enabling targeted interventions, more efficient fungicide use, and proactive disease control to safeguard yields.

1.1.2 Current spore detection methods are limited

Accurate and timely disease detection is critical in crop production, yet data on pathogen occurrence and distribution is often fragmented, outdated, or insufficiently standardised [69]. This makes it difficult to assess impacts or allocate resources effectively, underscoring the need for rapid, scalable airborne pathogen detection systems.

Current monitoring still relies heavily on microscopy-based assessments of airborne spores. While these methods can reveal broad trends, they are labour-intensive, require

specialist expertise, and lack the sensitivity of molecular approaches [203]. More recently, molecular assays such as polymerase chain reaction (PCR), loop-mediated isothermal amplification (LAMP), and enzyme-linked immunosorbent assay (ELISA) have shown promise in proof-of-concept studies, but adoption in commercial agriculture remains limited. Similarly, imaging, sensor networks, drones, and deep learning models have been trialled to detect plant disease symptoms in the field [41, 87, 337, 412], yet these approaches focus on visible infections and provide little insight into the airborne spores that initiate outbreaks.

Together, these limitations highlight the need for a more timely, comprehensive, and integrated approach to plant disease surveillance. Airborne eDNA sequencing offers such a possibility, detecting fungal pathogens directly from the air before symptoms appear on crops and providing growers with a proactive tool for disease management.

1.2 AirSeq Technology

AirSeq is a methodology that integrates sample capture and molecular analysis to enable the detection of airborne plant pathogens. Airborne particles are collected using a high-volume sampler, concentrated by filtration, and subjected to DNA extraction and sequencing to identify the taxa present. This provides a non-invasive, rapid, and accurate approach to pathogen surveillance, while the use of whole-genome sequencing (WGS) allows the entire airborne microbiome to be characterised, enabling simultaneous detection of multiple pathogens.

A preliminary version of this method was applied by Giolai et al. using the Coriolis µ sampler and Illumina sequencing [124]. In contrast, all data presented in this thesis were generated using Oxford Nanopore Technologies (ONT) sequencing. As both sampling and sequencing technologies continue to advance, the work presented here seeks to refine the original AirSeq pipeline and further evaluate its efficacy for airborne pathogen detection.

1.3 Thesis Overview and Research Objectives

This section provides a brief overview of the research undertaken in this thesis. The objectives of this PhD were to optimise and validate the AirSeq pipeline, encompassing airborne sample capture, wet-lab processes, and bioinformatic analysis, with the goal of enabling reliable detection of fungal plant pathogens that threaten crops. A secondary objective was to investigate the presence of potentially emergent airborne pathogens not yet established in UK crops but detectable in the air.

Chapter 2 is a literature review introducing the field of airborne eDNA, outlining methodologies, identifying potential sources of bias, and reviewing current applications and research gaps before considering future directions. Relevant plant pathogens are introduced in the following research chapters.

Chapter 3 presents a year-long proof-of-concept study conducted at a commercial strawberry production site, where monthly AirSeq sampling was used to compare airborne spore abundance with environmental conditions and visual disease assessments.

Chapter 4 details a series of experiments used to refine the protocol, covering both sample collection and wet-lab processes. These include comparisons of air samplers and storage methods, a distance-from-source trial, an evaluation of different cell lysis approaches, and inclusion of negative and positive controls.

Chapter 5 presents results from regular AirSeq collections in an untreated wheat plot at a research farm over two growing seasons. Alongside broad taxonomic comparisons, it examines fluctuations in the abundance of nine target pathogen genera in greater detail.

Chapter 6 presents back-to-back collections from the same wheat field over three different 24-hour periods, demonstrating the highly dynamic nature of airborne eDNA.

Chapter 7, the final research chapter, outlines the design of the MaRMOT (Metagenomic Alignment and Reporting for Monitoring of Threats) pipeline, presenting parameter optimisation tests and results from all the datasets in this thesis, as well as additional AirSeq samples processed with the pipeline to identify potential emergent pathogens.

Finally, Chapter 8 brings together the research as a whole, discussing limitations, challenges, and opportunities for future development, both to strengthen AirSeq and to expand its role in crop protection strategies.

This work establishes AirSeq as a powerful tool for the early detection of airborne plant pathogens, validated through disease scoring and pathogen biology. At the same time, it reveals the complexity of the airborne microbiome and the diverse factors influencing its composition, underscoring both the promise and challenges of this emerging field.

Chapter 2

Literature Review

2.1 Abstract

Airborne eDNA refers to DNA present in the air, derived from a range of bacteria, fungi, viruses, and vertebrates. Historically, microscopy and culture-based techniques were used to analyse the taxa present, but advances in molecular biology, such as PCR and next-generation sequencing (NGS), now enable researchers to generate vast amounts of data from airborne eDNA. Typical workflows begin with the capture of airborne material, followed by nucleic acid extraction, library preparation, sequencing, and ultimately taxonomic identification to determine the species present. Airborne eDNA analysis has been applied across diverse fields, ranging from pathogen detection in agriculture to human health, air quality monitoring, bioterrorism detection, and biodiversity assessment. Because of this diversity of applications, a wide variety of methodological approaches are employed, each tailored to the specific requirements of an experiment. This review summarises current methods in the literature, highlights how methodological choices influence the identified taxa, and explores both applications and ongoing challenges in the field. The review expands upon material published in [36], with revisions and additional content adapted for the scope of this thesis.

2.2 Introduction

Airborne eDNA originates from diverse sources, encompassing bacteria, fungi, and viruses suspended in the air, as well as vertebrates that shed skin or other materials [80, 228], and plants that release pollen or spores [363]. This airborne eDNA plays important roles globally, both via the transmission of human and plant diseases [124, 144, 210, 213] and influencing weather events such as ice-nucleating particles that affect cloud formation [163].

Temporal variation in airborne eDNA occurs across multiple scales, from hourly fluctuations [71] to seasonal patterns [29, 274]. Given current knowledge of airborne eDNA and its sensitivity to land use [221], variation is also expected over both very short intervals (seconds to minutes) and longer timescales spanning decades to centuries. In addition to temporal dynamics, airborne eDNA exhibits spatial heterogeneity, with distinct communities detected at different altitudes within the stratosphere [220, 233] and across horizontal scales ranging from cities to continents [274, 289]. These spatial patterns are often linked to local land use, with species composition reflecting surrounding environments [336].

Early studies in this field relied on passive traps and the identification of pollen and

fungal taxa by microscopy. These approaches provided valuable insights but were labour intensive, required specialist expertise, and captured a fraction of the diversity present. Advances in air sampling technology have since produced devices capable of processing thousands of litres of air per minute (e.g. SASS 4100), enabling the collection of sufficient material for molecular analysis. At the same time, improvements in DNA extraction, amplification, low-input library preparation, and the high sensitivity of NGS have greatly enhanced detection. Together, these developments make it possible to generate sequence data from collections lasting as little as one minute [3]. As a result, processing has become faster and more sensitive, allowing for the identification of a broader range of taxa [17, 244]. This technological shift has supported the expansion of airborne eDNA research across diverse fields, including real-time pollen monitoring [60, 71], tracking forest disease outbreaks [7], biodiversity assessment [80, 228], surveillance of SARS-CoV-2 [305], and the detection of plant pathogens [124].

This review covers the process of airborne eDNA analysis, from sampling through to DNA extraction, sequencing, and bioinformatic interpretation. It also discusses current applications, knowledge gaps, and potential future developments in the field (Figure 2.1), with particular attention to aspects most relevant to this thesis, including Oxford Nanopore Technologies (ONT) and alternative methods for plant pathogen detection.

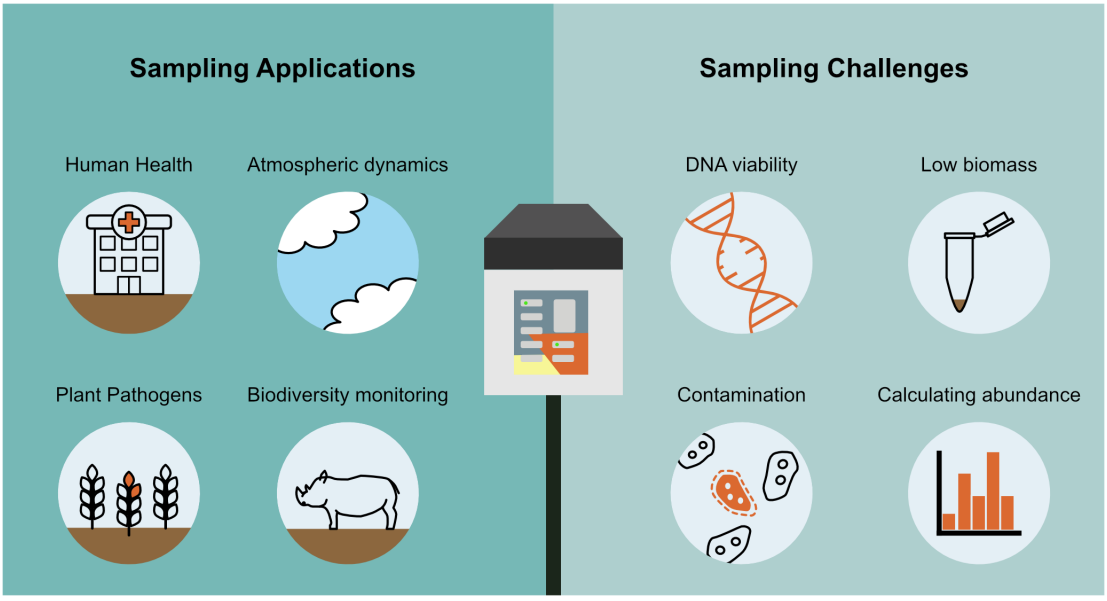


Figure 2.1: Representation of the applications and challenges of air sampling for eDNA analysis, from [36].

2.3 Stages in Airborne eDNA Analysis

Ordinarily, airborne eDNA analysis proceeds through four stages: sample collection, DNA extraction and isolation, amplification and sequencing, and bioinformatic processing. This general pipeline is illustrated in Figure 2.2, which highlights how the specific approaches at each stage vary between experiments. Methodological choices depend on factors such as time, budget, target taxa, and equipment availability. For instance, a study in a low-biomass environment may require a week-long collection to obtain sufficient DNA, whereas investigations of temporal dynamics in an urban setting might employ multiple 60-second

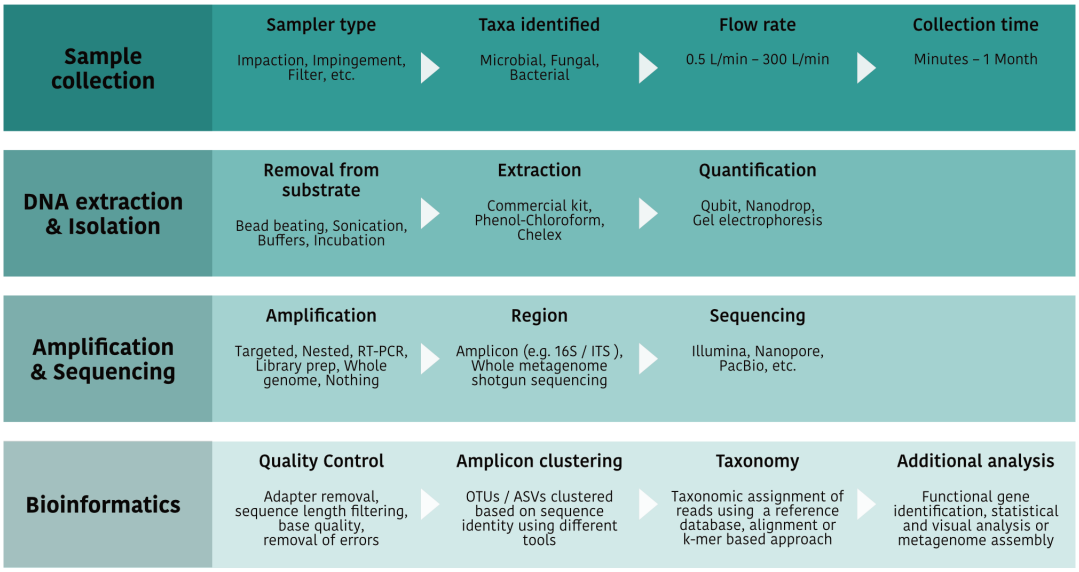


Figure 2.2: The four main stages of airborne eDNA analysis: sample collection, DNA extraction and isolation, amplification and sequencing, and bioinformatics. Each stage includes several steps with options that vary according to the research objectives. From [36].

snapshot samples across a single day.

The complexity of metagenomic data means that methodological choices inevitably introduce biases, as different techniques vary in their ability to detect particular taxa. These biases must therefore be considered when comparing results across airborne eDNA studies.

2.3.1 Air sample collection

Air sample collection is the first step in airborne eDNA analysis, requiring sufficient material for sequencing and to obtain a representative metagenomic profile. Outcomes are shaped by controllable factors such as sampler type, collection duration, and sampler height, as well as by sampler-specific constraints including flow rate, particle size capture, and efficiency.

2.3.1.1 Air sampler type

Air samplers operate through several mechanisms, including impaction, impingement, and filtration [386]. They can be either passive, without active airflow, or active, where air is drawn through the device. Each design has advantages and disadvantages (Table 2.1), with detection influenced by factors such as efficiency, flow rate, and collection surface. Consequently, sampler choice should be guided by the research question. For instance, active samplers are often preferred for DNA sequencing because they capture larger volumes of material. Whilst the use of a selective agarose medium as a collection surface can identify viable fungal spores, as in a study on the relationship between *Exobasidium* spore concentration and disease where colonies were counted after incubation [167], although this approach is limited to culturable taxa.

The focus of this review is on approaches that integrate DNA sequencing with air sampling for the identification of airborne taxa, rather than on detailed evaluations of

Table 2.1: Types of air samplers, from [36], originally adapted from [106, 386].

Category	Samplers available	Description	Advantages	Disadvantages
Impaction	Burkard 7 Day sampler, Andersen sampler, MicroPEM, AirPort MD8 364 device	Particles are impacted onto a surface (adhesive, agar, membrane filter etc.) transverse to airflow	Can select which particle size is recovered by adjusting the flow rate Direct collection onto agar for viability testing	Particle bounce can reduce collection efficiency
Impingement or cyclone	Bertin Coriolis, SKC Biosampler, VIVAS	Particles become suspended in a cyclone of air/liquid	Higher flow rates than filter samplers	Lower collection efficiency compared to filter samplers
Filtering	Thermo Scientific MFC-PM10 High Volume Air Sampler, Innovaprep Bobcat/Cub, SASS 3100/4100 Dry Air Sampler	Air is passed through a filter (fibrous, membrane or flat), which may include electrostatically charged filters ('electret')	Easy-to-use filter and recovery system Electret filters can collect particles smaller than the pore size	The choice of filter size determines collection efficiency for various particle sizes Not all particles captured (could also be an advantage)

individual sampler models. Comparative assessments of sampler performance have been provided elsewhere [106, 386], alongside studies dedicated to specific mechanisms such as passive samplers [237] and rotating arm impaction devices [74]. The following section instead outlines general observations on the attributes of the three principal sampler types.

Air sampler comparison An ideal air sampler, capable of capturing all airborne particles, representing the full diversity of taxa present, and detecting rare organisms, has not yet been developed. As a result, existing devices are typically evaluated in terms of particle collection efficiency, the diversity of taxa recovered, and their sensitivity. Collection efficiency can be assessed through measures such as total DNA concentration or, for bacterial communities, by quantifying 16S rRNA gene copy number. The extent to which a sampler reflects overall community composition, including rarer taxa, is determined during the sequencing and analysis stages.

A comparison of liquid impingement and filter impaction samplers used the 16S rRNA gene copy abundance for analysis. Filter impaction recovered more than an order of magnitude greater copy numbers than impingement, reflecting the higher particle retention efficiency of filter-based approaches [106]. Collection surfaces may also influence yield and subsequent molecular analyses. In a separate study, DNA recovery was evaluated from culture-based and culture-independent collections using membrane filtration, liquid impingement, and an electrostatic collector. The electorstatic collector achieved the highest yields, likely due to its substantially greater flow rate (10–100 times higher than the other devices tested) [250].

When comparing samplers with the same mechanism, the choice of capture material was shown to influence collection efficiency. For example, a comparison of five membrane filters demonstrated substantial differences in DNA recovery, with sequencing revealing corresponding variation in the diversity of recovered taxa [177]. Similarly, a study comparing microbial communities obtained from five air samplers and from settled dust found that

sampler type exerted a stronger effect on community composition than collection location [155]. Such variation is likely attributable to differences in flow rate, particle size cut-off, sampling height, and collection medium.

Studies examining sampler sensitivity are relatively limited, largely because such assessments require controlled conditions such as clean rooms. One comparison of two impingement samplers reported that the VIVAS exhibited higher sensitivity, detecting viruses not recovered by the BioSampler [277]. In a broader study of eight devices, high-flow rate samplers captured the greatest number of viruses, whereas lower-flow rate samplers, including the VIVAS, more accurately reflected airborne viral concentrations [304].

Research on bacterial pathogen detection has also demonstrated differences among sampler types. A comparison of impaction (AirPort MD8), impingement (BioSampler), and cyclone (Coriolis µ) devices showed that all were capable of detecting *Coxiella burnetii*, although performance varied across concentration levels [2]. Similarly, in the context of airborne vertebrate DNA, the Burkard spore trap recovered a greater number of vertebrate species than cyclone samplers (Coriolis µ and Burkard multi-vial cyclone) [290].

2.3.1.2 Influence of sampling volume and duration

The volume of air collected is determined by the flow rate and length of sampling, and directly affects downstream DNA yield. While sufficient DNA is required for sequencing and analysis, optimal collection duration depends on whether the aim is to capture the full community present or to monitor temporal changes in airborne eDNA. Increasing the volume of air sampled, either through higher flow rates or longer collection periods, generally enhances DNA recovery, [106, 226] and likely the diversity of identified taxa although there is limited published data on this relationship.

Comprehensive characterisation of airborne taxa often requires extended sampling in low-biomass environments. For example, in polar regions, 7-day samples produced 16S rRNA gene concentrations comparable to negative controls, indicating that substantially longer collections would be necessary to recover sufficient DNA for sequencing [92].

Species accumulation curves provide a useful tool to assess whether the majority of taxa at a site have been captured. In one study, species richness continued to increase over sampling intervals ranging from 1 to 403 minutes, with no plateau observed [3]. Accordingly, studies aiming to describe the full airborne microbiome frequently employ collection periods of up to seven days [7].

By contrast, investigations into temporal or climatic dynamics require shorter "snapshot" collections. High-flow rate samplers facilitate this approach by recovering sufficient DNA within short timeframes. A recently developed personal impingement sampler, for instance, was able to detect microbes from 10-second samples using quantitative PCR (qPCR) [6], although this technique requires less DNA input than NGS.

2.3.1.3 Sampler placement and environmental factors

The positioning of air samplers is a critical consideration, as variables such as height, wind direction, and surrounding habitat strongly influence the taxa detected. Although there is no universally optimal placement, decisions are typically guided by local site characteristics, meteorological conditions, and the biological system under investigation [231].

In indoor environments, placement depends on study objectives. Samplers may be located in bedrooms to monitor overnight allergens, positioned near infected patients to assess airborne pathogen load, or placed in hospital corridors to investigate microbial transmission of clinical concern [42, 197, 311]. Comparisons of indoor samplers at different heights have shown no significant differences in community composition or diversity, likely reflecting the relatively contained and homogeneous nature of indoor air [311].

For outdoor studies, positioning is more dependent on the ecological question. Samplers placed close to the soil surface are more effective for detecting spores of soil-borne pathogens [231], whereas those above the floral canopy are more likely to capture airborne inoculum arriving from external sources [386]. A comparison of rooftop and canopy-level samplers found fungal species richness to be greater at rooftop level, reflecting the dominance of local crop-derived spores closer to the canopy and a more diverse mixture of particles higher in the air column [386].

Topography further influences airborne eDNA dispersal, as wind speed and direction are shaped by valleys, mountains, and large built structures. In particular, obstacles can generate sheltered "quiet zones" where airflow is reduced, extending to heights up to twenty times that of the structure [231].

2.3.1.4 Timing of sample collection

The timing of air sampling is a key determinant of airborne eDNA composition, as both seasonal and diurnal variation strongly influence the taxa detected. Seasonal patterns are well documented, with pollen from flowering plants dominating in spring, grasses and wood-rotting fungi becoming more prevalent in summer, fungal spores peaking in autumn, and winter characterised by reduced pollen alongside cold-tolerant bacteria [28, 138, 355]. Accurate interpretation and comparison of results therefore require precise records of when collections were undertaken.

Diurnal variation further shapes airborne communities, driven by environmental factors such as temperature, humidity, and ultraviolet radiation. These drivers produce consistent daily patterns, including midday peaks of powdery mildew, nocturnal sporulation in many tropical fungi, and afternoon maxima of *Alternaria spp.* [125, 206]. Other shifts can occur more sporadically in response to weather events: rainfall both washes out existing particles and disperses new material via splash effects [172, 264]; wind speed and direction has been linked to increased microbial diversity [113, 370]; and humidity commonly triggers fungal spore release [20].

Evidence from urban and natural settings further highlights taxon-specific differences. In one city-based study, human pathogenic bacteria were more abundant during the day than at night [160]. While in Siberia fungal communities peaked in abundance and diversity at night, decreasing during daylight hours, with bacterial communities showing the reverse trend of evening maxima [138].

These dynamics underline the importance of recording detailed metadata, including time of collection and local meteorological conditions. Incorporating such contextual information into analyses enhances interpretation and provides deeper insight into the environmental drivers shaping airborne eDNA communities.

2.3.1.5 Effect of particle size on dispersal and capture

Particle size is a major determinant of airborne eDNA dispersal, affecting both atmospheric residence time and sampling efficiency. Although biological particles can occur at the nanometre scale, those commonly detected in air range from approximately 0.65 to 12 μm [81]. Smaller particles remain suspended for longer periods, while larger ones settle more rapidly [368]. Efficient capture of larger particles with short airborne residence may therefore require placement closer to the source, extended sampling durations, or targeting periods when release is most likely.

Sampler design also plays an important role, as devices are typically optimised for particular particle size ranges. Selection of an appropriate sampler is thus critical for accurately characterising community composition. Although literature on airborne eDNA particle size remains limited, reviews on fungal spore detection provide guidance for optimising collection strategies [231], and several studies have compared community composition across particle size fractions [12, 107, 171, 264, 399].

2.3.2 DNA extraction and isolation

Following air collection, the next stages involve concentrating the sample, lysing cells and spores, and extracting DNA. These steps are critical for downstream sequencing and taxonomic analysis, but each can introduce bias into the recovered community. Most studies rely on commercially available extraction kits, often with modifications such as extended incubation or additional filtering steps. There are few direct comparisons of extraction protocols for air samples, and those that exist are usually limited to a particular sampler or collection medium.

Obtaining sufficient, high-quality DNA is essential, as low yields can compromise sequencing depth and taxonomic assignment, while degraded DNA may result in misidentifications or reduced accuracy. Consequently, the choice of extraction protocol must balance maximising yield with preserving DNA integrity. Prior to extraction, samples usually require concentration by either filtration or centrifugation. A comparison of these approaches showed that filtration yielded three orders of magnitude more DNA than centrifugation and recovered nine fungal taxa absent from centrifuged samples when assessed using the ITS gene [244]. This suggests filtration provides a more comprehensive view of fungal communities in liquid samples.

Once concentrated, DNA is typically extracted using kits originally designed for other sample types (e.g. soil, water, or tissue), as none are currently optimised for air. A study comparing four commercial kits on quartz-filter samples reported that soil and water kits produced higher 16S rRNA gene copy numbers than blood and tissue kits, although taxonomic comparisons of diversity were not performed [92]. An additional complication is contamination from the so-called "kitome", background DNA present in commercial kits that may be mistaken for genuine sequences [319].

Bias may also arise from species-specific differences in lysis efficiency or DNA stability. Cells that lyse more readily or adhere less to tube surfaces will be overrepresented in extracts, whereas sporulating fungi, for example, are often underrepresented [27, 56]. Similarly, DNA extraction using the Zymo Fungal/Bacteria DNA Microprep Kit from a mock community produced both over- and under-represented species relative to their actual abundance [27]. A comparison of three extraction protocols with differing lysis and

incubation steps found that only half of the top ten most abundant species were shared across methods, highlighting the strong influence of protocol choice on community composition. Furthermore, separate analysis of the pellet and supernatant after centrifugation revealed substantial taxonomic differences between fractions [47].

These studies collectively show that extraction protocols introduce systematic biases, leading to discrepancies between observed and true airborne communities. While some degree of bias is unavoidable, it can be reduced by selecting appropriate protocols and, importantly, by applying consistent procedures throughout a project to ensure comparability. More comprehensive methods, such as the 'three peaks' protocol developed for faecal samples, which combines chemical, enzymatic, and mechanical lysis while retaining longer DNA fragments [299], may hold promise for future application to air samples, although their use remains uncommon.

2.3.3 Amplification & sequencing

Following DNA extraction, airborne eDNA can be analysed using either metabarcoding or WGS. Metabarcoding employs PCR-based amplification of marker genes, such as the 16S rRNA or ITS regions, and is commonly used to characterise microbial community composition. WGS, with or without prior WGA, is less frequently applied but provides broader insights, including the detection of functional genes. The choice between metabarcoding and WGS depends on factors such as desired taxonomic resolution, sequencing costs, and the specific research question. For example, studies focusing on bacterial community composition typically employ 16S amplicon sequencing, whereas investigations into antimicrobial resistance genes are better suited to WGS approaches.

2.3.3.1 Metabarcoding

Metabarcoding targets conserved genomic regions that contain variable sites, enabling taxa to be distinguished using a shared set of primers. For bacteria and fungi, this is most commonly the 16S rRNA and ITS regions, and amplicon sequencing based on these markers has been widely applied to airborne eDNA studies [7, 106, 221, 311]. The approach is frequently selected because it is relatively inexpensive and, through PCR amplification, can generate usable data from low DNA inputs. However, amplification introduces primer-driven biases and the resulting data typically offer lower taxonomic resolution than WGS [192, 226].

Primer choice strongly influences community profiles. For example, comparisons of ITS1 and ITS2 regions for fungal metabarcoding revealed distinct community compositions, with certain taxa only detectable using one marker: *Armillaria spp.* from ITS1, and *Hymenoscyphus fraxineus* and *Melampsora larici-populina* from ITS2 [76, 226]. Furthermore, a comparison of five different ITS primers with the same sample detected different fungal communities with each set, with each primer pair able to detect ∼50% of the fungal diversity, increasing to 70–80%, when results from two different primer sets were combined [327]. Similarly, bacterial studies contrasting 16S and 12S amplicons identified different sets of unique taxa [228]. Consequently, reliance on a single marker region limits the characterisation of airborne communities, as some species remain undetected.

Another limitation is the frequent inability to resolve sequences to species level. Amplicon reads are often assigned only to higher ranks such as genus [26, 228, 311]. This

presents challenges for applications where species-level identification is critical, such as the monitoring of fungal plant pathogens. Closely related taxa within the same genus can differ in host range or pathogenicity, meaning genus-level identification may be insufficient to inform management decisions such as fungicide application.

Long-read metabarcoding, using platforms such as ONT and Pacific Biosciences (PacBio), enables full-length marker genes to be sequenced, improving taxonomic resolution compared with short-read approaches [225, 325]. These platforms are outlined further in section 2.3.3.4. Falling PacBio costs and decreasing ONT error rates are making such methods increasingly practical [89]. Yet, long-read metabarcoding requires distinct bioinformatic pipelines, and applications to airborne eDNA remain rare. Despite this, the approach holds considerable potential for improving species-level detection in future studies.

2.3.3.2 Bias in amplicon sequencing

Amplification bias is a well-recognised issue in metagenomic studies, leading to the over- or under-representation of taxa in mixed samples following PCR [56, 340]. Several factors contribute to this effect, including primer choice, annealing temperature, and the number of amplification cycles [112, 193, 280, 282, 340]. Increasing the cycle number has been shown to markedly reduce apparent community richness, sometimes by four- to ten-fold, as taxa with efficient primer binding sites become disproportionately amplified while others are lost [193, 282].

Copy number variation also introduces distortion, as organisms differ in the number of target gene copies. In bacteria, for instance, the 16S rRNA gene can range from one to fifteen copies per genome [380]. Species with higher copy numbers will therefore appear artificially abundant in amplicon datasets.

Even single-base mismatches between primers and template DNA can affect amplification efficiency. Such mismatches can lead to preferential recovery of some taxa, altering relative abundance estimates, and may also impact taxonomic resolution by producing variable amplicon lengths that reduce classification accuracy [280].

Most work on these issues has been carried out in microbial mock communities or within gut and stool microbiome studies, but the same principles are likely applicable to airborne eDNA. Given these limitations, it is improbable that any single PCR protocol can amplify all taxa with equal efficiency. These challenges emphasise the need for careful primer selection, optimisation of reaction conditions, and awareness of potential biases when interpreting metabarcoding results.

2.3.3.3 Whole genome sequencing

WGS enables the simultaneous detection of taxa across all kingdoms and can provide both species-level identification and insights into functional gene content. In principle, it offers a more comprehensive profile of airborne eDNA than metabarcoding, capturing a broader range of organisms and functional information [47, 124, 179, 197, 215, 298].

A major advantage of WGS is that it typically requires little or no PCR amplification, thereby reducing biases associated with primer choice and amplification cycles. Consequently, relative abundance estimates from WGS datasets are generally closer to the true community composition than those obtained from amplicon sequencing. However, the approach requires a larger amount of starting DNA, which is often a limitation when working

with air samples characterised by low biomass. One strategy to address this challenge is to reduce reaction volumes during library preparation, as shown in a ONT sequencing study that successfully generated data from only 6.25% of the recommended DNA input [148].

Despite these advantages, WGS remains constrained by less comprehensive reference databases compared with metabarcoding markers. For instance, the UNITE database (v10) contains ITS sequences representing approximately eleven times more fungal species than are currently available as reference genomes in NCBI RefSeq (release 228). This lack of representation limits the accuracy of taxonomic classification and remains a key challenge for the broader application of WGS in airborne eDNA research.

2.3.3.4 Sequencing platforms

Further to the choice between amplicon or WGS is the choice of sequencing platform. The platform influences the resolution, accuracy, and practicality of airborne eDNA studies, with technologies varying in read length, throughput, cost, and portability.

First developed in 1977 [321], Sanger sequencing became the earliest DNA sequencing method to be widely adopted and was later applied in airborne eDNA studies, including those published in 2009 [51] and 2014 [115]. Despite the development of newer platforms, some studies in the 2020s have continued to employ this method [80, 210]. The technique is based on chain termination, where DNA polymerase incorporates fluorescently labelled dideoxynucleotides (ddNTPs) alongside unlabelled nucleotides; when a ddNTP is added, elongation stops, generating fragments of varying length that can be separated by capillary electrophoresis and read through detection of fluorescence. Although capable of producing highly accurate reads of over 500 bp, Sanger sequencing is now less commonly used due to its low throughput compared with NGS platforms.

Another early sequencing technology was pyrosequencing with the GS FLX (Roche 454) platform, used in airborne eDNA studies from 2012 - 2018 [3, 76, 155, 399]. Although now discontinued, it was a short-read sequencer widely used for amplicon studies, employing a sequencing-by-synthesis approach that detected nucleotide incorporation through the light signal released during the reaction.

A less common approach is the PhyloChip microarray, which was applied by Brodie et al. in 2007 [55] to identify bacteria and archaea through variation in the 16S rRNA gene. The method works by hybridisation, where fluorescently labelled sample DNA binds to complementary oligonucleotide probes on the chip, and the resulting signal indicates the presence of specific taxa. PhyloChip can rapidly and accurately detect more than 8,000 microbial strains, including low-abundance organisms, but is restricted to the 16S region and provides less reliable information on relative abundance compared with sequencing approaches.

Ion Torrent sequencing was employed in fungal ITS airborne eDNA studies published in 2018 and 2020 [28, 29]. Like pyrosequencing, it uses a sequencing-by-synthesis approach, but instead of light signals, a semiconductor is used to detect the release of hydrogen ions as nucleotides are incorporated into the template strand. Depending on the specific system, Ion Torrent typically generates reads of 200–600 bp.

Across the airborne eDNA studies reviewed here, the majority used Illumina technology, spanning from 2016 to present day [142, 171, 197, 222, 274]. Illumina employs sequencing-by-synthesis, detecting fluorescently labelled nucleotides as they are incorpo-

rated into template strands. The process is highly parallelised, with many fragments sequenced simultaneously, and paired-end reads providing consensus to enhance accuracy. Typical read lengths range from 100–600 bp.

PacBio is a long-read sequencing platform with high accuracy, which has rarely been used in airborne eDNA studies [306]. Compared with other long-read technologies such as ONT, it is more expensive and requires larger instruments, leading either to high start-up costs or the need to send samples away for sequencing. In PacBio HiFi sequencing, a DNA polymerase moves along a circular DNA template, incorporating fluorescently labelled nucleotides that emit light pulses used to determine the sequence. This circular consensus approach allows each read to be sequenced multiple times, generating high-accuracy reads of up to 30 kb. Most airborne eDNA studies to date have focused on amplicon sequencing, which reduces the relevance of long reads. Nonetheless, PacBio has potential for future applications, particularly in long-read metabarcoding as outlined in section 2.3.3.1, and in WGS studies.

Oxford Nanopore Technologies ONT is another long-read sequencing technology which so far has been used in one airborne eDNA study to construct metagenome-assembled genomes (MAGs) from environmental data [308]. ONT devices provide portable, real-time, and comparatively low-cost sequencing that can be run from a laptop and support multiplexing of samples. These features make nanopore sequencing particularly suited to rapid detection and the long-term goal of *in situ* airborne eDNA analysis. For this reason, all sequencing in this thesis was carried out using ONT platforms.

Nanopore sequencing devices use flow cells containing arrays of nanopores embedded in a membrane, each connected to an electrode and sensor chip. As DNA is unwound, a single strand passes through a pore, disrupting the ionic current in a characteristic way, with each nucleotide affecting the signal differently. These current traces, known as "squiggles" are then decoded by basecalling algorithms to determine the sequence in real time (Figure 2.3). ONT offers different library preparation kits, flow cells, and basecalling options, enabling researchers to tailor sequencing according to the objectives of each study.

During library preparation, ONT offers two main kit types: rapid and native. The rapid kit uses a transposase to fragment DNA and attach adapters in a single step, allowing sequencing to begin within minutes. The native kit requires ligation of sequencing adapters, a longer process but one that preserves DNA integrity and supports longer read lengths. In airborne eDNA studies, researchers must balance the need for fast results with the advantages of longer reads when selecting the kit to use.

The prepared library is pipetted onto a flow cell for sequencing. ONT offers several flow cell types that differ in pore number and data yield. Because the pores are biological, not all are active and ONT sets warranty levels below which a flow cell is replaced. The Flongle has a theoretical maximum of 126 pores (warranty level: 50) and can generate up to 2.8 Gb of data, while the MinION flow cell has up to 2,048 pores (warranty level: 800) with a yield of 50 Gb. The largest, the PromethION, contains up to 12,000 pores (warranty level: 5,000) and can produce as much as 290 Gb of data. Both the Flongle and MinION are portable devices that connect to a laptop via USB, unlike the PromethION which requires a dedicated instrument. The choice of flow cell depends on the sequencing depth required and the number of samples multiplexed. For example, sequencing six samples on a Flongle would generate $\sim$ 0.46 Gb per sample, whereas the same six samples on a PromethION

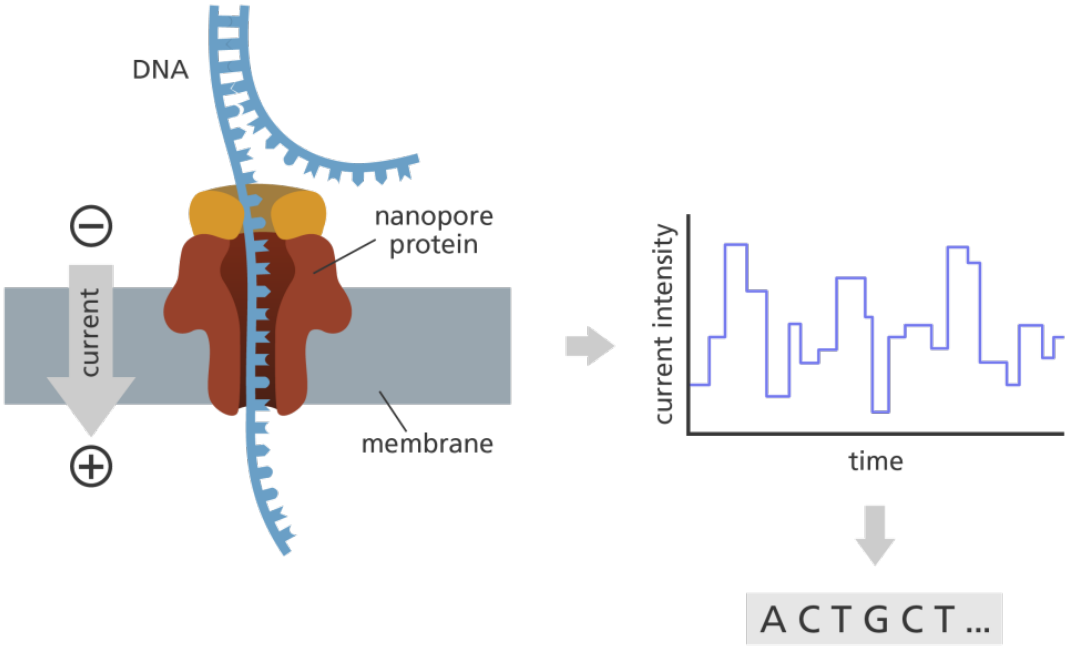


Figure 2.3: Schematic representation of nanopore sequencing technology. Image credit: Laura Olivares Boldú, Wellcome Connecting Science [270].

would produce ∼ 48 Gb each. ONT also offers barcoding kits (24- and 96-plex), allowing users to balance the number of samples with per-sample data yield, cost, and experimental goals.

Basecalling can occur concurrently with sequencing. ONT provides three main basecalling models: Fast, high accuracy (HAC), and super accuracy (SUP). The Fast model prioritises speed and can keep pace with data generation, enabling real-time sequence output. The HAC model offers improved raw read accuracy but requires more computational resources. The SUP model delivers the highest accuracy, though at the greatest computational cost. Data initially basecalled with Fast can later be rebasecalled using HAC or SUP.

Beyond library preparation, flow cell and basecalling there are other considerations with nanopore sequencing. Since it is a direct DNA sequencing technology, the extraction protocol is critical: the read length is limited by the length of DNA strands provided, contamination within the sample can damage the flow cell and reduce sequencing capacity, and tertiary structures on DNA can block pores and lower yield. Another limitation is accuracy, which remains lower than other next-generation sequencing platforms. ONT reports raw read accuracies exceeding 99% with its most recent SUP models (v5), yet independent benchmarking typically finds lower values. For example, testing of ONT basecalling models across nine bacterial genomes identified mean accuracies of 91% with the Fast model, 96% with HAC, and 97.1% with SUP using v4.2.0, improving to 97.7% with the updated v4.3.0 SUP model [390, 391]. These results highlight accuracy differences between the models and the improvements point to progress, though real-world performance often falls below ONT's headline values.

Historically, Illumina has been the platform of choice for amplicon-based airborne eDNA studies, whereas ONT provides unique advantages for whole-genome and real-time applications. For airborne plant pathogen detection, additional benefits of ONT include

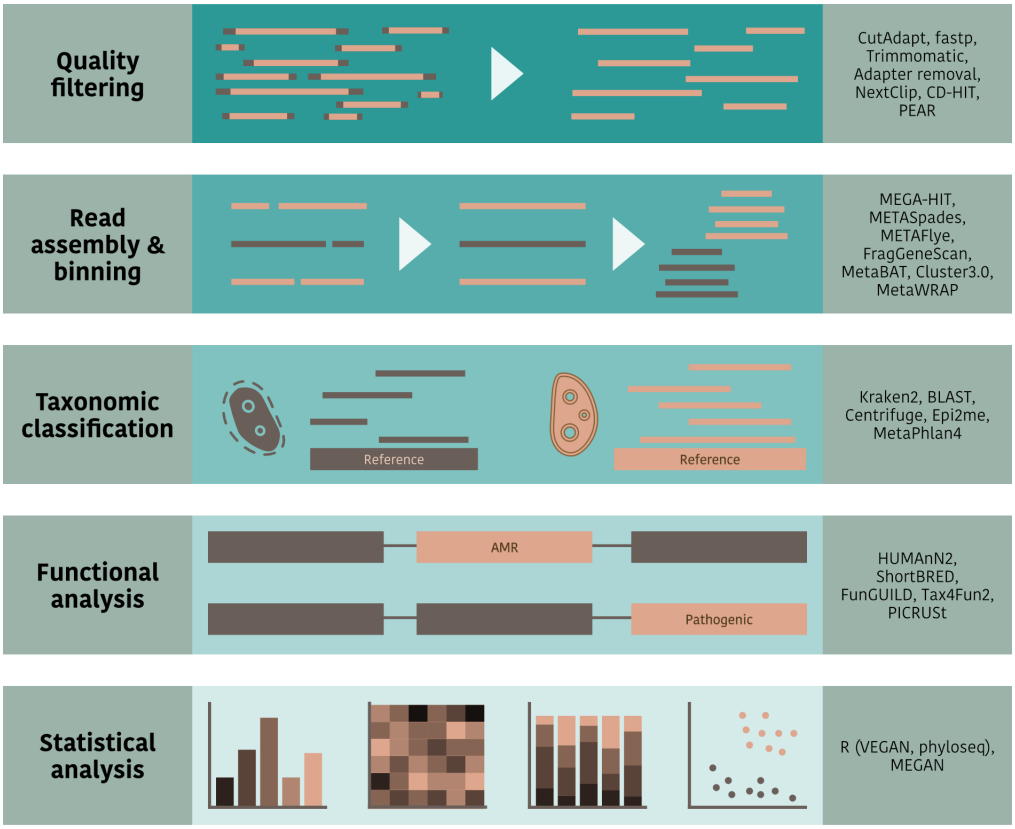


Figure 2.4: Stages in bioinformatic analysis for sequence data generated from air samples. On the right of each row is listed a selection of commonly used tools. Further details of the tools can be found in Table 2.2. From [36].

portability, affordability, and real-time sequencing outputs, which together support the goal of rapid *in situ* identification. As this thesis focuses on WGS, the longer reads produced by ONT are particularly valuable for achieving accurate taxonomic assignments.

2.3.4 Bioinformatics

As with other metagenomic studies, analyses of airborne eDNA typically address two key questions: the taxonomic composition of the sample ("what is present?") and the functional potential of the community ("what is it doing?"). The following section outlines the principal bioinformatic steps involved and illustrates how these have been applied to airborne datasets. Figure 2.4 summarises these stages alongside commonly used tools, with supporting references provided in Table 2.2. For comprehensive reviews of metagenomic data processing and its implications for both taxonomic and functional inference in amplicon and WGS studies, see [50, 195, 316, 402].

Table 2.2: Summary of commonly used tools for bioinformatic processing. Further details can be found in [359] and in the review articles referenced in the first column. Table from [36].

Stage	Specific purpose	Tool name	Ref.
Quality filtering See also [150]	Remove adaptor sequences	CutAdapt	[241]
		fastp	[75]
		Trimmomatic	[48]
		Adapter removal	[326]
	Remove duplicated reads	NextClip	[212]
		CD-HIT	[116]
	Map reads to human reference and remove from sequences	Minimap2	[217]
		BowTie	[208]
	Merge paired-end reads	PEAR	[411]
Read assembly See also review of assembly tools [93, 163]	Generate longer contigs from short reads to generate MAGs	MEGAHIT	[216]
		metaSPAdes	[266]
		metaFlye	[202]
		metaMDGB	[35]
	Identify fragmented genes in short reads	FragGeneScan	[310]
Binning See also the discussion of binning tools for WGS [234] and amplicon [139]	Group contigs together based on sequence similarity	MetaBAT	[190]
		MaxBin	[396]
		CONCOCT	[11]
		Cluster3.0/Pycluster	[156]
		MetaWRAP	[374]
Taxonomic classification See also review of classification tools [361]	Identify taxa from sequences by alignment to reference genomes	minimap2	[217]
		BLAST	[13]
		BWA	[219]
	Identify taxa from sequences with a k-mer-based approach	Kraken2	[395]
		Centrifuge	[196]
		MASH	[272]
	Identify taxa from sequences with a marker-gene approach	MetaPhlan4	[114]
Functional analysis	Identify genes with known functions	HUMAnN2	[114]
		ShortBRED	[189]
		FunGuild	[260]
		Tax4Fun2	[384]
		PICRUSt	[93]
Statistical analysis See also [227, 276, 385]	Perform abundance and diversity measures	VEGAN in R	[269]
		Phyloseq in R	[246]
Visualisation	Plot graphs of the analysis	Ggplot2 in R	[377]
Multiple stages	Used in combination with blast to classify taxa and can be used to carry out statistical analysis and visualisation	MEGAN and MALT	[151, 166]

Continued on next page

Stage	Specific purpose	Tool name	Ref.
	Bioinformatic workflows specific to Oxford Nanopore reads (basecalling, alignment and assembly)	Epi2me	https://labs.epi2me.io/
	Real-time classification and AMR analysis for nanopore sequencing	MARTi	[281]
	Amplicon-specific software tool that requires fastq files as inputs, which are then quality filtered, clustered into ASVs and assigned to taxa	DADA2	[62]
	A microbiome platform that is open-source, free and community-developed; the tool records the steps in bioinformatic pipelines to ensure they are reproducible and reusable by others	Quantitative Insights Into Microbial Ecology (QIIME 2)	https://qiime2.org/
	A tool designed to bin >100 bp reads; can also assign taxonomy, build phylogenetic trees, perform functional analysis and compare datasets	MetaBin web server	[333]

2.3.4.1 Quality filtering

Quality filtering is an essential pre-processing step that reduces the likelihood of misidentification arising from low-quality reads containing sequencing errors. The process typically begins with the removal of sequencing adaptors, the merging of paired-end reads where applicable, and the demultiplexing of libraries containing multiple samples. Subsequent filtering is usually applied on the basis of read length and quality scores, with specific thresholds determined by the sequencing platform and the aims of the study. A detailed discussion of the importance of quality filtering for downstream analyses is available at [150].

2.3.4.2 WGS read assembly

An optional step prior to taxonomic classification of WGS data is read assembly, in which short reads are combined into longer contigs to provide greater genomic context and thereby improve taxonomic resolution, particularly among closely related species [23]. Assembly is routinely employed in airborne eDNA data analysis, where it enables gene and protein prediction as well as alignment to reference databases for fungal spore identification and species confirmation [124, 287, 298]. Assembled contigs can subsequently be grouped, or binned, into MAGs for further analysis. Numerous computational tools are available for these tasks, and detailed reviews are provided elsewhere [122, 234, 252].

2.3.4.3 Amplicon binning

Metabarcoding data are typically clustered into operational taxonomic units (OTUs) or amplicon sequence variants (ASVs) prior to taxonomic classification. OTUs are defined

according to a sequence similarity threshold, which reduces the impact of sequencing errors but may also collapse closely related species into a single unit. By contrast, ASVs incorporate sequence abundance to differentiate true biological variation from sequencing error, offering higher resolution but with the risk of discarding rare taxa as noise.

ASVs generally provide greater taxonomic resolution, particularly at the species level, but may underestimate diversity in samples where low-abundance organisms are biologically important. This is especially relevant for airborne eDNA studies, where rare taxa may be underrepresented. In an airborne bacterial pathogen study, assembly methods have enhanced OTU-based clustering of 16S reads, improving species-level classification [160].

The choice between OTUs and ASVs, as well as the parameters applied during clustering, has a direct impact on downstream community composition analyses and therefore constitutes a critical design decision. One comparison of OTU- and ASV-based clustering of the same airborne eDNA dataset reported little agreement at the species level, but greater consistency at the genus level [7]. This highlights how methodological choices influence biological interpretation.

Comprehensive discussions of denoising and clustering strategies for metabarcoding data, are provided by Hakimzadeh et al. [139].

2.3.4.4 Taxonomic identification

Following clustering of OTUs/ASVs or processing of WGS reads, sequences are assigned to taxa by comparison with reference databases. A variety of approaches are available for this step. Early studies typically employed alignment-based methods such as BLAST [13] and similar tools. While accurate, these methods have become increasingly computationally intensive as reference databases have expanded. More recently, k-mer–based classifiers (e.g. Kraken2 [395]) and marker-gene approaches (e.g. MetaPhlAn4 [45]) have been widely adopted due to their greater efficiency. However, the accuracy of these methods is strongly dependent on the quality and completeness of the reference taxonomy, and errors in the underlying database can result in misclassifications.

The choice of database is therefore critical, as only species represented in the reference set can be detected. If an organism is absent, reads may be incorrectly assigned to a closely related taxon, generating false positives. Comparative studies have shown that community composition can differ markedly depending on the database used [346]. This problem is particularly acute in airborne eDNA studies, where many airborne organisms are absent from reference collections. As a result, some sequences are classified at higher taxonomic ranks or misassigned to related taxa. Additional complications arise from poor-quality reference genomes, contamination, or annotation errors, all of which can distort taxonomic assignments.

Some classification methods employ lowest common ancestor (LCA) assignment, whereby reads are placed at the lowest taxonomic rank consistent with all reliable matches. Although expanding databases have increased the proportion of reads that can be classified, they have also made species-level assignments more challenging. This is because the number of new species added has outpaced the addition of higher-level taxa, creating greater ambiguity at finer taxonomic ranks. Consequently, LCA approaches are now more likely to assign reads to the genus level rather than to species [259]. The increasing size of reference collections also places heavier demands on computing resources and complicates efforts to

maintain fully updated databases. Further discussion of these challenges is provided in [78].

Airborne eDNA studies frequently map WGS reads to large resources such as the NCBI nucleotide or non-redundant databases [138, 197, 228, 262, 290, 308], though some studies use smaller curated datasets focused on particular taxa or research aims. Metabarcoding studies typically rely on gene-specific databases (e.g. ITS, 16S, 18S) to reduce computational load and minimise misalignments. For fungi, the UNITE database [263], which contains only ITS sequences, has been applied in studies of airborne *Alternaria spp.* [18], rice pathogens [113], and forest fungal communities [233]. For bacteria, the SILVA rRNA database [339], which includes both 16S and 18S sequences, has supported investigations of airborne bacteria in urban Tokyo [370], analyses of viable airborne microorganisms [127], and surveys of marine bacterial aerosols [243].

In some cases, studies have developed custom reference databases tailored to their research aims. For instance, viral genomes from NCBI have been used to investigate seasonal dynamics of the airborne virome [295], while the Pathogen–Host Interaction database (PHI-base) [372] was employed to construct a pathogen-specific database for the identification of airborne plant pathogens [124].

Regardless of the database selected, taxonomic assignments—particularly at species or strain level—should be interpreted with caution. Post-classification filtering can help to reduce misassignments. For example, a study of airborne eDNA in Siberia improved classification accuracy by only accepting species-level assignments when sufficient reads aligned uniquely to that species [138]. Appropriate thresholds, however, are study-dependent and influenced by factors such as amplification strategy and sequencing depth. In studies without DNA amplification, rare species may be represented by only a few reads, and applying stringent thresholds could exclude genuine low-abundance taxa, leading to an underestimation of diversity.

2.3.4.5 Functional analysis

In addition to taxonomic classification, functional analysis of airborne eDNA sequence data provides insight into the potential activity of detected organisms. Metabarcoding data can only be used to infer function indirectly, through the known traits of identified taxa, whereas WGS enables the detection of specific genes or transcripts. WGS therefore offers greater functional resolution, though its application is often constrained in airborne studies by limited DNA yield and sequencing depth.

Most functional analyses of air samples have focused on antimicrobial resistance (AMR) genes, commonly using resources such as the Comprehensive Antibiotic Resistance Database [178] and the Structured Antibiotic Resistance Gene Database [404]. For example, WGS has been applied to investigate airborne AMR genes in public transport systems [215] and hospitals [147], while qPCR has been used for real-time detection of AMR genes in poultry sheds [349].

Functional analysis has also been used to explore plant pathogenic potential in airborne fungi [7, 76], with ITS data analysed through tools such as FUNGuild [260]. Similarly, studies of airborne bacteria along the Antarctic coast have applied 16S rRNA gene analysis to infer functional profiles associated with biosynthesis and metabolism [64]. At a broader scale, a global meta-analysis of airborne microbiomes demonstrated that anthropogenic

activity correlates with an increased abundance of pathogenicity genes [181], consistent with findings from China where reduced air quality was associated with higher relative abundance of pathogenic bacteria, as measured by qPCR of specific genes [400].

There is considerable scope to extend functional analyses to include additional traits such as virulence and toxicity, which have been characterised in marine and soil metagenomes [34, 105]. However, to date no published studies have applied these approaches to airborne metagenomic data. Expanding functional investigations of airborne communities has the potential to provide valuable insights into antimicrobial resistance, pathogenicity, and metabolic capacity, thereby advancing understanding of how these organisms interact with and respond to their environments.

2.3.4.6 Statistical analysis

Following taxonomic classification, data can be analysed using statistical software and custom bioinformatic pipelines to explore diversity, identify patterns, and generate ecological insights relevant to environmental health and microbial dynamics. Read counts per taxon are commonly used to calculate alpha and beta diversity across sample sets. Several reviews provide detailed discussion of statistical approaches to metagenomic data, which are applicable to airborne eDNA data [227, 276, 385].

Airborne studies often extend beyond standard diversity metrics to include bespoke analyses. These include the detection of specific pathogens [113, 123, 392], correlation of microbial community composition with meteorological and environmental variables [124, 287, 370], and assessments of how airborne eDNA changes with distance from sources such as wastewater treatment plants [221] or from land across the ocean [243].

2.4 Current Applications of Airborne Sampling

Sequencing of airborne eDNA has been applied across a wide range of fields, including plant pathogen surveillance, high-altitude and oceanic air monitoring, human health applications, and biodiversity assessment (Figure 2.1).

2.4.1 Detection of plants and plant pathogens

Airborne eDNA has been widely applied to monitor plant pathogens in agricultural and forestry systems. Traditionally, these environments relied on passive samplers, where spores landed on adhesive tapes and were then identified through microscopy. Although effective, this approach is labour-intensive and requires specialist expertise. The advent of sequencing-based methods has enabled faster, more affordable, and more accurate detection.

Direct comparisons of microscopy and metagenomics on spore-trap samples demonstrated that both approaches detected major cereal pathogen genera, but sequencing additionally revealed pathogenic species not identified microscopically [287]. This highlights the capacity of DNA-based methods to enhance the resolution of air sampling.

Molecular approaches also facilitate large-scale monitoring without extensive field surveys. For instance, qPCR targeting known forest pathogens across 12 sites successfully identified wind-dispersed diseases [7], while WGS of UK air samples detected several agriculturally important crop pathogens [124]. Similarly, seasonal monitoring of rice pathogens

using qPCR and ITS metabarcoding enabled links to be drawn between pathogen abundance, disease severity, and environmental conditions [113].

The large datasets generated by sequencing can be integrated with climate and meteorological variables to improve models of pathogen spread [378] and to investigate broader interactions between biological particles and the atmosphere, such as ice nucleation processes [51]. In some cases, these analyses reveal reciprocal relationships; for example, a global air-sampling network showed that annual mean temperature significantly influences fungal spore community composition [274]. Airborne sampling has also been applied to plant pollen, as demonstrated in a study of genetically modified maize in which real-time qPCR was used to quantify the ratio of modified to wild-type pollen along a distance gradient, providing insights into pollen dispersal [111].

2.4.2 Sampling at high altitudes and above the ocean

Airborne eDNA sampling at high altitudes and over oceans provides opportunities to investigate microbial diversity in environments that are otherwise difficult to study. Sampling height above ground is a key determinant of community composition [94]. While most collections occur near ground level, some studies have classified aerosols at altitudes up to 12,200 m above sea level, using aircraft fitted with adapted air-sampling devices programmed to operate at different heights and times [94, 171, 251, 348].

High-altitude studies have shown that microbial communities change markedly with elevation. For example, above 1,000 m microorganisms no longer exhibited the diurnal cycles observed nearer the ground [94]. In-flight sampling has also demonstrated the influence of environmental events, with collections above high-intensity forest fires showing increased microbial concentrations and enhanced diversity of viable cells [201]. Such findings highlight the importance of considering sampling height when comparing airborne eDNA studies.

Oceanic air sampling has similarly revealed novel insights [243, 336]. For instance, a study along the Antarctic coast identified airborne bacteria carrying genes associated with growth and survival at the ocean surface [64]. These results illustrate how airborne eDNA reflects interactions between terrestrial, atmospheric, and marine systems.

Together, high-altitude and oceanic surveys broaden the scope of airborne eDNA research beyond conventional environments. Exploring these less-studied settings has the potential to advance understanding of microbial contributions to ecological processes and atmospheric dynamics.

2.4.3 Sampling for human health applications

Airborne eDNA sampling has been applied in healthcare settings to establish baseline microbiome profiles, providing reference points against which unusual increases in pathogen abundance can be detected [197]. During the SARS-CoV-2 pandemic, air sampling was carried out across hospitals to identify high-risk transmission zones [42, 210, 305].

Wearable samplers have also been trialled in clinical contexts, such as for detecting monkeypox in hospitals [130]. Although healthcare workers were equipped with personal protective equipment in these studies, wearable devices have the potential for monitoring occupational exposure to airborne pathogens. Similarly, air samples collected from rooms of norovirus patients have provided insights into transmission dynamics, showing that

viable viral particles can remain airborne and that their abundance correlates with recent vomiting events [12].

While many of these studies employed amplicon-based approaches, WGS has also been used to identify widespread airborne AMR genes, which were found to be more abundant and diverse inside hospitals compared with nearby outdoor environments [147].

Collectively, these findings highlight the value of air sampling in clinical settings, offering tools to monitor pathogen dynamics, assess worker exposure, and track the distribution of AMR genes.

2.4.4 Biodiversity assessment from sampling

Molecular approaches have become increasingly important for biodiversity monitoring, and airborne eDNA offers particular advantages. Compared with conventional surveys, air sampling can reduce labour requirements, extend monitoring to otherwise inaccessible areas (e.g. via drones), and detect species that are difficult to observe directly. Several proof-of-concept studies have demonstrated its potential.

Airborne detection of vertebrates has been shown in zoo settings, where non-native species were identified at distances of up to several hundred metres, and DNA from the animal feed was detected [80, 228]. A separate study compared Burkard spore-trap samples with camera trap images, showing complete overlap in species identified by both methods, alongside additional species detected only from air samples [290]. Air sampling in a botanical garden identified 67 plant species out of 1,585 known to be planted in the garden, with the remainder likely undetected due to seasonal variation in pollen release or insufficient sampling effort [363].

Beyond targeted biodiversity surveys, existing air quality monitoring infrastructure has also been repurposed. For example, a 34-year archive of weekly filters from a radionuclide monitoring station (each processing over 100,000 L of air) was analysed by WGS, revealing more than 2,700 genera and demonstrating links between land-use change and biodiversity decline [355]. Similarly, amplicon sequencing of filters from air quality stations identified over 180 taxa, including vertebrates [222]. These examples highlight the potential to assess biodiversity retrospectively, provided filters have been appropriately preserved.

Together, these findings illustrate the versatility of airborne eDNA as a biodiversity monitoring tool, enabling both real-time species detection and the reconstruction of historical biodiversity patterns through existing monitoring networks.

2.5 Current Research Gaps and Challenges

2.5.1 Is captured DNA viable?

Detection of airborne eDNA does not necessarily indicate that the source organisms are viable, an important distinction for both agricultural and medical applications. In crop protection, fungicide treatments are only warranted if detected spores remain viable, while in human health, public safety measures depend on whether airborne pathogens pose an actual infection risk.

The most established approach to assessing viability is culturing, whereby airborne particles are collected directly onto a growth medium and colonies are counted after incubation. This can be achieved using selective media targeting a specific pathogen [167] or

in combination with sequencing to link viability to broader community composition [171]. However, culturable fungi typically represent less than 1% of the taxa detected in airborne eDNA studies [244], meaning the viability of most airborne organisms cannot be assessed in this way.

Alternative molecular approaches have therefore been explored. One strategy leverages the fact that metabolically active cells contain higher levels of Ribosomal RNA (rRNA). Therefore, species with a high rRNA-to-DNA ratio are more likely to be active than those with a low abundance of rRNA. A study applying this principle to airborne eDNA identified a significant difference in the taxonomic abundance of DNA and RNA communities from the same sample, suggesting that not all the identified species were active [127]. However, this approach may be less applicable to organisms that disperse in a dormant state, such as spores, which remain viable but exhibit little or no rRNA activity.

Emerging technologies may offer new solutions. Nanopore sequencing of native DNA can detect modifications and damage as molecules pass through the pore, producing electrical signals that differ between viable and non-viable cells. Recent work has used machine learning models trained on DNA from live and dead cells to distinguish viability signatures in sequencing signals [373]. Although still at an early stage, this method has potential to provide culture-independent viability assessments within metagenomic workflows.

2.5.2 Low DNA biomass in the air

Air typically contains very little biological material, resulting in low concentrations of DNA available for sequencing. For example, airborne concentrations of SARS-CoV-2 have been estimated at only 0.87 genome copies per litre of air, which was considered insufficient for NGS but suitable for detection with Sanger sequencing [210]. More generally, outdoor air has been reported to contain approximately 10^5 bacterial and 10^6 virus-like particles per cubic metre [294].

DNA concentrations also vary temporally and spatially. In Siberia, yields were 170-fold higher in summer than in winter [138]. Whilst another study reported greater read numbers in spring compared with autumn, likely reflecting plant growth, flowering, and pollen release during spring months [183]. Location is also important, in Antarctica, samplers operating continuously for two weeks and processing 16,000 m^3 of air still yielded DNA concentrations comparable to negative controls [92].

Given these challenges, several studies have developed methods to improve DNA recovery from low-biomass samples. Optimising sampling, storage and extraction has been shown to increase DNA accumulation by 8–170-fold compared with earlier protocols [226]. Similarly, the MetaSUB project introduced a multi-step extraction pipeline combining chemical lysis buffers, centrifugation to remove DNA from filters, and subsequent enzymatic and mechanical lysis. This approach substantially increased DNA yield and diversity in complex airborne samples such as subway air, although it was less effective in simple mock communities [47].

2.5.3 Preventing contamination

Airborne eDNA studies are particularly vulnerable to contamination due to the inherently low biomass of samples. Even small amounts of extraneous DNA can persist through sequencing pipelines and distort community profiles. Amplification steps used in metabar-

coding further increase this risk by propagating trace contaminants [228]. Contamination is a recognised challenge across eDNA research, and several reviews outline strategies to minimise both cross-contamination between samples and external contamination originating from laboratory environments, reagents, or personnel [63, 330].

External contamination can be reduced through sterile laboratory practices and monitored by including negative controls at all stages of sampling and processing [179]. Such controls are essential, as DNA extraction kits and consumables themselves may contain exogenous DNA, which will appear even under sterile conditions [319, 387]. Taxa detected in negative controls can later be scrutinised during analysis to help determine whether they represent genuine signals or contaminants [215].

Human DNA presents an additional complication. Contamination may arise from researchers or individuals present in the sampling area, a phenomenon sometimes described as "human genetic bycatch". Beyond technical concerns, this raises ethical considerations. For example, one study reported that haplotypes of individuals present in a room could be reconstructed after just five hours of sampling [389]. In public spaces it is not feasible to obtain informed consent from all individuals, and current best practice is to remove reads aligning to the human genome before data publication.

2.5.4 Linking sequence data to species abundance

A major limitation of current eDNA analysis is the difficulty in translating sequencing data into absolute measures of species abundance. Although DNA concentrations in air samples can be quantified following extraction, the amount of DNA varies widely between taxa, preventing a linear relationship between DNA yield and organism abundance.

Additionally, airborne eDNA datasets are inherently compositional, reflecting relative rather than absolute abundances. As a result, an increase in one taxon's representation can artificially reduce the apparent abundance of others, even in the absence of real biological change. This complicates interpretation in longitudinal studies, where shifts in relative abundance may reflect broader community dynamics (e.g. seasonal pollen release) rather than true population changes. For a detailed discussion of compositionality in eDNA datasets, see [126].

Several strategies have been proposed to address this challenge. The addition of synthetic DNA "spike-ins" at known concentrations can enable semi-quantitative estimates of airborne DNA yield, and this approach has been applied to fungal monitoring studies [274]. Empirical evidence also supports a partial relationship between relative abundance in sequencing data and organism abundance. For example, in agricultural systems, pathogen abundance inferred from WGS correlated with environmental factors over the growing season [124], while wind tunnel experiments demonstrated that higher spore release rates led to greater relative representation of those taxa in sequencing data [124].

In principle, with sufficient genomic coverage, it should be possible to estimate effective population sizes from airborne eDNA. This has already been achieved in aquatic systems [338], but the low concentrations of DNA typically recovered from air currently preclude such analyses. Advances in sampler sensitivity and sequencing technologies may eventually make population-level inference from airborne eDNA feasible, opening new opportunities for ecological and epidemiological studies.

2.5.5 Microbial source identification

The dispersal of airborne biological particles is influenced by particle size, wind speed and direction, and a range of meteorological factors. For taxa that rely on aerial transport, such as fungal spores and pollen, these dynamics reflect evolutionary adaptations to dispersal. Some species remain close to their origin where conditions are favourable, while others disperse over long distances, in some cases reducing host density to limit pathogen pressure. Because multiple interacting variables affect particle movement along a trajectory, it is often extremely challenging to trace airborne material back to its source using meteorological data alone.

Source attribution is nevertheless critical in applications such as disease control and bioterrorism monitoring. For example, detection of foot-and-mouth disease DNA in an air sample would necessitate rapid identification of the source farm to enable containment, a principle equally relevant to plant and human pathogens. In many airborne eDNA studies, source inference is based on ecological context—for instance, attributing wheat pathogens to nearby wheat fields. More detailed analyses combine biological detection with modelling approaches. One study, for example, tracked *Alternaria* spores across 38 sites from the Pannonian Plain to Poland using Hirst traps, identifying transport events through elevated source concentrations, atypical spore signatures at the destination, and atmospheric footprint modelling of air mass trajectories [133].

Atmospheric footprint modelling commonly employs particle dispersion models such as Hybrid Single Particle Lagrangian Integrated Trajectory (HYSPLIT), which use meteorological data to reconstruct likely air mass pathways [351]. In practice, however, accuracy is constrained by uncertainties in particle size, topography, wind patterns, humidity, and precipitation. Despite these limitations, dispersion models have been successfully applied in airborne pollen and eDNA studies to improve source attribution [77, 133, 142, 251]. These models can also inform strategic sampler placement, as discussed in Section 2.3.1.3.

Ongoing developments in aerobiology, including the establishment of wider sampling networks and increasingly refined dispersion models, are enhancing the resolution of microbial source tracking. While pinpointing exact origins remains difficult, integrating meteorological modelling with airborne eDNA analysis offers a promising pathway towards more accurate source attribution and improved sampling strategies.

2.6 Future Directions in Airborne eDNA Analysis

Airborne eDNA research is progressing rapidly, transitioning from traditional culture-based and microscopy methods towards real-time molecular detection. Continued technological development is expected to expand the scope and resolution of this field, deepening understanding of microbial communities and their environmental and health impacts. Anticipated advances include continuous monitoring, *in situ* testing, plant pathogen surveillance networks, and improved sequencing approaches for low-input samples.

Proof-of-concept studies already suggest that real-time pathogen surveillance in agricultural fields, livestock facilities, or high foot-traffic environments such as airports is feasible, with the potential to provide early alerts to farmers or public health authorities [144, 215, 386]. Portable air samplers could also support personalised exposure monitoring [6], while endangered species may be tracked non-invasively through detection of their airborne DNA

[80, 228].

In healthcare settings, airborne samplers could play an important role in mitigating disease transmission, particularly in high-risk areas. For example, continuous monitoring in wards housing immunocompromised patients could provide early warnings of harmful bacteria, enabling rapid intervention. A detailed review of indoor air sampling for public health, including available devices, is provided by Memon et al. [248].

Advances in hardware miniaturisation, power efficiency, and computational capability are making *in situ* testing increasingly feasible in places such as agricultural fields, hospitals, and public transport networks. The principal barrier to widespread deployment remains the automation of DNA extraction and library preparation, processes that are still largely laboratory-based. While portable amplification methods such as PCR and LAMP are already available. There are a number of emerging digital microfluidic devices, including ONT's TraxION and Integra Biosciences' MIRO CANVAS, which offer the potential to automate upstream protocols. Such innovations could enable fully integrated, field-deployable sequencing platforms.

In situ testing has the potential to deliver faster diagnostics and support rapid decision-making by eliminating delays associated with sample transport and laboratory processing. This is particularly critical in responding to human health threats or plant pathogen outbreaks. Direct, on-site testing also removes the need for storage and transport, thereby reducing the risks of sample degradation and contamination.

Over the next decade, sequencing-based pathogen surveillance could transform agricultural management. Existing fungal spore monitoring networks, such as those used in French vineyards [209], could be scaled up to model and predict disease severity across wider regions. These networks could also be integrated with established air quality monitoring infrastructure [222], enabling early detection of emerging pathogens, an increasingly urgent challenge in the context of climate change and globalisation [344]. Within the supply chain, air samplers could be deployed in storage facilities to monitor postharvest diseases, reducing losses and preventing the accumulation of harmful mycotoxins. Looking further ahead, integration of *in situ* sequencing with precision agriculture could allow real-time application of treatments, for example by linking airborne pathogen detection directly to tractor-mounted spray systems.

As sequencing technologies advance, with improved samplers, low-input protocols, and reduced error rates, airborne eDNA analysis is likely to become increasingly precise, sensitive, and scalable. These developments could enable the detection of antimicrobial, fungicide, and herbicide resistance genes within pathogens [215, 221, 298], facilitate strain-level taxonomic resolution [106, 164], and support single nucleotide polymorphism (SNP) based analyses for estimating effective population sizes, an approach with particular value for monitoring endangered species [338].

Realising this potential will require continued interdisciplinary collaboration and technological innovation. Improvements in sequencing accuracy, expansion of reference databases, and refinement of source-tracking models will all enhance analytical reliability. At the same time, overcoming current challenges within *in situ* processing, viability assessment, and sensitivity of detection will broaden the range of applications across health, agriculture, and environmental monitoring.

Chapter 3

AirSeq in a horticultural setting

3.1 Abstract

This study investigates the ability of AirSeq to detect airborne crop pathogens in a horticultural setting. Traditional manual disease scoring was carried out alongside air sampling to assess the presence of three strawberry pathogens *Botrytis cinerea*, *Podosphaere aphanis* and *Phytopthora spp.*. Air samples were collected monthly over the course of a year from three locations at Wilkin & Sons Ltd in the East of England including an open field and two covered greenhouses. DNA was extracted from the air samples, whole genome sequenced, and analysed in combination with the disease score and environmental data to better understand pathogen dynamics throughout the growing season and under varying environmental conditions. The results revealed a strong correlation between AirSeq data and manual disease scores, demonstrating that AirSeq can accurately identify pathogens and that read count fluctuations reflect changes in disease incidence. Furthermore, fungicide application did not reduce pathogen prevalence in the air or on crops, suggesting limited efficacy. Additionally, a comparison of *P. aphanis* fungal spores collected directly from the plants and *P. aphanis* material collected from the air showed high genome similarity. This work highlights the potential of AirSeq as a reliable, non-invasive tool for monitoring airborne plant pathogens. By detecting multiple pathogens and tracking their prevalence over time, AirSeq shows promise for broader applications across diverse crops and environments, offering a valuable approach for improving disease management and forecasting in agriculture.

3.2 My Contributions

In this chapter, I contributed to multiple aspects of the research. I participated in the fieldwork with Dr Leggett or Dr Heavens to collect samples and, under Dr Heavens' guidance, performed the laboratory work from DNA extraction through to sequencing. In 2023, Dr Heavens independently collected additional powdery mildew samples and processed them in the laboratory. I was solely responsible for all bioinformatics and analysis, receiving feedback and advice from my supervisors. Additionally, I used ChatGPT (GPT-5) [273] to refine my code and improve the grammar and flow of my writing.

3.3 Introduction

In strawberry cultivation, infections caused by *Botrytis cinerea* (grey mould), *Podosphaera aphanis* (powdery mildew) and *Phytophthora spp.* can render fruit unmarketable, reducing yields and creating substantial economic losses. These diseases are responsible for some of the most serious threats to strawberry production worldwide, with grey mould alone causing annual losses estimated at $10–100 billion and yield reductions of up to 80% in untreated crops [312]. Powdery mildew is similarly damaging, reducing UK yields by 20–70%, equating to market losses in the tens of millions of pounds each year [140], while *Phytophthora* root rot accounts for an estimated $150 million in annual losses in the United States [320].

Such substantial impacts highlight the need for effective control strategies, where early detection and accurate pathogen monitoring are essential to prevent disease spreading through the crop. While environmental control within greenhouses can help reduce pathogen spread [194], these measures are often insufficient because the conditions that optimise strawberry yield, such as warmth and humidity, also favour fungal sporulation and infection [401]. As a result, many growers rely heavily on fungicides, often applied preventatively or indiscriminately, despite growing resistance, chemical bans, and consumer demand for sustainable alternatives. Additionally, current detection and prevention strategies depend on environmental indicators such as humidity and temperature, alongside visual symptoms, to guide fungicide use. A more targeted approach, enabled by advanced detection methods such as AirSeq, could provide early warning of pathogen presence, reducing both disease-related losses and the excessive use of costly fungicides and fertilisers.

3.3.1 Strawberry diseases of interest

The three taxa (*B. cinerea*, *P. aphanis*, and *Phytophthora* spp.) were selected because Wilkin & Sons Ltd routinely conduct manual disease scoring for these pathogens, enabling a direct comparison between their disease assessments and the data obtained through AirSeq. This comparison serves to evaluate the reliability of AirSeq as a tool for pathogen detection.

3.3.1.1 *Botrytis cinerea* - grey mould

B. cinerea is a necrotrophic ascomycete pathogen, meaning it grows preferentially on damaged or senescing material and causes tissue death. The disease can affect fruit both pre- and post- harvest making them unmarketable [286]. *B. cinerea* is the causative agent of

grey mould or Botrytis fruit rot, it infects all aerial parts of the plant and is characterised by large brown lesions on the fruit, a characteristic grey fuzz of conidia on these lesions and the collapse and soaking of soft tissues [249, 293, 382, 394]. *B. cinerea* is heterothallic [314], which means it requires two compatible mating types for sexual reproduction but can also reproduce asexually.

B. cinerea is a broad host-range pathogen that infects over 400 plant species, including strawberries, other soft fruits, flowers, and leafy vegetables [205]. It is of major economic significance, as grey mould infection leads to high rates of fruit rejection by growers, retailers, and consumers alike [286]. Globally, *B. cinerea* is estimated to cause crop losses valued between $10–100 billion annually [161]. In open-field strawberry production, losses can exceed 80% in the absence of fungicidal treatment [312]. Although greenhouse systems reduce the initial risk of infection, elevated temperatures in protected environments can accelerate pathogen spread once established [37].

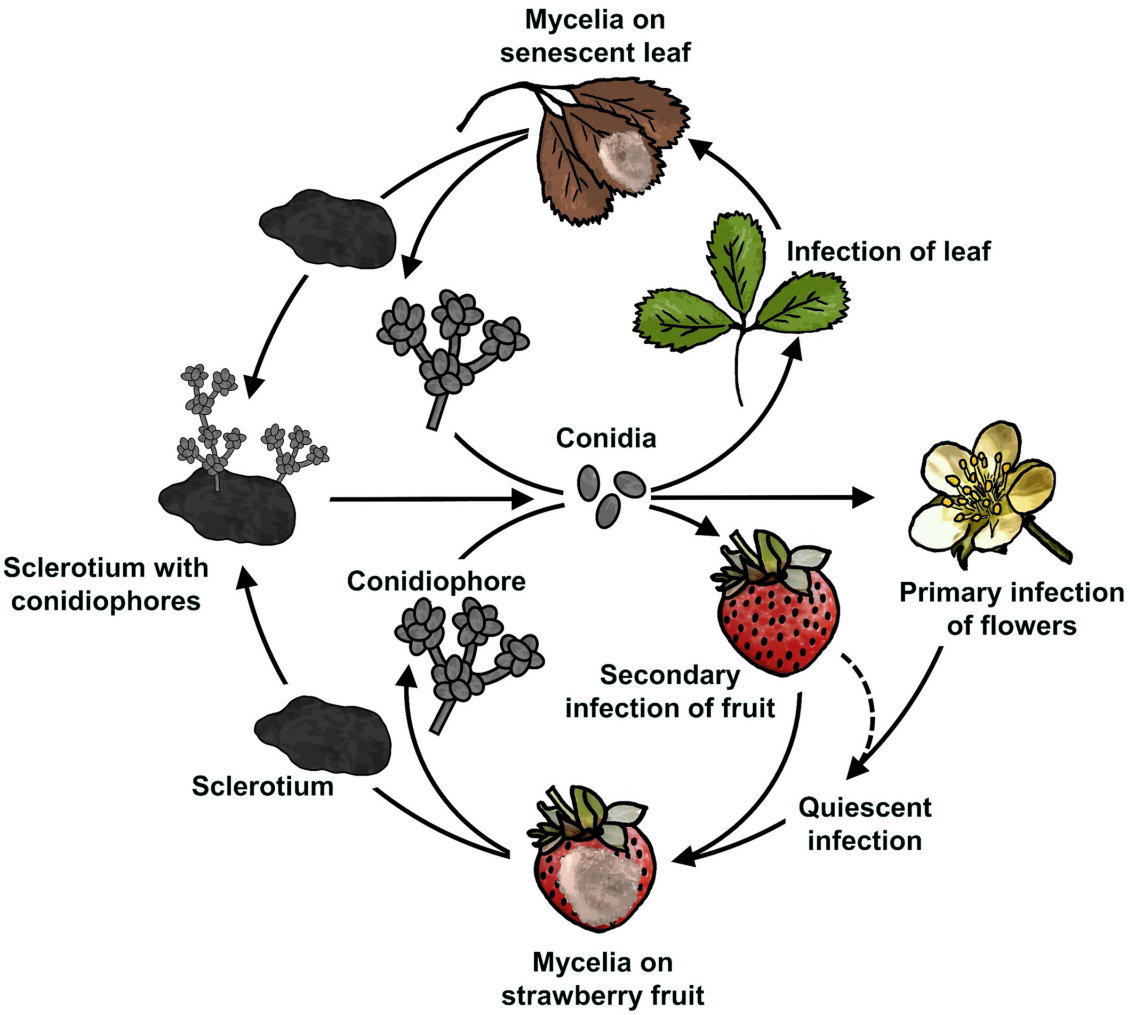


Figure 3.1: Lifecycle of *B. cinerea*, from Petrasch et al., 2019 [286]

The lifecycle of the fungus is shown in Figure 3.1, *B cinerea* over-winters as sclerotia on plant debris or in the soil. Sclerotium are a compact mass of hardened mycelium [286, 394], and in spring they begin to grow and produce conidiophores. Conidia are released from conidiophores as the primary inoculum, typically in the early morning when conidiophores dry out due to decreases in the relative humidity (RH) and warming temperatures [175]. Conidia are then dispersed by airborne currents to neighbouring plants where they infect flowers, germinate and develop hyphae which grow into the receptacle of the plant. The

fungi remain in a quiescent phase within the receptacle, which is the part of the stem the fruit develops from, until the fruit begins to grow [286, 394].

When the fruit is ripe the fungus transitions to a more aggressive necrotrophic phase. This is when the plant begins to decay due to the fungus causing tissue death, visible from the brown lesions appearing and growing on the fruit [286]. Additionally conidiophores are produced and spread conidia during this decay phase, conidia cause secondary infections in the ripe fruit of nearby plants without the quiescent phase [286, 382].

The optimal conditions for *B. cinerea* conidia germination are high RH (93%) or water droplets, reduced light and moderate temperatures (15°C - 25°C) [249, 393, 394].

The current control strategy for this disease is the application of pre- and post- harvest fungicides, although resistance is developing against these [15, 232, 382]. Airborne populations of *B. cinerea* have been monitored in previous studies using cyclone samplers coupled with an ELISA-based assay [129].

3.3.1.2 *Podosphaera aphanis* - powdery mildew

Strawberry powdery mildew, is caused by the obligate biotrophic ascomycete fungus *P. aphanis*. Infection affects the leaves, flowers, and fruit of the strawberry plant, leading to reduced quality and yield [125, 182]. Infected plants will develop white patches of fungal spores which cause the berries to be unmarketable [275]. Further disease can lead to reddish or purple blotches developing on the leaves which also begin to curl limiting the photosynthetic capability of the plant and thereby reducing fruit yields [9, 275].

Strawberry powdery mildew is a significant global concern. In the UK, a 2016 report by the Agriculture and Horticulture Development Board (AHDB) estimated that the disease can reduce crop yields by 20–70%, with a 20% loss corresponding to a market value loss of approximately £56.8 million [140]. This underscores the substantial economic burden associated with the disease and the need for effective control measures. Similarly, in Canada, severe infections have been reported to cause yield losses of up to 30% [65]. Previous microscopy-based studies have detected the spores of *P. aphanis* in the air at experimental and commercial strawberry farms [43, 66].

P. aphanis is heterothallic and the fungus survives over winter as chasmothecia, created from sexual reproduction [118]. In late summer it is possible to see chasmothecia on the underside of leaves of infected plants as small black dots [9]. In spring, when the environmental conditions are conducive, chasmothecia begin to sporulate producing conidia which are spread by wind and cause further infections on host plants [9, 285]. The lifecycle of powdery mildew on strawberries is shown in Figure 3.2

The ideal conditions for powdery mildew infection and development on strawberry plants are moderate temperatures (15°C - 25°C) and high RH which has been reported as 35% [254] or 75% - 98% [16]. Different conditions are favoured throughout the fungal life cycle. After initial infection higher temperatures (18°C - 25°C) and increased RH (97% - 100%) lead to the growth of lesions and production of conidia. The germination and spread of conidia has been shown to be favoured at 22°C and RH of 45% - 55% in laboratory conditions [170], wind speeds >0.5 m/s are also required to spread spores [9]. Additionally the production of chasmothecia, the over winter survival structure, is only initiated at temperatures below 13°C [9].

The primary control strategy for powdery mildew is the continuous application of fungi-

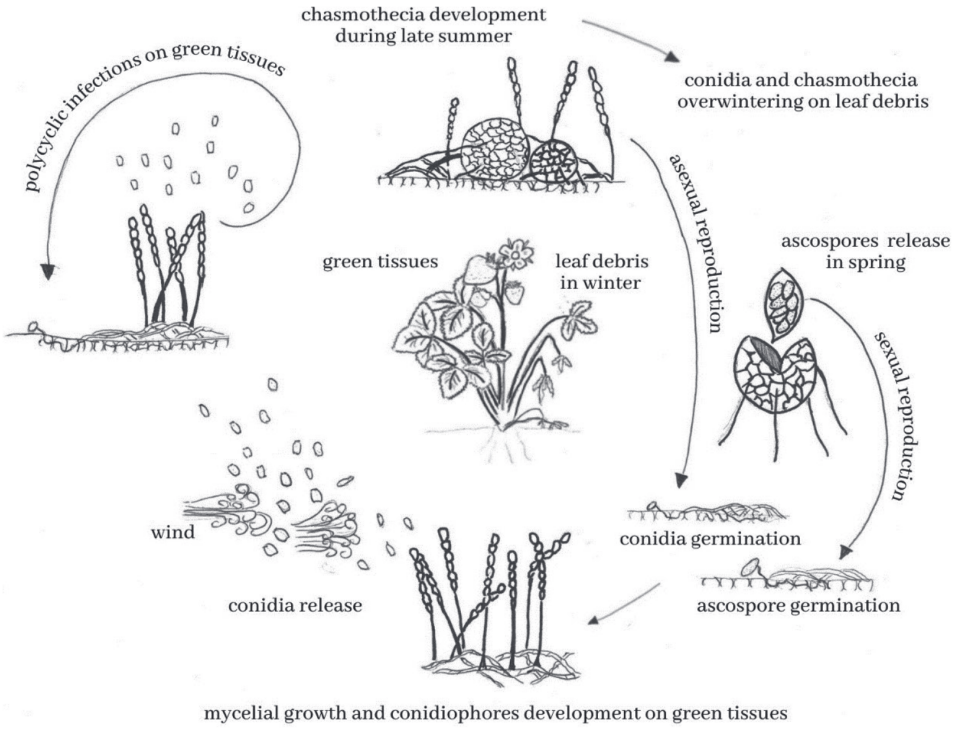


Figure 3.2: Lifecycle of *P. aphanis*, from Aldrighetti et al., 2023 [9]

cides from the emergence of new leaves until the final harvest [322]. Since young tissue is more susceptible to powdery mildew infection and strawberry plants remain in the ground for up to four years, there is a continuous production of new susceptible tissue and older debris harbouring inoculum, thereby increasing host vulnerability to *P. aphanis* [66] and making it more difficult to control.

Furthermore, interactions between diseases have been observed, with late-stage infection by *P. aphanis* potentially increasing the susceptibility of strawberry plants to *B. cinerea* [66].

3.3.1.3 *Phytophthora spp.* – crown & leather rot

Diseases caused by *Phytophthora spp.* pose a major threat to strawberry production, with estimated losses of $150 million annually in the United States alone [320]. Two species are implicated in strawberry rots, *P. cactorum*, which causes Phytophthora crown rot (PhCR) and also contributes to Phytophthora leather rot (PhLR), and *P. nicotinae*, which is involved in PhLR [240]. PhCR typically causes losses early in the season, linked to irrigation and transplantation of nursery plants, while PhLR occurs throughout the season, particularly after heavy rainfall [238].

The symptoms of PhCR are the stunting of plant growth, leaves turning blueish and wilting followed by collapse of the plants crown and the whole plant wilting [238, 332]. PhLR occurs on the berries, clearly visible on green unripe fruits as brown patches but less visible on mature fruit which may have no colour change or only slight purple or brown patches. As PhLR progresses the fruit becomes tough and leathery, hence its name. When there is high moisture, a layer of white mycelium may develop. Ultimately the infected fruits will dry out and shrivel into mummified berries [230]. Even if there is only a small part of the fruit affected by PhLR, the whole berry will have an unpleasant taste, rendering

it unmarketable, therefore farmers have a low tolerance for PhLR in their crop.

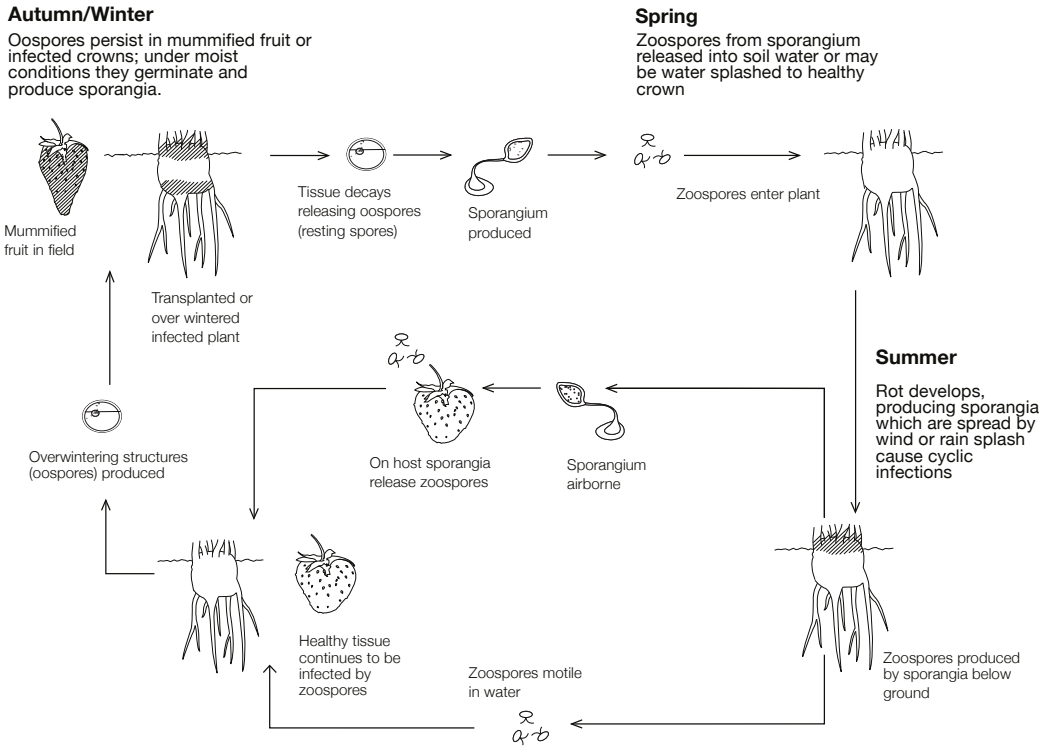


Figure 3.3: Lifecycle of *Phytophthora spp.*, adapted from Irving & Wedgewood 2007 [168]

P. cactorum and *P. nicotinae* are oomycete pathogens that are homothallic, meaning they can reproduce sexually by selfing, Figure 3.3 outlines the *Phytophthora* life cycle. The pathogens overwinter on the mummified berries, plant debris or in the soil as oospores. In warm, wet conditions, typically in spring, oospores germinate and produce sporangium containing zoospores. The zoospores then spread by free water in the soil or splash dispersal caused by rainfall or irrigation to infect new hosts [240, 309]. Zoospores enter the host by wounds, natural openings or by developing appresoriums, once infected new sporangia and zoospores are created and dispersed from the host causing secondary infections [230, 332].

In addition to over-wintering structures imported nursery plants can be a source of primary inoculum [238]. The optimal conditions for disease development are wet conditions from irrigation or high rainfall and warm temperatures (17°C - 35°C) [25, 332], the optimum temperature for sporangium production is 20°C [230].

The current control strategies for *Phytophthora spp.* diseases are chemical (soil fumigation) and cultural (improved drainage) [238]. The most effective chemical used in soil fumigation, methyl bromide, was banned in the EU in 2010 and now less effective chemicals are used [332].

Rotating arm samplers have been widely used to detect airborne spores of *Phytophthora spp.* However, studies have primarily focused on *P. capsici* and *P. infestans* [101, 102, 132, 211], rather than species which are known to infect strawberries.

3.3.2 Disease monitoring

At Wilkin & Sons, disease monitoring relied on visual inspection of plants, where symptoms such as white mycelial patches from *P. aphanis* or grey fuzz and lesions from *B. cinerea* were scored for severity. Although essential for guiding fungicide applications, these inspections were labour-intensive and so they were carried out only every other week. Furthermore, visual assessment is inherently subjective, introducing potential bias and variability between assessors. Underscoring the need for more reliable and standardised disease monitoring tools.

Increasingly advanced technology has been employed to enhance strawberry protection. For instance, *Phytophthora spp.* and *B. cinerea* have been identified using lateral flow devices, and LAMP or ELISA-based assays [203, 364, 365]. Moreover, LAMP assays have been utilised to detect fungicide resistance in *B. cinerea* populations to quinone outside inhibitors [159]. Other protection strategies include the deployment of sensor networks for disease mitigation in strawberry farming systems [87, 412] and the application of deep learning models to detect diseases such as grey mould and powdery mildew from plant images [41, 337]. While these technologies show significant potential, they remain in the developmental stages and are not yet widely adopted in commercial agricultural settings.

Airborne monitoring provides an alternative route. Spore samplers combined with microscopy have shown that higher concentrations of *B. cinerea* spores in the air precede outbreaks of grey mould [44, 129]. However, microscopy lacks sensitivity and requires considerable expertise. Coupling airborne capture with molecular approaches could overcome these limitations, offering both faster and more accurate detection.

AirSeq extends this concept by using sequencing to identify airborne pathogens before symptoms appear on crops, providing growers with an early warning system to support more sustainable disease management.

3.3.3 Fungicide efficacy and developing resistance

Fungicides are an integral component of modern agriculture, frequently applied to manage fungal pathogens such as *B. cinerea*, *P. aphanis* and *Phytophthora spp.* [238, 275, 347, 382]. However, their extensive and often indiscriminate use has led to the emergence of fungicide-resistant populations. Resistance develops as a result of strong selection pressures imposed by fungicides, exacerbated by factors such as over application and inadequate rotation of active ingredients [242].

Resistance in *B. cinerea* is particularly well documented, with numerous studies reporting reduced fungicide sensitivity in global populations [15, 232, 382]. Sequencing of resistant isolates from agricultural strawberry systems has revealed specific mutations associated with fungicide resistance (Table 3.1). These findings underscore the role of genetic mutations in driving resistance and highlight the value of molecular approaches in identifying and monitoring resistant strains.

Fungicide resistance has also been observed in *P. aphanis* and *Phytophthora spp.*, though prevalence and severity vary between regions and farming practices [238, 347]. For example, studies comparing *P. aphanis* isolates from organic and conventional farms found greater fungicide sensitivity in organic isolates, reflecting lower chemical use [322]. Such variability highlights the need for site-specific management strategies to mitigate resistance development.

Table 3.1: Fungicide groups, mutation types identified in *B. cinerea*, and references

Fungicide group	**Mutation type and location**	**Reference**
Fenhexamid	Multiple point mutations in erg27	[14, 109, 131]
Succinate dehydrogenase inhibitors & Quinone outside inhibitors	Multiple point mutations in CytB sdhB	[14, 214, 379]
Anilinopyrimidines	Mutations in many mitochondrial function genes, only two mutations detected in field populations	[256]
Phenylpyrroles	Loss-of-function mutations in bos1 only observed in laboratory populations	[110]
Broad/multi-resistance	Rearrangement in the promoter region of mfsM2 and mutations in Mrr1	[204]

Development of fungicide resistance is further complicated by the simultaneous presence of multiple pathogens within a single crop system. For instance, the use of fungicides targeted at *B. cinerea* may inadvertently select for resistance in *P. aphanis* or other pathogens [275]. This phenomenon underscores the interconnected nature of pathogen management and the potential for resistance to emerge as an unintended consequence of targeted treatments.

To address the growing threat of fungicide resistance, organisations such as Fungicide Resistance Action Committee (FRAC) have been established. FRAC provides guidelines for fungicide use, monitors resistance risks, and compiles information on known resistance-associated mutations (`https://www.frac.info`). Such initiatives play a critical role in informing growers and policymakers about sustainable fungicide practices to safeguard crop yields.

3.3.4 Research objectives

The aim of this study was to evaluate the effectiveness of AirSeq in detecting *B. cinerea*, *P. aphanis* and *Phytophthora spp.* in a commercial strawberry farm. The performance of AirSeq was assessed by comparison to traditional disease scoring, the current standard for disease monitoring.

To achieve this, air samples were collected monthly from three locations at Wilkin & Sons Ltd from December 2021 to August 2022. The samples were processed, sequenced, and analysed to identify the presence of airborne pathogens. Additionally, environmental metadata including disease scores, temperature, humidity, and fungicide applications were incorporated to further understand pathogen dynamics.

To further build on this work *P. aphanis* samples were collected from the Strawberry plants and the sequenced material was compared to the *P. aphanis* reference genome and the sequenced air samples.

By addressing these objectives, this research seeks to explore the broader potential of AirSeq as a tool in agricultural pathogen monitoring. The findings have significant

implications for improving early disease detection, enhancing targeted disease management practices, and optimising fungicide application to mitigate the development of fungicide resistance in pathogen populations.

3.4 Methods & Analysis

3.4.1 Field site & sampling

The experimental work in this chapter began with the collection of samples from a commercial farm in Tiptree, Essex, UK. Baseline samples were collected in December 2021 and then monthly air samples were collected from February 2022 – August 2022, using the Coriolis µ (Bertin Instruments). At the June collection the AirPrep Cub (InnovaPrep) and Coriolis Compact (Bertin instruments) were used to collect additional samples. Some samples were collected on my behalf by Dr Heavens. The dates of each collection and the sampler used can be seen in Table 3.4.

Replicate samples (n = 2) were collected monthly from three distinct locations: an open field and two greenhouses (Greenhouse 1 (G1) and Greenhouse 2 (G2)), all of which were used for strawberry cultivation (Figure 3.4).

Each location was planted with a different strawberry variety, categorised as either June-bearing or everbearing. Details of the varieties used and their known susceptibilities are provided in Table 3.2. In G1, the June-bearing varieties were removed at the end of the 2020-21 growing season, and the area remained fallow until March 2022, when it was replanted with the Favori variety. In the field, the Little Scarlet variety was replanted in July following crop turnover.

Table 3.2: Strawberry variety, bearing types, and disease susceptibilities by location

Location	**Variety**	**Bearing Type**	**Disease Susceptibility**	**Source**
Field	Little Scarlet	Everbearing	A cultivated form of *Fragaria virginia*, susceptibility unknown	Wilkin & Sons
G1	Favori	Everbearing	Tolerant to root diseases; strongly resistant to powdery mildew	FlevoBerry
G2	Malling Centenary	June-bearing	Susceptible to crown rot and Botrytis grey mould; intermediate resistance to powdery mildew	NIAB
	Prize	Everbearing	Moderately susceptible to powdery mildew and Botrytis grey mould	Patent [354]

On each collection day, samples were obtained between approximately 10:30 and 14:30, with the order of locations varying between visits. At each location, two consecutive samples were collected using the Coriolis µ air sampler. A sterilised collection cone was filled with 15 ml of sterile nuclease free water and attached to the sampler. The sampler ran for 20 minutes at a flow rate of 300 L/min, a new sterile swan neck was fitted to the

Figure 3.4: Images from the Strawberry Disease sampling location. A) Satellite image showing the three sampling locations, B) Strawberry plants growing in the field where sampling took place C) Coriolis µ sampler in the greenhouse.

machine at each location change. After collection, the cone was removed, and the collection fluid filtered with a syringe through a Swinny filter, containing a 13 mm 0.22 μm PVDF membrane filter (Durapore). The PVDF filter was then removed and stored on dry ice in a 2ml tube for transfer to the lab where it was stored at -80°C until further processing.

3.4.1.1 Collection of *P. aphanis* tissue

In 2023, *P. aphanis* tissue was collected from infected strawberry leaves by Dr Heavens. A scalpel was used to scrape the material into an Eppendorf tube containing DNA/RNA-Shield (Zymo Research). All samples were collected from G2, and one air samples was collected at the same time following the procedure described above. Figure 3.5 shows an infected plant from which some of the material was obtained.

Figure 3.5: Strawberry fruit infected with powdery mildew (*Podosphaera aphanis*).

3.4.2 DNA extraction and isolation

The PVDF membrane filters from December 2021 were processed and sequenced independently as a proof-of-concept that the methodology was sufficient to obtain sequence reads from the pathogens of interest. All other PVDF membrane filters (collected in 2022)

underwent DNA extraction and isolation within 7 days of being brought back from the farm.

To extract the DNA, 125 μl of lysis solution and 250 mg of beads from the DNeasy Power Soil DNA Isolation kit (Qiagen) were added to the 2 ml tube containing the filter or *P. aphanis* spore sample and beat for 20 seconds at speed code 20 (SuperFastPrep-2™, MP Biomedical). The tube was then centrifuged at 13.2 rpm for 1 minute to separate the lysate from the filter and beads.

Next a bead clean up step was performed using a 1:1 ratio of KAPA Pure beads (Roche) and lysate (75 μl of each), the beads were added to the lysate in a 1.5 ml Eppendorf and vortexed. The tube was then left to incubate for 5 minutes at room temperature before being placed on a magnetic particle concentrator to pellet. The supernatant was removed, and the pellet washed twice with 70% EtOH, before being spun down and any excess EtOH removed. 10 μl of sterile nuclease free water was then added to the pellet and vortexed until the pellet was fully suspended. Following an additional 5 minutes room temperature incubation the tube was placed back on the magnetic particle concentrator to pellet and the supernatant DNA was removed and stored in the −20°C freezer.

DNA concentration was quantified using the Qubit fluorometer (Invitrogen, Thermo Fisher Scientific). For each assay, 1 μl of DNA, 1 μl of High Sensitivity (HS) dye, and 198 μl of HS buffer were combined in a Qubit assay tube according to the manufacturer's instructions. Fluorescence was then measured with the Qubit instrument to determine DNA yield. Based on these results, all samples in this study subsequently underwent a WGA step.

Once all the samples had undergone DNA extraction and quantification, WGA was performed in randomised batches. WGA began by adding 3.5 μl of DNA, 0.5 μl of EquiPhi reaction buffer, and 1 μl of Exo primers (Thermo Fisher Scientific) to a PCR tube, followed by incubation for 3 minutes at 95°C. Then, 1.5 μl of reaction buffer, 2 μl of 10 mM dNTPs, 0.2 μl of 100 nM DTT, 10.3 μl of water, and 1 μl of EquiPhi polymerase were added to each sample and incubated for 90 minutes at 45°C, followed by 20 minutes at 80°C. This 45°C incubation step was increased for samples whose yields did not improve after the first round of WGA. After incubation, the samples were cleaned using a 1x KAPA bead clean-up as described above, but this time eluted into 20 μl of water. Following elution, samples were quantified using the Qubit assay.

The next step in WGA was debranching using T7 endonuclease (New England Biolabs). Debranching is commonly required after WGA, as the amplification process generates branched DNA molecules that can obstruct nanopores during sequencing. A total of 17 μl of DNA from the WGA step was combined with 2 μl of reaction buffer and 1 μl of T7 endonuclease in a 1.5 ml Eppendorf tube. Samples were then incubated at 37°C for 30 minutes, followed by an additional 1x bead clean-up. The DNA was eluted into 20 μl of water and quantified using the Qubit assay.

3.4.3 Library preparation & sequencing

All libraries were prepared from WGA samples following standard ONT protocols, using different kits and flow cells depending on the sequencing batch. December 2021 samples were prepared with a native barcoding kit (SQK-NBD110-96) and sequenced on a MinION flow cell. For samples collected between February and August 2022, a rapid barcoding kit

with Q20 chemistry was used (SQK-RBK114-24); these were divided into two pools and sequenced on PromethION flow cells.

Spore samples were processed with a rapid barcoding kit (SQK-RBK114-96) and also sequenced using PromethION flow cells. The air sample collected at the same time as the spore samples was sequenced on a Flongle with the SQK-LSK114 ligation kit.

WGA DNA samples were diluted to 2 ng/μl, and 10 μl of each dilution was used for library preparation by combining with 1 μl of barcode and mixing by pipetting. The tubes were then incubated at 30°C and 80°C for 2 minutes each, followed by placement on ice and centrifugation. The barcoded samples were subsequently pooled, combined with an equal volume of AMPure XP beads, and mixed by flicking before being incubated at room temperature for 5 minutes.

The pooled sample was then centrifuged and placed on a magnetic particle concentrator to pellet the beads, and the supernatant was carefully pipetted off and discarded. Fresh 80% EtOH was used to wash the pellet twice, after which any residual ethanol was removed. The tube was then taken off the magnet, and the pellet was resuspended in elution buffer, followed by a 10-minute incubation at room temperature (elution buffer volumes are provided in Table 3.3). The beads were subsequently pelleted again using the magnetic particle concentrator, and the clear eluate was pipetted off and retained for downstream steps. This eluate was divided into 11 μl and 4 μl aliquots for use in different sequencing runs (see Table 3.3). The two pools were prepared separately.

To sequence the library, 0.5 μl of rapid adaptor was diluted with 1.16 μl of adapter buffer. Once combined, 1 μl of the diluted rapid adaptor was added to the library, mixed by flicking, and incubated at room temperature for 5 minutes. The flow cell was then prepared following the standard ONT procedure, and sequencing was initiated. Each sequencing run lasted for 24 hours, after which the library was recovered, the flow cell flushed, and either a reused or fresh library was loaded. Specific details of the libraries and sequencing run durations are provided in Table 3.3. Library recovery and flow cell flushing were performed by Dr Heavens, following the ONT protocol.

During two of the sequencing runs, a power outage occurred after 24 hours, which prevented further sequencing and resulted in the loss of some sequencing data. In one of the Pool 2 runs, the recovered library exhibited low sequencing efficiency; the run was therefore paused, and an additional 4.5 μl of library was added. This adjustment is indicated by an asterisk in the recovered library cell of Table 3.3.

Reads were basecalled using the HAC model: Guppy v6.2.11 for the 2021 runs and Guppy v6.3.9 for the 2022 runs.

3.4.4 Metadata

Wilkin & Sons Ltd provided metadata, including temperature (recorded every 10 minutes), humidity (recorded every 5 minutes for the greenhouses but not in the field), fungicide application schedules, and fortnightly disease scores for Phytophthora, powdery mildew, and grey mould across all locations (collected between 15/02/2022 - 16/08/2022).

Weekly disease scores were assessed independently by two agronomists using a standardised 0–5 scale, where 0 indicated no disease, 1–2 were not of concern, 3 typically triggered fungicide application, and 4–5 indicated widespread disease requiring immediate intervention. These scores were visualised in R to examine temporal trends.

Table 3.3: Table summarising the different sequencing runs

Samples	**Elution Buffer (μl)**	**Eluate sequenced (μl)**	**Sequence time (hrs)**	**Flow cell used**	**Recovered library**	**Power cut**
Pool 1	15	4	24	Flongle	No	No
		11	24	MinION	Yes	No
		11	24	MinION (reused)	No	Yes
Pool 2	15.5	11	24	MinION	Yes	No
			1	MinION	No*	No
		4.5	24	MinION (reused)	No	Yes

* Recovered library exhibited low sequencing efficiency, run was paused and an additional 4.5 μl of library added.

Active ingredients and targets of the fungicides used were identified from manufacturer data to classify them as broad-spectrum or targeted treatments for *Botrytis* or *Podosphaera*. Additionally, the resistance risk associated with each fungicide's mode of action was evaluated using the FRAC classification system. Two chemical treatments (Rigel WP and Batavia) were excluded from the analysis due to uncertainty about their classification or relevance to fungal disease management.

3.4.5 Bioinformatics

All scripts utilised for the following analysis are available at `https://github.com/Mia-FGB/Strawberry_Analysis`

3.4.5.1 Airborne eDNA sequencing data

Raw read data was transferred from the GridION to the Earlham Institute's High-performance Computer (HPC) for analysis, reads were base-called to high accuracy using Guppy during sequencing and classified into pass and fail based on standard Guppy parameters. Because barcodes were conserved across pools, reads for each barcode were concatenated across sequencing runs, with pass and fail files maintained separately for downstream analysis.

Reads were length filtered, those <300 bp were removed from analysis before alignment, using an Awk script from Dr Martin. The threshold was determined to retain enough reads for future analysis whilst removing those unlikely to be correctly classified.

Reads were aligned to a database of pathogen reference sequences based on PHI-base [372] and described in [124] with the addition of the *Podosphera aphanis* referece genome (provided by Thomas Heaven at NIAB) [148] using *minimap2* (v2.24) [217]. *Minimap2* was run with the standard parameters and the `-x map-ont` flags. Alignments were output in PAF format and filtered to remove those with a mapping quality (MQ) below 5 but retain those with a MQ of 0, using a custom Python script, `dict_paf_parse.py`. The MQ threshold was adopted from a similar metagenomic study [10].

Alignments may map to multiple reference genomes within the database, in this scenario

they are assigned a MQ of 0. In such cases, additional analysis was performed to assess the best alignment or whether the alignment should be removed from further analysis. The logic for processing these reads is shown in Figure 3.6. Reads with identical MQ values and identical reference genomes were retained. Reads with identical MQ values but multiple reference genomes were excluded from downstream analysis. Where multiple MQ values and multiple reference genomes were present, only the alignment with the highest MQ score was retained. In cases where an alignment was retained, the taxa count was increased by +1 for the identified species. The output of the script is a taxa count table.

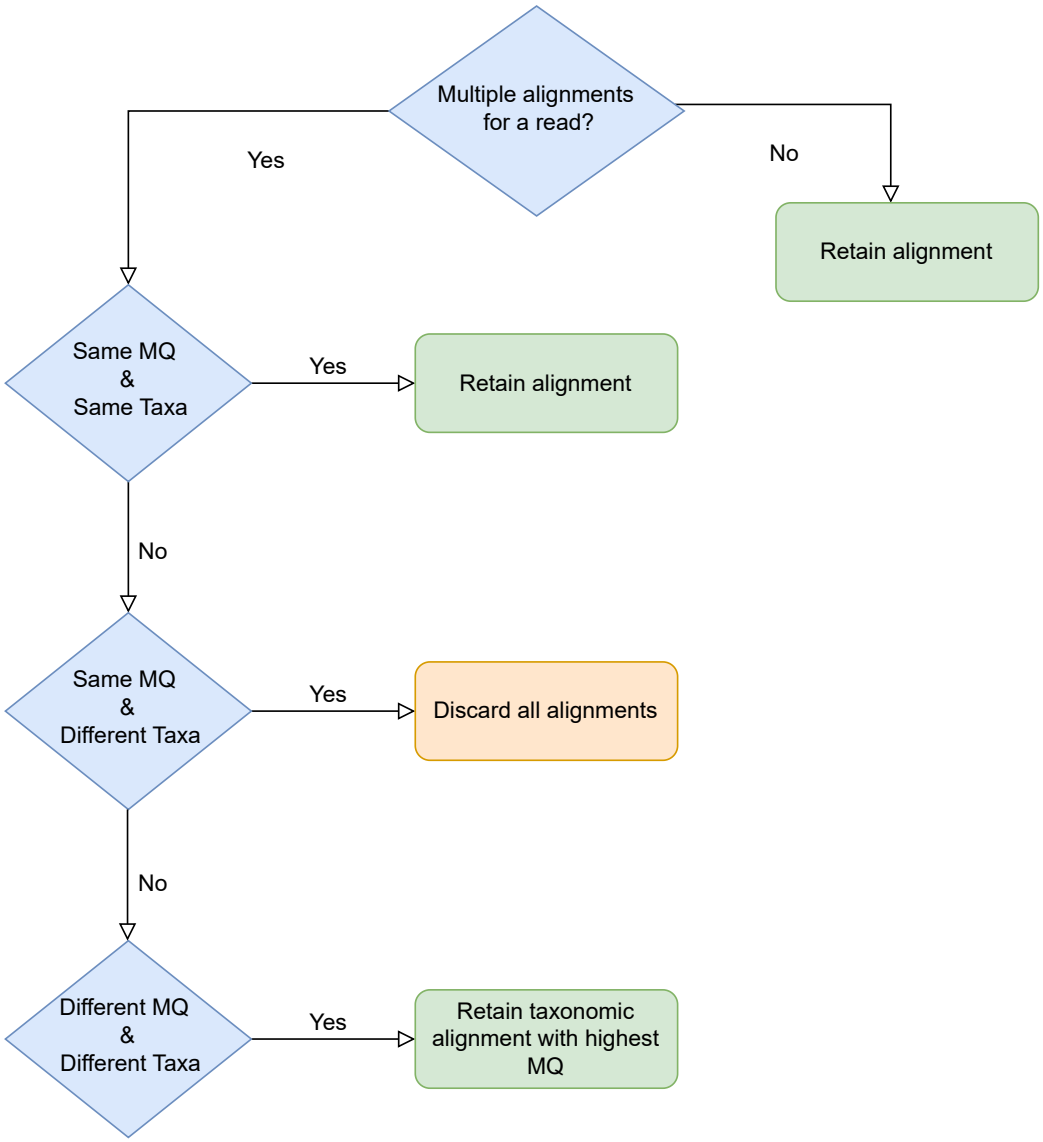


Figure 3.6: Decision tree used to resolve reads with multiple alignments. Alignments were evaluated based on whether they had identical or differing MQ scores and reference genome assignments.

Analyses on the taxa count table were performed in R using packages from the tidyverse collection, including `dplyr` for data manipulation and `ggplot2` for visualisation. Taxonomic count data were imported and normalised to hits per 100k reads per barcode, correcting for variation in sequencing depth between samples, with the script `PHIbase_analysis.R`. Following normalisation, heatmaps were generated to visualise species abundance across

all samples, grouped either by individual sample or by collection month. Abundance values were log-transformed (log10) to facilitate visualisation across orders of magnitude, and only taxa with a read count greater than 10 that were also detected in at least 10 samples were included.

The same script (`PHIbase_analysis.R`) was used to further examine the most abundant taxa, normalised read counts were aggregated across all samples. The top 20 species were selected based on total abundance, and horizontal bar plots were produced to show their mean normalised read count across samples, with standard error bars indicating variability. A separate subset containing the top 10 most abundant species was used to investigate temporal and spatial trends in pathogen presence. Stacked bar plots were generated to illustrate top 10 species abundance grouped by sampling month and by collection location.

For further analysis into the three scored diseases, reads assigned to the genera *Botrytis*, *Podosphaera*, and *Phytophthora* were filtered and summarised by sample, location, and date. For each genus, the mean, minimum, and maximum log-transformed normalised abundance values were calculated and visualised as line plots, faceted by location and overlaid with error bars to reflect the two repeat samples, using the script `Tiptree_all_data_analysis.R`.

3.4.5.2 Metadata

Environmental data analysis was conducted in R using the `ggplot2` and `lubridate` packages. The temperature and humidity data were processed to calculate daily averages, which were then visualised as line graphs with standard error bars.

Disease scoring data were cleaned and plotted by location and sampling date to visualise symptoms over time with the script `Tiptree_all_data_analysis.R`. Similarly, environmental data were processed to calculate daily mean temperature and humidity values, with standard deviations used to generate shaded error bands. Fungicide application records were also processed and joined with additional data describing their target pathogens. Application dates were added as vertical lines to all relevant plots. Finally, for each focal genus, pathogen abundance, disease severity, and environmental conditions were assembled into combined, panelled plots to enable integrated interpretation of biological and environmental patterns.

The fungicide data was imported into R and plotted as bar charts using the `fungicide_data.R` script.

3.4.5.3 *P. aphanis* tissue

Sequence data generated from the powdery mildew and air samples were rebasecalled with Dorado using the Super Accuracy model (v4.3.0).

Four independent spore samples were processed. To assess the proportion of contamination in the read data, the Metagenomic Alignment and Reporting for Monitoring of Threats (MARMoT) pipeline was used to map the reads with *minimap2* against a curated pathogen database (see Chapter 7 for a detailed description of the pipeline).

Genome assemblies were generated for each spore sample with Flye (v2.8.1) using the `nano-raw` flag. The AirSeq data was then aligned to each assembly, along with the *P. aphanis* reference genome (accession: GCA_022627015.2), using *minimap2* with the flags `-N 100 -c -x map-ont` to produce pairwise alignment format (PAF) outputs.

The PAF files were subsequently imported into the Jupyter notebook `air_mildew_individual.ipynb`, where alignments were filtered by identity ($> 80\%$) and mapping length (> 150bp). A bar chart was then generated to visualise the number of alignments for each assembled sample and the reference genome.

3.5 Results

The results of this study include the sample information, environmental conditions, disease scores, airborne sequencing results and fungicide application patterns observed. First, we describe the sample characteristics, including DNA yield and the effects of quality filtering. This is followed by an analysis of the pathogen diversity present in the samples. The airborne pathogen sequence data for the three pathogens of interest in this study are then analysed in conjunction with environmental variables, including temperature, humidity, and disease progression over time. Additionally, fungicide efficacy is assessed based on the observed pathogen dynamics. Finally, a comparison of plant collected *P. aphanis* spores to the airborne sequencing data is conducted, showing a higher alignment between the plant collected spores and the airborne samples than with the *P. aphanis* reference genome.

3.5.1 Sample information

The sample metadata are summarised in Table 3.4, including sample ID, date of collection, air sampler used, collection location, and whether an insect was present on the filter after sampling. DNA concentration, is reported both after the initial extraction (DNA ext.) and following WGA and T7 debranching (DNA WGA). Sequencing metrics include the number of pass, fail and total reads, as well as the N50 value of the pass reads (N50). The table also reports the percentage of pass reads longer than 300 bp (>300 bp %) and the proportion of reads removed due to alignment to multiple reference genomes (Filtered %).

Table 3.4: Metadata for strawberry greenhouse and field samples, including collection date, sampler type, location, insect presence, DNA yield, and counts of pass and fail sequence reads.

Sample ID	Date	Sampler	Location	Insect	DNA ext.	DNA WGA	Pass	Fail	Total	N50	>300 bp %	Filtered %
1.1	10/02/2022	Coriolis µ	Field		too low	25	53476	6424	59900	2172	96.47	1.14
1.2	10/02/2022	Coriolis µ	Field		0.254	32.2	65932	5689	71621	1848	95.94	1.04
1.3	10/02/2022	Coriolis µ	Greenhouse 1		1.8	0	3117	1327	4444	1160	90.28	1.09
1.4	10/02/2022	Coriolis µ	Greenhouse 1		1.62	0.213	2027	1695	3722	1053	85.74	2.02
1.5	10/02/2022	Coriolis µ	Greenhouse 2		0.97	88	44917	6474	51391	3176	96.58	2.11
1.6	10/02/2022	Coriolis µ	Greenhouse 2		0.485	115	3364	694	4058	4128	93.88	3.42
2.1	09/03/2022	Coriolis µ	Greenhouse 1		too low	2.41	64003	8723	72726	816	93.01	0.49
2.2	09/03/2022	Coriolis µ	Greenhouse 1		0.29	0	1342	1952	3294	464	66.02	0.82
2.3	09/03/2022	Coriolis µ	Greenhouse 2		too low	0	125	395	520	2745	91.2	2.4
2.4	09/03/2022	Coriolis µ	Greenhouse 2		0.16	0.709	2605	1159	3764	522	89.9	0.54
2.5	09/03/2022	Coriolis µ	Field		0.395	110	66259	7963	74222	3294	97.49	1.91
2.6	09/03/2022	Coriolis µ	Field		too low	38.6	54657	5022	59679	2331	97.43	0.15
3.1	05/04/2022	Coriolis µ	Field		0.843	124	32225	5876	38101	4770	97.4	4.25
3.2	05/04/2022	Coriolis µ	Field		1.86	85.7	26314	4265	30579	6536	97.16	8.35
3.3	05/04/2022	Coriolis µ	Greenhouse 2		0.901	35.4	47590	6416	54006	4193	97.54	3.4
3.4	05/04/2022	Coriolis µ	Greenhouse 2		0.379	7.01	46029	5227	51256	2281	96.66	1.11
3.5	05/04/2022	Coriolis µ	Greenhouse 1		0.317	22.9	83310	16300	99610	1881	97.08	0.97
3.6	05/04/2022	Coriolis µ	Greenhouse 1		0.289	3.97	23663	6059	29722	843	92.67	0.16
4.1	11/05/2022	Coriolis µ	Field		1.91	17.6	39514	5531	45045	5889	97.9	3.73
4.2	11/05/2022	Coriolis µ	Field		2.09	17.8	44428	4652	49080	1450	96.36	1.04
4.3	11/05/2022	Coriolis µ	Greenhouse 1		0.817	0	3460	1165	4625	667	92.98	0.52
4.4	11/05/2022	Coriolis µ	Greenhouse 1		0.322	0	46029	789	46818	1881	1.17	0
4.5	11/05/2022	Coriolis µ	Greenhouse 2		0.5	2.19	35971	6248	42219	4168	97.04	3.12
4.6	11/05/2022	Coriolis µ	Greenhouse 2		0.352	0	1372	679	2051	617	88.05	0.95
5.1	07/06/2022	Coriolis µ	Field		0.61	26.2	62203	7143	69346	4651	97.85	0.5
5.2	07/06/2022	Coriolis µ	Field	yes	7	110	77162	5367	82529	4303	97.12	0.81
5.3	07/06/2022	Coriolis µ	Greenhouse 1		2.48	131	77457	8969	86426	6185	97.21	0.76
5.4	07/06/2022	Coriolis µ	Greenhouse 1		1.79	136	130799	13706	144505	3672	97.11	0.63
5.5	07/06/2022	Coriolis µ	Greenhouse 2		2.97	144	53480	6626	60106	3251	96.78	0.96
5.6	07/06/2022	Coriolis µ	Greenhouse 2		1.39	73	57846	4650	62496	2095	96.57	0.81

Sample ID	Date	Sampler	Location	Insect	DNA yield (ng/μl)	DNA yield after WGA (ng/μl)	Pass Reads	Fail Reads	Total Reads	N50	>300 bp Reads (%)	Filtered Reads (%)
5.7	07/06/2022	Coriolis Compact	Field		0.19	0	2011	1348	3359	967	87.47	1.39
5.9	07/06/2022	Coriolis Compact	Greenhouse 1		0.29	5.3	53747	3729	57476	3175	96.97	0.75
5.11	07/06/2022	Coriolis Compact	Greenhouse 2		0.4	30.2	81949	7346	89295	3287	97.23	0.98
5.13	07/06/2022	InnovaPrep Cub	Field	yes	4.65	259	160896	14195	175091	2519	97.23	0.64
5.14	07/06/2022	InnovaPrep Cub	Greenhouse 1		1.92	0.14	7315	2317	9632	883	92.4	1.38
5.15	07/06/2022	InnovaPrep Cub	Greenhouse 2		2.29	94.5	39121	3937	43058	2723	97.32	0.93
6.1	13/07/2022	Coriolis µ	Field		0.984	36.4	66531	7561	74092	5381	97.14	0.98
6.2	13/07/2022	Coriolis µ	Field	yes	13.8	148	78166	6750	84916	3178	97.52	1.31
6.3	13/07/2022	Coriolis µ	Greenhouse 1		0.616	52	71817	7771	79588	1853	97.18	0.93
6.4	13/07/2022	Coriolis µ	Greenhouse 1	yes	4.29	187	35297	3540	38837	3077	97.17	0.8
6.5	13/07/2022	Coriolis µ	Greenhouse 2		0.322	0	3278	2290	5568	1999	83.83	0.61
6.6	13/07/2022	Coriolis µ	Greenhouse 2		0.486	6.96	37765	6291	44056	1548	93.98	0.65
7.1	11/08/2022	Coriolis µ	Field		0.908	117	56430	6884	63314	3591	96.56	1.53
7.2	11/08/2022	Coriolis µ	Field		0.775	25.5	51153	5464	56617	3857	95.93	1.54
7.3	11/08/2022	Coriolis µ	Greenhouse 1		0.446	12.4	31807	4070	35877	1631	96.8	1.48
7.4	11/08/2022	Coriolis µ	Greenhouse 1		0.428	68.5	34612	3159	37771	5013	97.42	1.47
7.5	11/08/2022	Coriolis µ	Greenhouse 2		0.383	55.5	46196	6293	52489	4140	97.18	1.65
7.6	11/08/2022	Coriolis µ	Greenhouse 2		0.588	100	59361	5058	64419	2087	97.29	1.31

3.5.1.1 DNA yield

A total of 48 samples were collected, with an average DNA yield of 1.39 ng/μl (range: 0 - 13.8 ng/μl). Four samples contained insects in the collection liquid, which inflated the DNA yields; when these results are excluded, the average DNA yield dropped to 0.84 ng/μl (range: 0 - 2.97 ng/μl). Due to low DNA yields, all samples underwent WGA to increase the amount of DNA for sequencing. If the DNA yield after initial amplification was <1 ng/μl, WGA was repeated and the replicate with the highest DNA yield was used for sequencing. This resulted in an average DNA yield of 53.07 ng/μl (range: 0 - 259 ng/μl) across all samples, and an average of 41.89 ng/μl (range: 0 - 144 ng/μl) in insect free samples.

The DNA yields from initial extraction and after WGA are shown as boxplots in Figure 3.7, grouped by the month of collection and location of sampling. WGA greatly increased the DNA yield across all samples compared to the initial extraction. A difference in DNA yield is clear between samples collected in different months, and appears more substantial after WGA (Fig 3.7a), though the variance within each month is also considerable. DNA yield was more consistent across samples collected from different locations (Fig 3.7b).

However, samples with extremely low DNA levels (March and August) did not increase as much after WGA, likely due to the limited DNA in the initial sample, which left insufficient material for amplification. Although, samples from August which had negligible DNA post extraction had yields increase to >50 ng/μl after WGA.

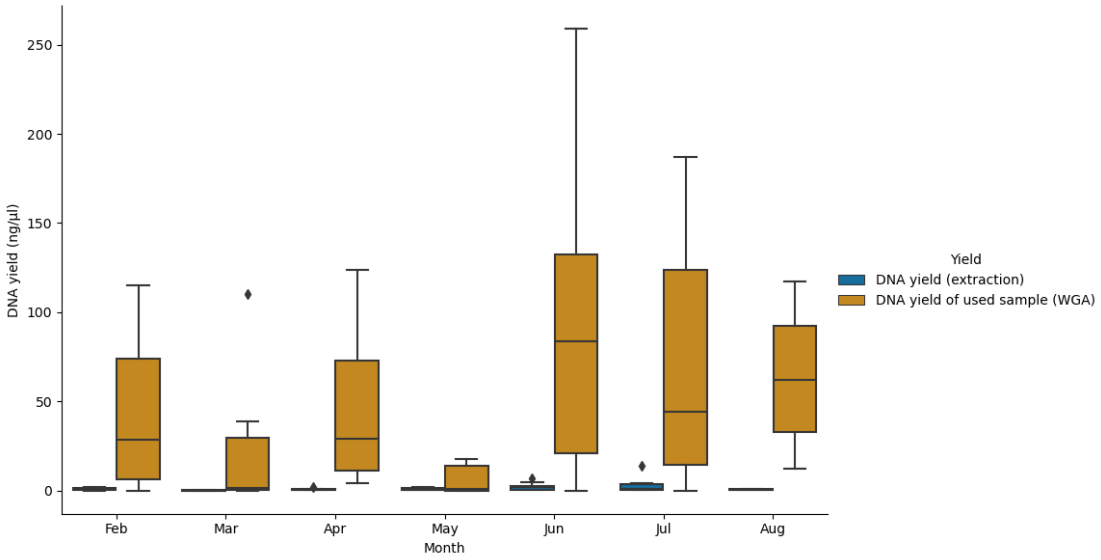
Given the considerable variation in DNA yields across samples, all samples were diluted to 2 ng/μl (where possible) before library preparation to minimise the impact of these variations on sequencing. Although this is below the amount recommended by ONT for library preparation, previous work has demonstrated that it is sufficient for successful sequencing [149].

3.5.1.2 Quality filtering sequence data

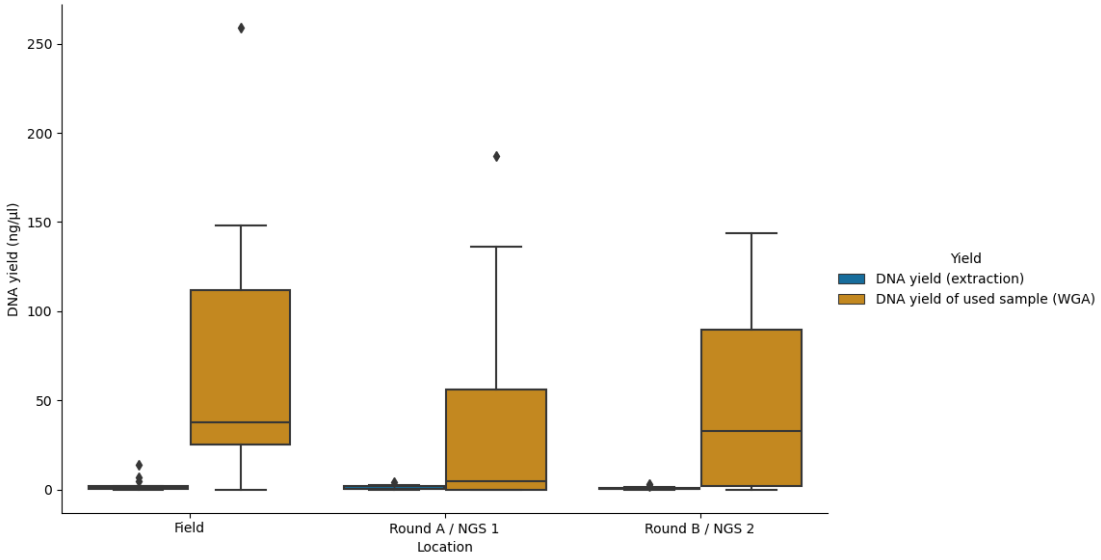
Following alignment with *minimap2* reads were filtered based on their MQ, as described in section 3.4.5. Reads that mapped to multiple taxa with the same MQ were excluded from the analysis, which accounted for an average of 1.45% of the total reads per sample in this analysis (range: 0% – 8.35%). Given that this represents a small proportion of the reads, it is unlikely to have impacted the results, and this filtering approach was deemed sufficient. If a larger proportion of reads had been lost during filtering, it might have been necessary to implement a LCA approach, as seen in other metagenomic studies [10, 70, 72].

3.5.2 Pathogen presence in airborne data

Alignment results from mapping the airborne sequence data to a curated reference database of pathogen genomes (based on PHI-base; see Section 3.4.5) were used to generate heatmaps illustrating changes in pathogen prevalence over time (Figure 3.8). The overall top 20 species have been plotted as a bar chart (Figure 3.9), and the top 10 species have been plotted as stacked bar charts grouped by month and location (Figure 3.10).



(a) Grouped by collection month



(b) Grouped by collection location

Figure 3.7: Box plots showing the DNA Yield (ng/μl) of the samples before and after WGA, grouped by collection month or location.

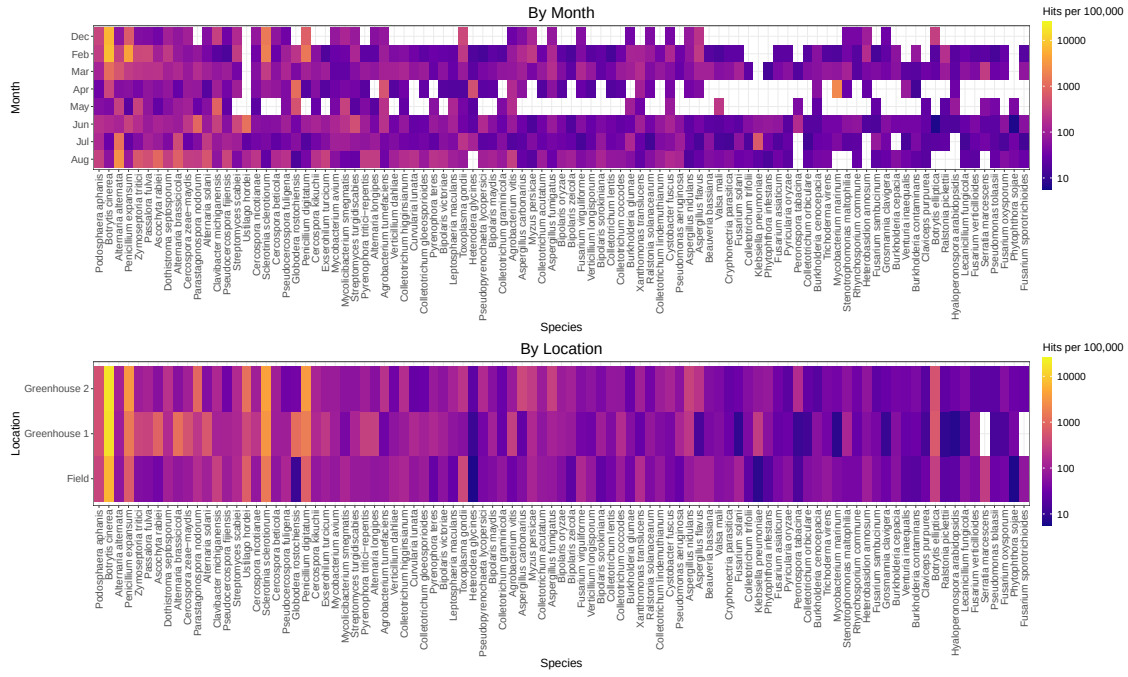


Figure 3.8: Heatmaps showing the hits per 100k reads of pathogens from PHIbase, grouped by month or location of collection. The data was filtered to only contain species with $\geq$ 10 aligned reads and that are present in $\geq$ 10 different samples.

3.5.2.1 Comparison of PHI-base pathogens in the dataset

The heatmaps represents log-transformed hits per 100k reads (Figure 3.8), with the colour scale ranging from black (low abundance) through red and orange to light yellow (high abundance); white indicates absence. Two heatmaps have been created, one where samples collected in the same location have been grouped together (Figure 3.8a) and another where they are grouped by the collection month (Figure 3.8b).

Tables 3.5 and 3.6 show the number of unique species observed by location and by month, respectively. Each table reports raw species counts, counts filtered to include only species with at least 10 total reads, and counts of species present with at least 10 reads in a minimum of 10 samples. As replicate samples were collected each month, a species detected in all three locations during three sampling months with at least 10 reads per sample would be retained, as it would appear in 12 samples. These filtering criteria are consistent with those applied to the heatmap data, which includes the same set of 95 species.

In both heatmaps it is clear that pathogens of strawberries are the most abundant (*B. cinerea* and *P. aphanis*), although the abundance does change over the months. The next most abundant pathogen is *A. alternata* which is known to be present at high levels in many air samples [287] and has been shown to infect Strawberries [247, 356].

Spatial Heatmap - When grouped by collection location (Figure 3.8a), both commonalities and differences in pathogen profiles become apparent. G1 shows the most distinct pattern, with higher abundances of fewer pathogens compared to G2 and the field (Figure 3.8b). As shown in Table 3.5, all three locations initially contain similar numbers of unique species (260 to 268), indicating comparable baseline diversity. However, filtering out species with fewer than 10 total reads disproportionately reduces the number of species detected in the greenhouses (G1: 143, G2: 150) compared to the field (190), suggesting the

Table 3.5: Number of unique species by location and filtering level.

Location	Number of species	Number of species with ≥10 reads	Number of species with ≥10 reads in ≥10 samples
Field	268	190	95
G1	260	143	93
G2	265	150	95

Table 3.6: Number of unique species by month and filtering level.

Month	Number of species	Number of species with ≥10 reads	Number of species with ≥10 reads in ≥10 samples
December	437	93	83
February	649	176	134
March	543	142	103
April	461	72	67
May	513	50	47
June	697	378	265
July	661	264	201
August	583	232	213

field harbours a greater number of low-abundance taxa. Applying a further filter to retain only species present at this level in at least 10 samples results in all locations converging to a similar species count (93 to 95). This stronger filter may favour the greenhouses, where species are more likely to be present in both locations, while under-representing the field. Although this level of filtering may exclude ecologically relevant low-abundance species, it was necessary to construct a readable heatmap, as visualisation becomes impractical when over 100 species are included.

B. cinerea, P. aphanis, and *Sclerotinia sclerotiorum* were abundant across all three locations. Several pathogens were more prevalent in G1 compared to the other sites, including *Penicillium expansum, Zymoseptoria tritici, Passalora fulva, Ascochyta rabiei, Alternaria brassicicola*, and *Globodera rostochiensis*. Conversely, two species (*Serratia marcescens* and *Fusarium sporotrichioides*) were detected at low abundance in the other locations but were absent from G1. *S. marcescens* is a bacterial pathogen of humans and *F. sporotrichioides* is a fungal plant pathogen, both are known to be present in a range of environments including soil and plant surfaces, so it is reasonable for them to be identified at some of the locations [5, 257].

There were fewer pathogens uniquely abundant in G2, with *Parastagonospora nodorum* showing the most distinct increase in abundance at this site. This is a fungal pathogen of wheat so may have come from a nearby farm as there was no wheat grown on the site.

Samples collected in the field displayed profiles broadly similar to those from G2. However, *Toxoplasma gondii* and *Serratia marcescens* were detected at slightly higher levels in the field, though still at relatively low abundance (approximately 100 hits per 100k reads).

Temporal Heatmap - Grouping by the month of collection (Figure 3.8b) reveals several trends in pathogen abundance across different months, which are more distinct than between locations. In December 2021 and February 2022, the most abundant pathogens

were *B. cinerea*, *Penicillium digitatum*, *Penicillium expansum*, and *Sclerotinia sclerotiorum*, with these species present at much higher levels compared to others. These pathogens were less prevalent in the subsequent months of 2022. In March, the abundance of these same pathogens decreased, although *B. cinerea* remained relatively prominent.

In April, the most abundant species shifted, with *Agrobacterium tumefaciens, Globodera rostochiensis,* and *Mycobacterium marinum* showing higher levels. May saw a marked decrease in pathogen levels across the board, potentially due to the lower DNA yield in samples collected that month (as shown in Figure 3.7).

In June, a variety of pathogens were present in the airborne microbiome, but at relatively low abundances. Notably, *Parastagonospora nodorum* and *Ustilago hordei* were present at higher levels than other species. In July, *Alternaria alternata*, which causes black leaf spot in strawberries, was observed at slightly elevated levels compared to March. Additionally, *Klebsiella pneumoniae* also appeared at relatively high levels. In August, several *Alternaria spp.*, including *A. alternata, A. brassicola*, and *A. solani*, were found in higher abundances.

Decreases in pathogen abundance over time may also reflect increases in airborne pollen, since abundances are expressed as relative rather than absolute values. Influxes of other taxa can therefore reduce the apparent abundance of pathogens.

3.5.2.2 Most abundant pathogens in the dataset

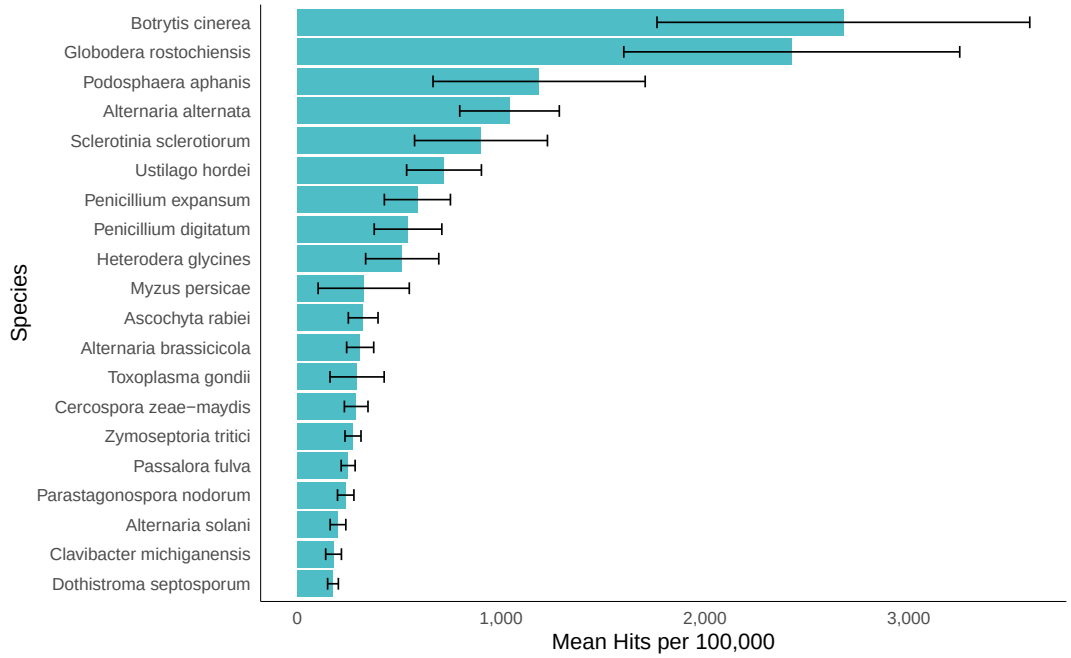


Figure 3.9: Mean abundance of the top 20 pathogen species across all samples. Bars represent the mean normalised read count (hits per 100k reads) for each species. Error bars indicate the standard error of the mean across samples.

The top 20 most abundant pathogen species were identified based on normalised read counts and are presented with associated variability across samples (Figure 3.9). *B. cinerea* emerged as the most dominant species closely followed by *Globodera rostochiensis* and *P. aphanis* is the third most abundant. The top three species, exhibit large standard errors, indicating substantial variability in abundance across the dataset. Other frequently de-

Table 3.7: Top 20 most abundant pathogen species across all samples, including their common names, taxonomic kingdoms, and typical hosts.

Scientific Name	**Common Name**	**Kingdom**	**Common Host**
Botrytis cinerea	Grey mould	Fungi	Strawberry, grapevine
Globodera rostochiensis	Golden potato cyst nematode	Animalia	Potato
Podosphaera aphanis	Powdery mildew	Fungi	Strawberry
Alternaria alternata	Leaf spot fungus	Fungi	Multiple plants
Sclerotinia sclerotiorum	White mould	Fungi	Multiple plants
Ustilago hordei	Covered smut	Fungi	Barley
Penicillium expansum	Blue mould	Fungi	Apple
Penicillium digitatum	Green mould	Fungi	Citrus fruits
Heterodera glycines	Soybean cyst nematode	Animalia	Soybean
Myzus persicae	Green peach aphid	Animalia	Peach, potato
Ascochyta rabiei	Chickpea blight	Fungi	Chickpea
Alternaria brassicicola	Black spot	Fungi	Cabbage, broccoli
Toxoplasma gondii	Toxoplasmosis agent	Protista	Cats, humans
Cercospora zeae-maydis	Grey leaf spot	Fungi	Maize
Zymoseptoria tritici	Septoria tritici blotch	Fungi	Wheat
Passalora fulva	Leaf mould	Fungi	Tomato
Parastagonospora nodorum	Septoria nodorum blotch	Fungi	Wheat
Alternaria solani	Early blight	Fungi	Tomato, potato
Clavibacter michiganensis	Bacterial canker	Bacteria	Tomato
Dothistroma septosporum	Red band needle blight	Fungi	Pine

tected species included *Alternaria alternata*, *Sclerotinia sclerotiorum* and *Ustilago hordei*, though at lower levels. The distribution reflects a skewed pattern of abundance, with a few species accounting for the majority of reads, while the remainder show progressively lower detection levels.

The common names, taxonomic kingdoms, and typical hosts of the top 20 most abundant pathogen species are provided in Table 3.7. The majority of these species are fungi (15 out of 20), while three are animals, and one species each belongs to the kingdoms Bacteria and Protista.

Figure 3.10 shows the top 10 most abundant pathogens across locations and months, revealing clear temporal variation in pathogen composition but minimal spatial differences. While the identities and relative abundances of dominant pathogens shift markedly across months, the three locations share broadly similar profiles within each time point. These graphs are plotted as the average hits per 100k for the duplicate samples collected in the

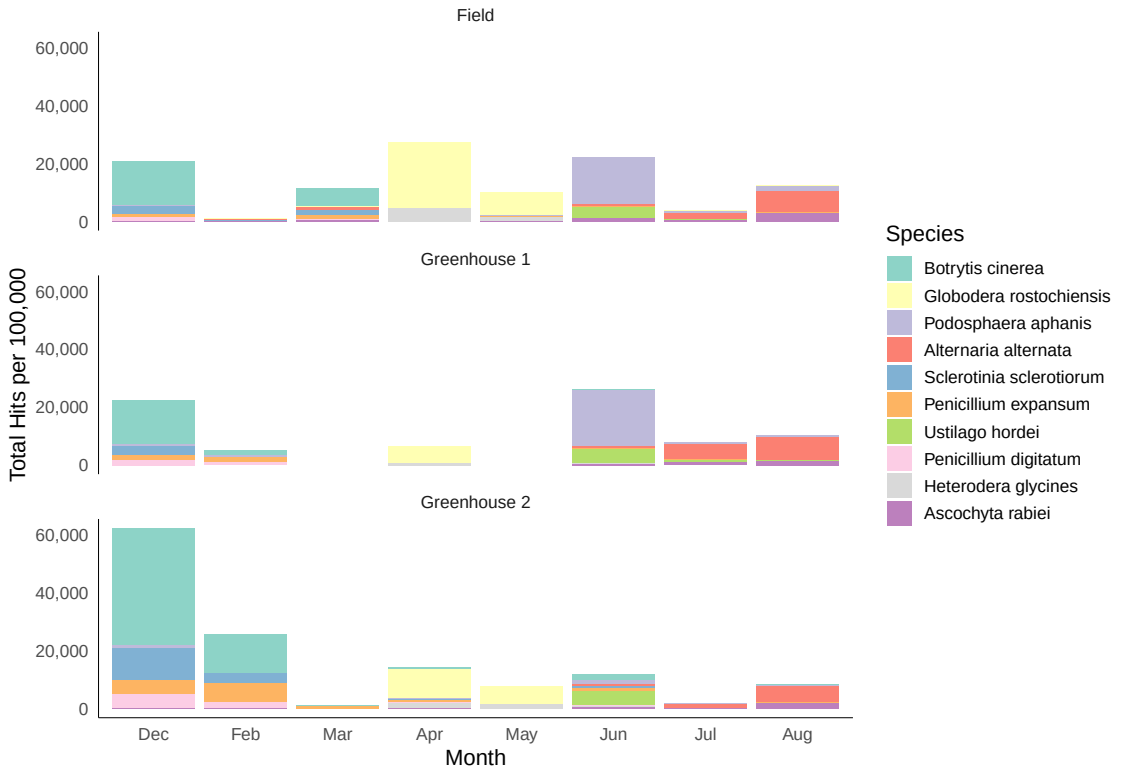


Figure 3.10: Stacked bar plots showing the average distribution of the top 10 most abundant pathogen species across collection locations and sampling months. Species were selected based on the highest total normalised read counts (hits per 100k) across the full dataset. Each bar represents the combined abundance of all top 10 species at a given location and time point, with individual species coloured distinctly.

same location each month.

December was dominated by *B. cinerea* in all locations, particularly in G2, which reached nearly 60,000 hits per 100k compared to ∼20,000 in the field and G1. *S. sclerotiorum* and *Penicillium spp.* were also abundant in all the December samples. This high proportion of *B. cinerea* in December, compared to very low levels in the rest of the year explains the large error bar seen on the top 20 bar chart (Figure 3.9).

In February, *B. cinerea* remained highly abundant in G2, albeit at one-third of December levels. G1 and the field showed minimal detection of top 10 pathogens during this month. In March, pathogen presence was low in the greenhouses but moderate in the field, with *B. cinerea* still present alongside low levels of *A. alternata*.

A major shift occurred in April and May, marked by a predominance of *G. rostochiensis* across all locations, with higher levels in April. *Heterodera glycines* was also detected in low abundance in many of the samples collected at this time. Interestingly, no top 10 pathogens were detected in G1 in May.

In June, the dominant pathogen in the field and G1 was *P. aphanis*, accompanied by low levels of *A. alternata* and *U. hordei*. G2 showed a different profile with predominantly *U. hordei*, though many pathogens were present at relatively low levels.

July and August exhibited another distinct profile, with *A. alternata* emerging as the predominant species in all locations. Overall abundance was low in July, especially in the field and G2. In August, the profiles became more uniform with *Ascochyta rabiei* appearing in all locations and *P. aphanis* only seen in the field.

3.5.3 Pathogen trends, environmental conditions, and fungicide usage

The detected abundance, disease scores, and environmental metadata across the three study locations are shown in Figure 3.11 (*Botrytis*), Figure 3.12 (*Podosphaera*), and Figure 3.13 (*Phytophthora*).

Across all locations, average daily temperature rose from below 10°C in December to around 20°C in September. In the greenhouses, relative humidity remained close to 90% from December to June before declining to 40–60%, where it stabilised until September.

Fungicides were applied between March and August, targeting *Botrytis*, *Podosphaera*, or both diseases. In the field, fewer applications were made within the narrower window of April to May, whereas in both greenhouses fungicides were applied more intensively. Application dates for each location are indicated by vertical lines in Figures 3.11, 3.12, and 3.13.

3.5.3.1 *Botrytis* - grey mould

The *Botrytis* data can be seen in Figure 3.11, the highest airborne *Botrytis* levels (hits per 100k reads) were recorded in December across all three locations, reaching a maximum of over 10,000 hits per 100k. In February, levels remained relatively high, though a larger decrease was observed in the field. By March, *Botrytis* levels in the field rebounded to December levels, while levels in the greenhouses continued to decline. The lowest airborne *Botrytis* levels were recorded in April (field) and May (greenhouses). From May to September, *Botrytis* levels gradually increased in the field, reaching levels comparable to February. A similar trend was observed in the greenhouses, although *Botrytis* peaked earlier in June in G2, while G1 remained more stable.

The disease score data is initially 0 in February, across all locations. Disease severity steadily increased to 2 in the greenhouses by April, before dropping back to 0. For the remainder of the season, the disease score fluctuates between 0 and 2 in both greenhouses. In the field, the disease score remained at 3 from April to July, then declined to 1, before increasing to 2 at the end of the season.

3.5.3.2 *Podosphaera* - powdery mildew

The airborne sequencing data and disease scores for *Podosphaera* can be seen in Figure 3.12.

Airborne *Podosphaera* levels in the field are relatively stable from December to March, but they decrease in April. Following this from April until June airborne *Podopshaera* levels in the field increase and remain high until the end of sampling in August. This decrease in *Podosphaera* levels around April are also seen in the greenhouses (G1: April, G2: May). Aside from these two sharp decreases, *Podosphaera* levels in the airborne sequencing data remain relatively stable for the rest of the season. *Podosphaera* levels are the highest in June reaching a max of 1,000 hits per 100k in G1.

At all three locations, initial disease scores are 0. Disease appears earlier in the greenhouses, with scores reaching 1–2 in March. There is a brief drop in both greenhouses to 0 in mid-April, but disease levels rise again through the rest of the season. By August, G1 reaches a score of 2–3, while G2 is is stable at a score of 3. Interestingly, in the final score of the season G2 is recorded as 0. In the field, disease is first observed in May with a score of 2. It then continues to increase steadily, reaching a peak score of 4 in August.

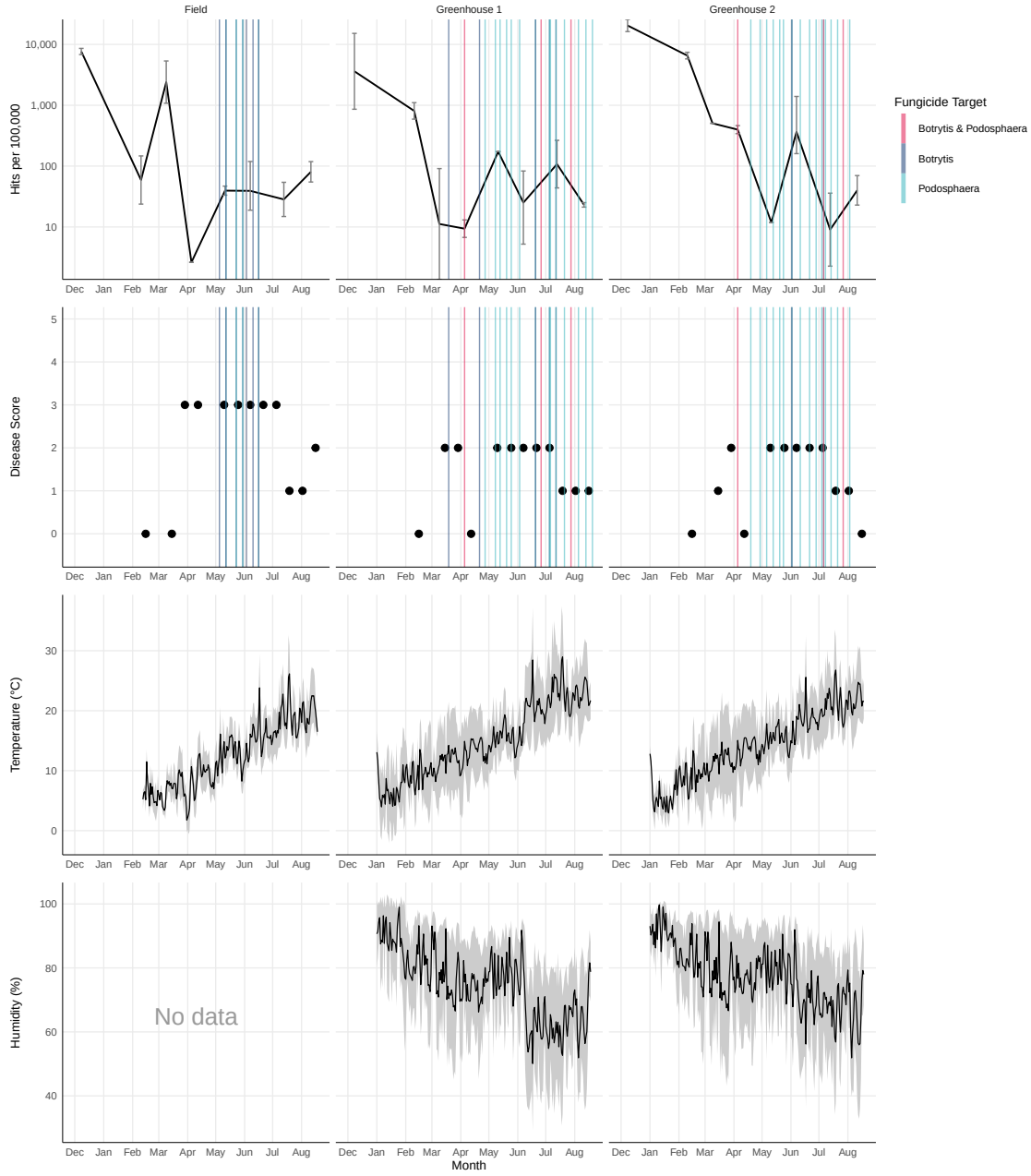


Figure 3.11: Visualisation of AirSeq read abundance, disease severity, fungicide application, and environmental data for *Botrytis*. The top row shows line graphs of the number of reads per 100k mapped to the genus of interest. Points represent the mean of two replicates, with error bars indicating individual sample values. The second row presents scatter plots of fortnightly disease scores (0–5 scale). Coloured vertical lines mark the dates of fungicide application, with colours representing the targeted fungal group. The bottom two rows show recorded temperature (°C) and relative humidity (%) data at each location. The black line represents the daily mean, and the shaded grey area indicates the standard deviation. Note that no humidity data are available for the field.

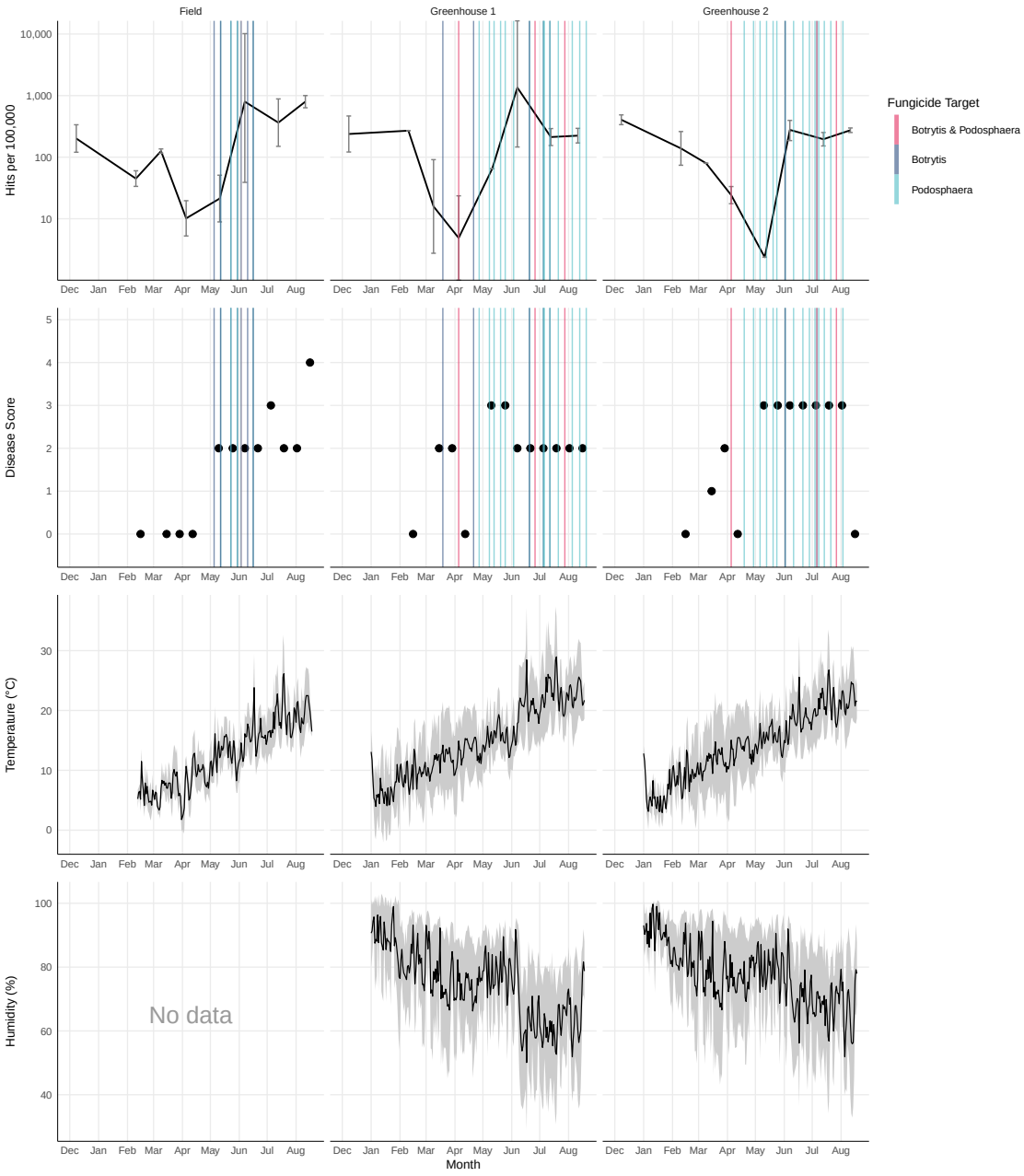


Figure 3.12: Same data structure as Figure 3.11, but for *Podosphaera*.

3.5.3.3 *Phytophthora* - crown & leather rot

The airborne sequencing data and disease scores for *Phytophthora* can be seen in Figure 3.13.

Airborne levels of *Phytophthora* are much lower than those of other pathogens, with a maximum of 100 hits per 100k. In the field, *Phytophthora* levels remain relatively stable from December to March but drop to zero in April (notably with a large error margin). Levels then rebound in May, returning to values similar to those observed before April, and peak in July before declining again to May–June levels.

A similar July peak is observed in both greenhouses. In G1, *Phytophthora* levels are lowest in December and March. In contrast, G2 shows a decline to zero hits in both February and March, followed by a sharp increase in July, mirroring the trend seen in the field.

Phytophthora also has the lowest disease scores of all the pathogens. Disease is not detected in the field throughout the season, except for a one-off spike to a score of 2 in late April. In G1, the disease score peaks at 3 in February, then declines to 1 in March and drops to 0 from April to August, with the exception of June, when it briefly rises to 2. G2 shows no *Phytophthora* symptoms at any point, maintaining a disease score of 0 throughout the season.

3.5.3.4 Fungicide protection and FRAC resistance risk

In addition to the fungicide application timelines shown in Figures 3.11, 3.12, and 3.13. Figures 3.14 and 3.15 summarise the number and types of fungicide applications across the three locations. Further details about the specific fungicides and application dates can be found in the Appendix (Tables A.1 and A.2).

Figure 3.14 shows monthly application counts by protection type, faceted by location. Bars are coloured according to the targeted pathogen group (*Botrytis*, *Podosphaera*, or both). This graph highlights variation in fungicide usage over time and across environments, with G1 receiving the highest number of applications. Additionally, fungicides targetting *Podosphaera* are applied predominantly in the greenhouses. Within the field there are fewer fungicide applications, which evenly target *Botrytis* or *Podosphaera*, and one application targeted both.

Figure 3.15 presents fungicide applications grouped by FRAC-defined resistance risk. Despite variation in application number, the relative proportions of different FRAC risk categories are broadly similar across the three locations. Grey bars (N/A) represent treatments without an assigned FRAC risk level, these are primarily biological or bacterial products, which are not currently classified by FRAC.

Two chemicals were excluded from the analysis due to uncertainty or irrelevance to the study's focus. "Rigel WP," listed by Wilkin & Sons Ltd, could not be verified as a registered fungicide and may have been a typographical error referring to either Rigel-G, a plant health enhancer, or Rovral WP, a broad-spectrum fungicide. Additionally, "Batavia," an insecticide, was excluded as this study focuses on fungal disease management. Both chemicals were applied only once to G1 (Rigel WP on 04/03/2022 and Batavia on 23/03/2022), and their exclusion is not expected to affect the overall conclusions.

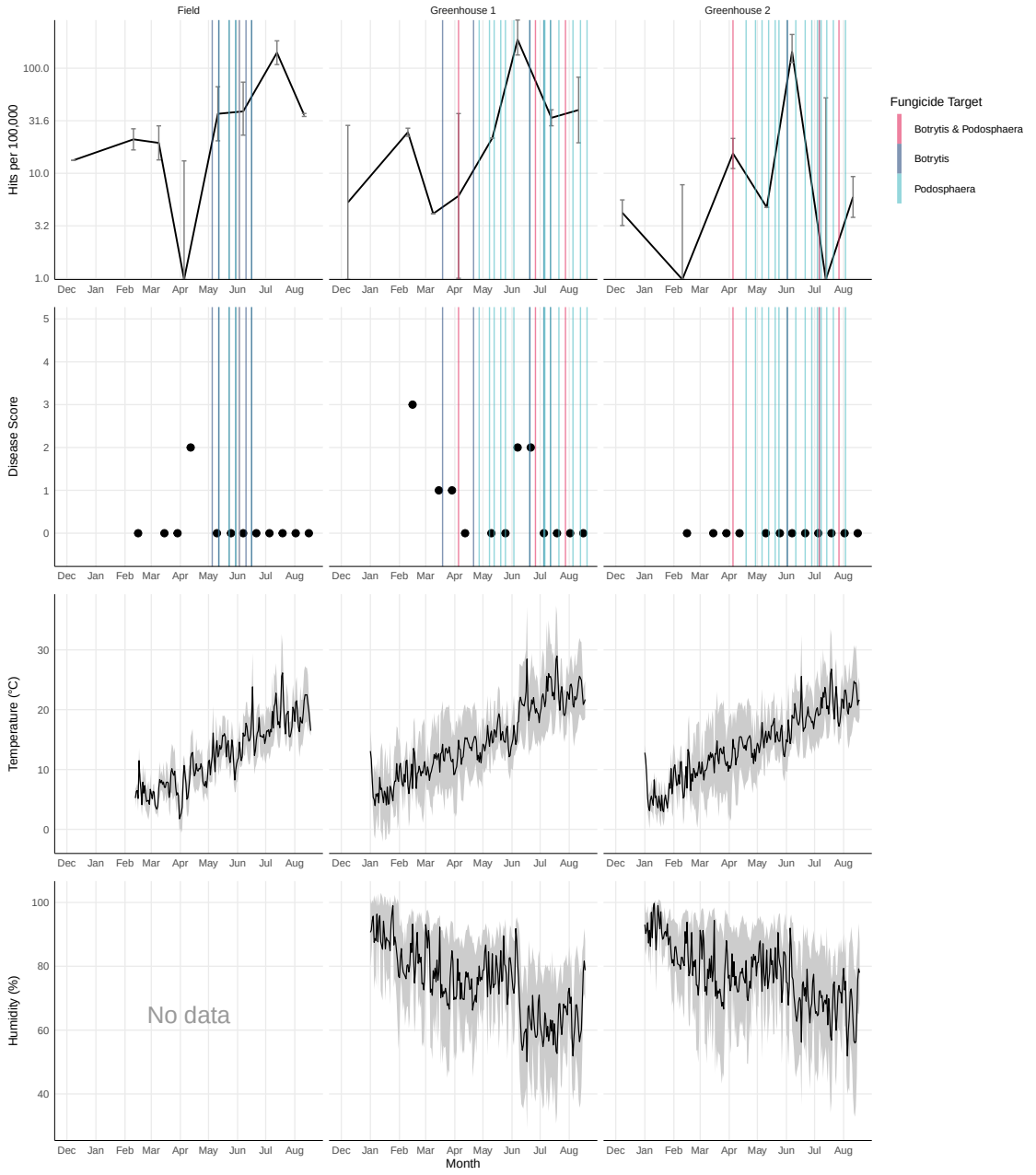


Figure 3.13: Same data structure as Figure 3.11, but for *Phytophthora*.

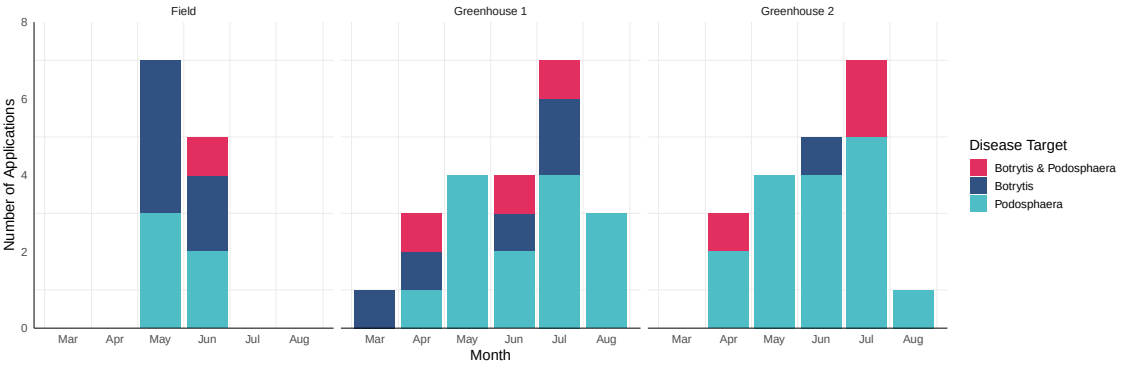


Figure 3.14: Fungicide spray counts by month, protection type, and location. Bars represent the number of sprays per month, grouped by protection type and faceted by location. Colours indicate the targeted disease group.

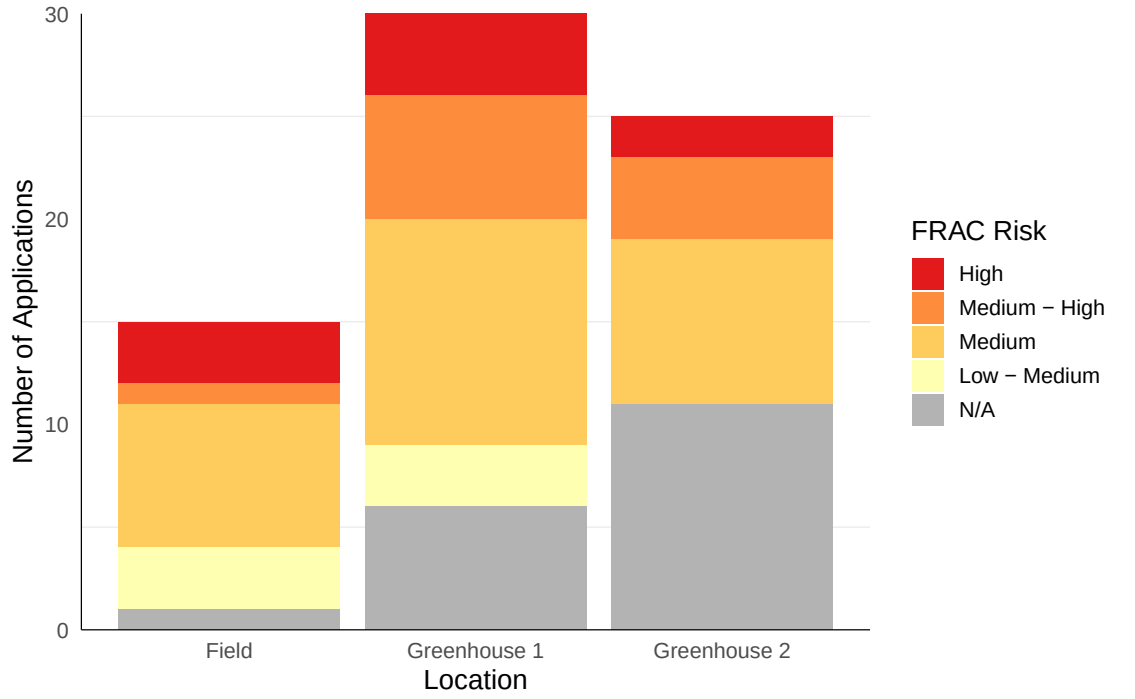


Figure 3.15: Fungicide spray counts by location and FRAC resistance risk. Bars show the number of spray events at each location, stacked by FRAC-defined resistance risk. Grey indicates treatments without an assigned risk level.

3.5.4 Alignment of AirSeq data to *P. aphanis* tissue assemblies and reference genome

Sequencing yield and read length metrics varied markedly across *P. aphanis* samples (Table 3.8). Sample 2 generated the largest dataset (4,100,415 reads) and the longest reads on average (mean 1,749 bp; N50 3,630 bp). Samples 1 and 4 were intermediate. Sample 3 was an outlier with far fewer and much shorter reads (698,707 reads; mean 150 bp; N50 144 bp). These input differences are consistent with the downstream assembly outcomes reported below.

Figure 3.16, demonstrates that the reads in the samples collected from powdery mildew infected leaves align to *Podosphaera* more than any other genera in the reference database. Sample 3 has the lowest abundance of *Podosphaera* aligning reads, but it is still the most abundant genera within that sample.

Assembly statistics for the four spore samples are summarised in Table 3.8. Sample 2 yielded the largest assembly at 97.2 Mb with the greatest number of contigs (49,801) and scaffolds (284), although mean coverage was moderate at 37. Sample 1 produced a 41.7 Mb assembly with 24,432 contigs, 68 scaffolds and the highest mean coverage at 53. Sample 4 assembled to 24.31 Mb with 14,139 contigs, 25 scaffolds and a mean coverage of 47. In contrast, Sample 3 assembled very poorly, with a total length of only 0.081 Mb from 18 contigs, no scaffolds and a mean coverage of 30. Fragment N50 values were in the low kilobase range for the longer assemblies assemblies (from 2,900 to 4,169), while the higher N50 for Sample 3 (7,990) reflects its very small assembly rather than good contiguity.

Across the four spore samples, the number of airborne sequencing reads that aligned varied widely, (Figure 3.17). From 419 to 6,510 alignments per sample, with Sample 4 the highest and Sample 3 the lowest; the reference genome produced 2,021 alignments. After applying the quality filter (identity $\geq 80\%$ and length ≥ 150 bp), Sample 2 retained

the most alignments (3,444), with the other samples ranging from 206 to 1,696, and the reference retaining 397. Among filtered alignments, mean identity clustered between 90 and 94%, and mean aligned length ranged between about 300 and 610 bp.

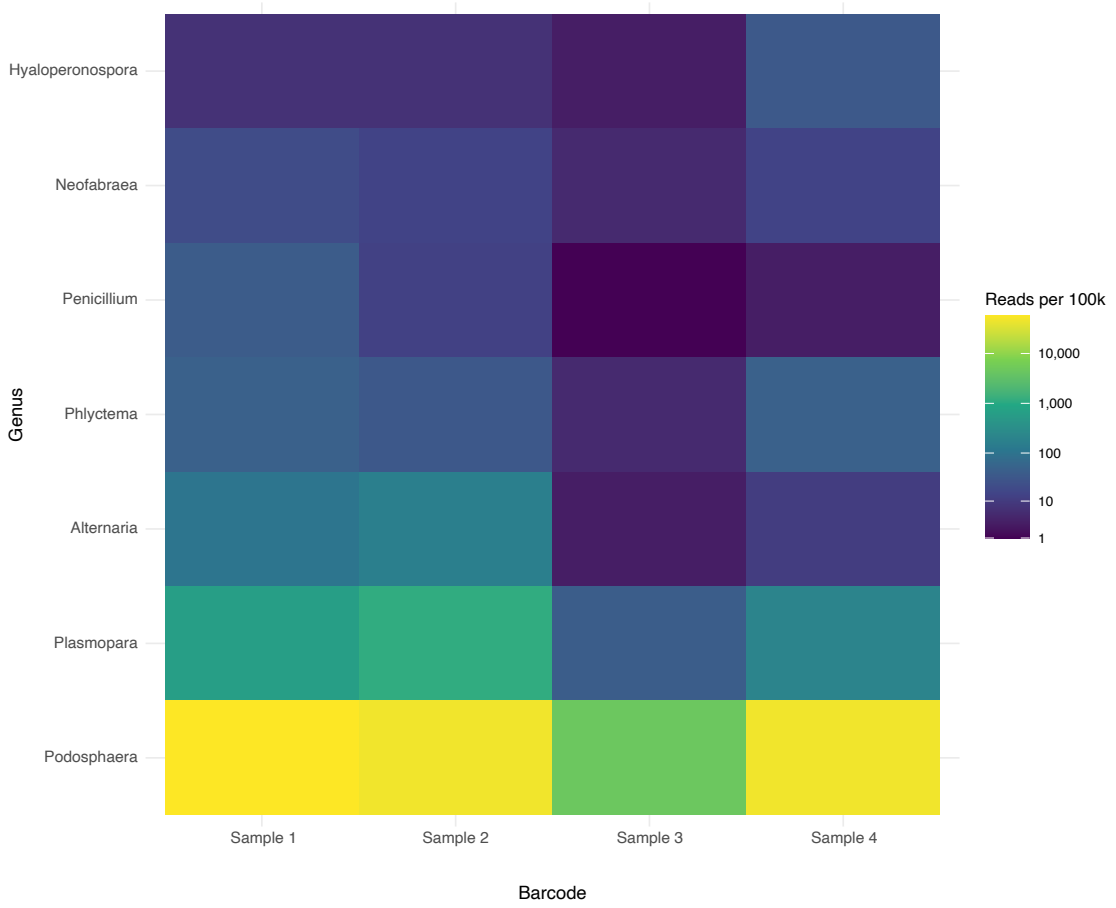


Figure 3.16: Genus level heatmap of *Podosphaera aphanis* spore sequencing reads aligned to a curated pathogen reference database, normalised to reads per 100k per sample with genera under 50 reads removed.

Table 3.8: Table summarising the sequencing data and Flye assembly metrics

Sample	**Read Number**	**Mean Length**	**N50 Length**	**Contigs**	**Contigs N50**	**Scaffolds**	**Mean coverage**
1	3,409,962	1,429	2,933	24,432	3066	68	53
2	4,100,415	1,749	3,630	49,801	4169	284	37
3	698,707	150	144	18	7990	0	30
4	2,344,700	1,301	2,727	14,139	2900	25	47

3.6 Discussion

This study examined the relationship between airborne fungal pathogen presence, environmental conditions, and disease scores at three locations on a commercial strawberry farm. Using AirSeq data, visual disease assessments, and metadata on climate and fungicide use,

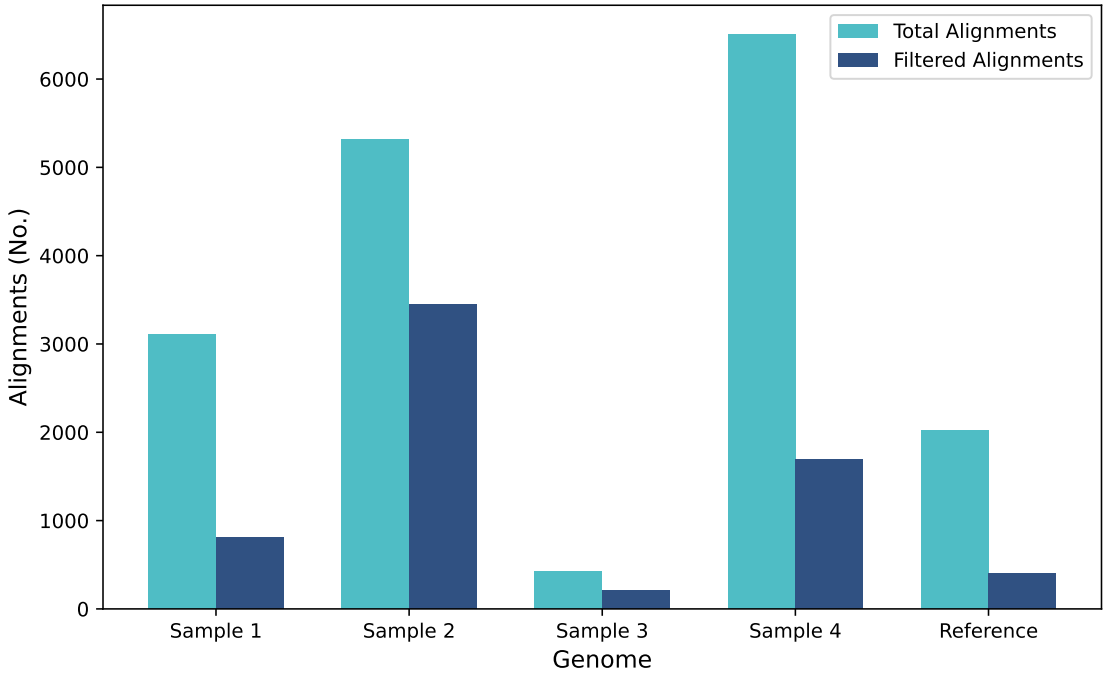


Figure 3.17: Bar chart showing, for each spore sample and the reference genome, the total number of alignments and the number passing the quality filter from the air sample mapping. Light blue bars show total alignments, dark blue bars show alignments with identity $\geq 80\%$ and length ≥ 150,bp.

several key patterns emerged. The most abundant airborne pathogens detected included *B. cinerea* and *P. aphanis*, both of which primarily infect strawberries (Figure 3.9). A more detailed analysis of these pathogens, along with *Phytophthora*, was conducted to compare their airborne presence with disease scores, assessing AirSeq's potential to detect spores before visible symptoms appear.

Initial alignment of the airborne data to a pathogen database showed the complexity of the airborne microbiome and the difficulties associated with correct taxonomic identification from limited DNA yields.

Following this, the airborne data was specifically analysed for three known strawberry pathogens, revealing differences in disease onset and severity across pathogens and locations (Figures 3.11, 3.12 and 3.13). *Botrytis* and *Podosphaera* were detected early in the season and showed strong correlations with disease scores, especially in the greenhouses. In contrast, *Phytophthora spp.*, infection on the plants was limited and detected at considerably lower airborne levels. Limited correlations were found between *Phytophthora spp.* disease scores and airborne sequences in G1, while similar airborne levels in G2 were detected despite no visible plant infection.

Environmental data from each location was used to further interpret these disease dynamics, guided by established literature detailing favourable conditions for infection and sporulation. This analysis demonstrated that despite the presence of airborne inoculum, disease establishment on the plants occurred only when environmental conditions were conducive. This insight is crucial for future AirSeq applications, as environmental conditions must be factored into any warning system, airborne inoculum alone is insufficient to cause infection.

Finally, fungicide applications (Figures 3.14 & 3.15) did not appear to reduce airborne

inoculum or disease incidence in a consistent way. In some cases, disease scores continued to rise despite repeated applications, particularly for *Botrytis* and *Podosphaera*, suggesting potential issues with application timing, efficacy, or resistance development in the pathogen populations.

These findings demonstrate both the potential and limitations of airborne pathogen monitoring as a tool for disease forecasting and management.

3.6.1 Pathogen presence in the airborne samples

To understand the dynamics of airborne pathogen presence, the sequencing data were aligned to a refined database of pathogen reference genomes using *minimap2*. Alignment results were then used to determine taxa counts per sample, which were visualised as heatmaps (Figure 3.8) and summarised in bar charts showing the top 20 and top 10 most abundant species (Figures 3.9 and 3.10, respectively).

The data show that a small number of pathogens were present at high abundance across the samples, including *B. cinerea*, *G. rostochiensis*, and *P. aphanis* (Figure 3.9). However, the dynamic nature of the airborne microbiome is evident, as many pathogens were detected only during specific time windows. For example, *B. cinerea* was most abundant from December to March, *G. rostochiensis* during April and May, and *P. aphanis* in June (Figures 3.8b and 3.10).

In contrast, spatial variation between the different sampling locations was less pronounced, with pathogen abundances generally consistent across sites. Nevertheless, notable exceptions were observed, such as G2 exhibiting more than double the *B. cinerea* reads per 100k compared to other locations, and G1 showing reduced *P. aphanis* abundance (Figure 3.10). G1 also displayed a lower overall diversity and abundance of detected pathogens relative to the other locations (Figure 3.8a).

3.6.1.1 Spatial and temporal comparison of airborne pathogens

The heatmaps (Figure 3.8) provide a clear visualisation of the variation in pathogen profiles across space and time. To improve interpretability, only species with more than 10 aligned reads and detected in at least 10 different samples were included. This filtering removed many low-frequency taxa that could otherwise obscure broader patterns, but may have reduced the diversity of the samples (Table 3.5 and 3.6).

The data were grouped by collection location and month to reveal the spatial and temporal patterns of pathogen abundance (Figure 3.8).

Spatial comparison of airborne pathogens Previous studies have shown that environmental conditions in greenhouses differ markedly from those in open fields, particularly with regard to humidity, temperature and airflow, all of which can influence pathogen communities [401]. Based on these differences, one might expect distinct pathogen profiles between the field and greenhouse environments. However, in this study, it was observed that G1 exhibited a more distinct pathogen profile, while G2 and the field shared greater similarity.

The lack of a clear distinction in the field's pathogen profile may be attributable to the filtering criteria applied during data processing. Although the field initially had a higher number of unique species compared to the greenhouses when filtering out taxa with

less than 10 assigned reads, this difference diminished when an additional criterion was applied, excluding species not present in at least 10 samples (Table 3.5). This suggests that the filtering approach is appropriate for visualising high-abundance pathogen species in heatmaps. However, it limits the capacity to compare overall diversity between locations, as a substantial number of taxa (ranging from 55 to 95) were removed from the analysis. These excluded species, despite their low frequency, may contribute meaningfully to the observed differences across locations.

Another possible explanation lies in the limitations of the reference database used in this study, which included only known pathogens. Numerous other airborne fungal or microbial taxa, not represented in the database, were therefore excluded from the analysis. Inclusion of a broader spectrum of airborne organisms could potentially reveal greater differences between sampling locations. While this study focused on airborne pathogen detection, this limitation is worth considering in future comparisons of air samples across locations.

Despite the overall similarity, there are differences in the pathogen profiles across the three locations, likely attributable to several factors. These include the different strawberry varieties cultivated (Table 3.2) [238, 297, 335], climatic contrasts between the field and greenhouses and microclimatic variation between the two greenhouses [53, 408]. Additionally, variation in the type and quantity of fungicides applied (Figure 3.14) is likely to influence pathogen abundance. While most research examines fungicide effects on soil microbiomes, similar impacts on airborne communities are likely, as fungicides have been shown to alter microbial diversity and structure [141, 229].

More detailed analysis of *B. cinerea* and *P. aphanis*, the two pathogens for which disease score data were available, revealed location-specific differences in disease prevalence and severity over time (Figures 3.11 and 3.12). Accordingly, it is reasonable to expect corresponding variation in the abundance of other airborne pathogens across locations.

Temporal comparison of airborne pathogens The heatmap illustrating the monthly presence of PHI-base pathogens (Figure 3.8b) highlights the dynamic nature of the airborne microbiome across the sampling period. Pathogen profiles from December 2021 and February 2022 show considerable similarity, with notably high abundances of *B. cinerea*, *Penicillium digitatum*, *Penicillium expansum*, and *Sclerotinia sclerotiorum* relative to other pathogens detected. Although these species remained detectable at lower abundances throughout the year, their abundances are greatly reduced outside this early period.

The early and abundant detection of *B. cinerea* is particularly relevant, confirming its established presence and potential inoculum reservoir at the farm in plants from the previous season. In contrast, the high abundance of *P. expansum*, typically associated with post-harvest fruit rot [104], is somewhat unexpected early in the growing season. This pattern may reflect residual spore populations from the previous season or reduced fungicide use during off-season periods allowing these fungal popualtions to grow. Similarly, *S. sclerotiorum*, a pathogen capable of causing wilting and white rot in strawberries [91, 240], was frequently detected. However, visual symptoms associated with *P. expansum* or *S. sclerotiorum* infection were not observed during plant monitoring so it is not possible to determine the accuracy of AirSeq as a method for monitoring these diseases.

In March the abundances of pathogens that were prominent in December and February declined, in general there are a larger number of pathogens present but at lower abundances.

B. cinerea has also declined in abundance from February but remains more prevalent than other pathogenic species. The second most abundant pathogen identified in March was *Alternaria alternata*, which can cause black leaf spot of Strawberry [247, 356]. Black leaf spot is another disease not scored by agronomists during the experiment, but the presence of *A. alternata* within a strawberry growing setting is not unexpected. Additionally *Alternaria spp.* have been consistently found at high abundance in other airborne eDNA studies [287, 345, 401] so it is expected that this study would also detect them at high abundance.

In April, several abundant taxa were detected, including *Agrobacterium tumefaciens*, *Globodera rostochiensis* and *Mycobacterium marinum*. *A. tumefaciens* does not routinely cause disease in strawberry, although in laboratory settings it has been used to induce genetic transformation by infection [279]. *G. rostochiensis* is a cyst nematode that predominantly infects potato and is typically found in soil [317]. *M. marinum* is a bacterium that infects fish and humans [143]. Consequently, detection of *M. marinum* in airborne samples from a strawberry farm is unlikely given the scarcity of suitable hosts, and further scrutiny of these alignments is required to verify whether the reads derive from these species rather than close relatives or misalignments.

In June, there are many pathogens present in the airborne microbiome at low abundance. Two species present at elevated abundance are *Parastagonospora nodorum* and *Ustilago hordei* neither of these are pathogens of Strawberry and are instead known to infect wheat and grains [120, 268], which will be growing in neighbouring fields in June, so they may have been detected as spores which had been transported greater distances through the air. The generally low abundance of pathogens observed in June may reflect increased pollen capture by the air sampler. Given the compositional nature of metagenomic data [126], a higher proportion of pollen reads would reduce the relative abundance of pathogens compared with other sampling dates.

In July *A. alternata* is detected at slightly higher levels than in March and *Klebsiella pneumoniae* levels are also high. *K. pneumoniae* is a known human pathogen [97], but it also occurs naturally in soil [19]. Its detection on the farm is therefore plausible, given both environmental presence and human activity.

In August, several *Alternaria spp.* were detected at higher levels, including *A. alternata*, which can infect strawberries; *A. brassicola*, which infects brassicas; and *A. solani*, a pathogen of tomato and potato [165, 284]. These are crops known to be grown in the East of England so it is plausible that spores could have blown over from neighbouring fields.

Overall, pathogen presence in the air appears to reflect seasonal changes, with both crop-related and external environmental influences. Some identifications, particularly of pathogens not typically associated with strawberry or agricultural environments (e.g. *M. marinum*), may result from sequence similarity with related taxa or database limitations. In addition, non-pathogenic taxa such as pollen, fungi, and environmental microbes also contribute to the airborne community, and their fluctuating abundance can influence the relative representation of pathogens in metagenomic datasets.

This analysis demonstrates the potential of AirSeq as a powerful tool for the simultaneous detection of hundreds of pathogens and for tracking their relative abundance over time. Such comprehensive monitoring would not be feasible through manual disease scoring alone. However, limitations in accurate taxonomic identification remain and must be addressed before the technology can be fully relied upon for routine diagnostic use.

3.6.1.2 Closer look at the most abundant pathogens

The most abundant pathogens included *B. cinerea*, *G. rostochiensis*, *P. aphanis* among others, see Table 3.7 and Figures 3.9 and 3.10. These species exhibited distinct temporal patterns, with different pathogens peaking in different months. Overall, the trends observed align well with existing literature, although a few unexpected patterns were also noted.

For example, *B. cinerea* is most abundant in December and remains dominant in February and March, before declining in later months. Given that strawberry is a known host, this classification is likely accurate. However, its reduced prevalence during the main growing season is notable. The presence and dynamics of *B. cinerea* are discussed in more detail in Section 3.6.2.1.

Additionally, two *Penicillium spp.*; *P. expansum* and *P. digitatum*, were detected across all three locations between December and March (Figure 3.10). *P. expansum* is known to infect both apples and strawberries [176, 406], whereas *P. digitatum* is primarily a pathogen of citrus fruit [86]. Given the host specificity of *P. digitatum*, it is likely that these reads were misclassified and instead represent *P. expansum* or other *Penicillium spp.* associated with strawberry that were absent from the reference database.

While the presence of *Penicillium spp.* in greenhouse environments is documented [73], data on their seasonal variation are limited. However, a study by Rodolfi et al., examining airborne fungal communities in a botanical garden greenhouse found that *Penicillium spp.*, including *P. expansum*, were most abundant in winter [315]. This finding aligns closely with the observations presented here.

In April and May, the most abundant taxon detected was *Globodera rostochiensis*, a cyst nematode that infects potatoes and other root crops, and is not pathogenic to strawberry. While specific data on the seasonal timing of airborne nematode dispersal are limited, wind has been identified as a vector for their movement [68, 296]. This may explain the higher relative abundance detected in field samples, where wind exposure is greater compared to the enclosed greenhouse environments.

Additionally, *G. rostochiensis* hatching is triggered by soil temperatures between 15 –27°C [187], which were reached during April and May. These warmer conditions may have increased nematode activity in the environment, potentially explaining their elevated detection during these months. An alternative explanation for the high abundance of nematode DNA could be soil disturbance and dust release caused by farm machinery, such as tractors, or by increased footfall associated with manual berry harvest as ripening began in the greenhouses. Although the cause is difficult to determine from a single visit per month, the consistent detection of *G. rostochiensis* across nearly all locations in both April and May suggests the presence of an environmental trigger. The detection of nematodes at high abundance highlights the broader utility of AirSeq as a tool for monitoring diverse airborne organisms, beyond fungal pathogens.

P. aphanis was detected at its highest abundance in June in both the field and G1, but was not observed in G2. Infection typically occurs at temperatures between $15 - 25$°C and relative humidity levels of 75–98% [16], conditions that were present in June (Figure 3.12) and likely contributed to the increased prevalence during this period. However, the absence of detection in G2 remains unclear. The occurrence of *P. aphanis* is explored further in a later section.

In July and August the predominant pathogen is *Alternaria alternata*, this is supported

by other literature of airborne pathogens that have detected higher levels of *Alternaria spp.* between June and October [113, 324, 345].

Accurate species-level identification using AirSeq can be challenging, particularly when highly abundant reads are misclassified, as these are potentially of greatest concern to growers. In the top 10 most abundant taxa, there are cases where species are likely misidentified due to limitations in the reference database. For example, reads were assigned to *Ascochyta rabiei*, a fungal pathogen of chickpea that is unlikely to be present in the UK [344], and to *Ustilago hordei*, the causal agent of barley smut. Neither species has closely related counterparts known to infect strawberry, but may belong to closely relates non-pathogenic species that were not included in the reference database. In the case of *U. hordei*, it is also possible that spores originated from nearby cereal fields and were transported by air. These issues highlight the limitations of AirSeq for precise taxonomic resolution, particularly at the species level, and are discussed in more detail in the limitations section (3.6.4).

The alignment between the seasonal appearance of detected pathogens and published literature on their respective genera provides reassurance that AirSeq is capable of accurately capturing expected airborne taxa across the year. However, in the absence of a definitive ground truth, it remains difficult to assess whether there are species present in the environment that are not detected with AirSeq. In this context, incorporating disease score data is valuable, as it offers insight into the actual condition of the plants at the time of sampling and supports the interpretation of airborne detection results.

3.6.2 Disease onset in relation to environmental conditions and fungicide application

To better understand the relationship between airborne pathogen presence and plant infection, disease scores and fungicide usage were analysed alongside sequencing data for three diseases: grey mould, powdery mildew, and Phytophthora rot.

For grey mould and powdery mildew, fungicide applications did not consistently prevent disease onset or progression (Figures 3.11 and 3.12). This is evident from the fact that neither manual disease scores nor sequencing data showed a notable decrease following fungicide treatments. In contrast, fungicide efficacy against *Phytophthora spp.* was harder to assess, as the disease was largely absent during the experiment (Figure 3.13).

Many of the fungicides used have medium to high resistance risk according to FRAC classifications (Figure 3.15). This raises the possibility that pathogen populations may already carry resistance to some of the active ingredients. Resistance to a range of fungicides has been documented in *B. cinerea* (Table 3.1), and similarly in *P. aphanis* and *Phytophthora spp.*, although the genetic basis of resistance in these pathogens remains poorly characterised [238, 239, 275, 322, 332, 347]. Even if resistance is not yet widespread in the populations in this study, continued reliance on high-risk fungicides may drive its development.

Due to the absence of a fungicide-free control location, it is not possible to conclusively determine whether the treatments were ineffective. It is more plausible that they suppressed disease severity to some extent, as disease scores never reached the maximum rating of 5 in any site. Still, the data suggest that the fungicides were not sufficient to fully eradicate disease symptoms. Future studies should include untreated control plots and/or pathogen culturing with sensitivity assays to more rigorously evaluate fungicide efficacy.

There is precedent for this approach: for example, mefenoxam resistance in *Phytophthora cactorum* was confirmed through culturing and testing [239].

Regular resistance monitoring and rotating fungicides with different modes of action remain essential to slow resistance development. However, with few new modes of action entering the market, growers increasingly face a difficult trade-off: prioritising short-term disease control or preserving fungicide efficacy for future use [84]. In this context, AirSeq could serve as a valuable disease management tool by enabling more targeted fungicide applications, potentially reducing unnecessary treatments and associated selection pressure.

The specific fungicides applied, and their effects on disease scores, airborne pathogen abundance, and associated environmental conditions, are further discussed in relation to each monitored disease.

3.6.2.1 *Botrytis* - grey mould

B. cinerea airborne DNA was detected at its highest airborne concentration in December 2021 across all three locations (Figure 3.10). However, no disease score data were available for this month, preventing a direct comparison between pathogen presence and disease incidence. Airborne *Botrytis* inoculum remained detectable throughout the entire growing season, with a dip observed in all locations in April and peaks between May and August (Figure 3.11).

While *B. cinerea* DNA was consistently detected during the growing season, the relative abundance was lower (<1,000 hits per 100,000) compared to the December peak (∼10,000 hits per 100,000) (Figure 3.11). *Botrytis* can overwinter on aboveground plant material (Figure 3.1), which likely explains the elevated DNA levels detected in December. Senescent plant tissues left in the field can harbour mycelia and contribute to early-season inoculum sources [382]. However, one might expect comparable airborne *B. cinerea* levels in spring when temperatures are again suitable for infection, but from March - April airborne *Botrytis* levels were below 1,000 hits per 100,000. This December spike may not solely reflect an increase in *Botrytis* biomass, but could also result from the compositional nature of microbiome sequencing, where relative abundance appears higher due to a decline in other airborne taxa [126].

There are also differences between locations in December when the top 10 most abundant pathogens are plotted (Figure 3.10), samples from G2 contained much higher levels of *B. cinerea* than the other locations (∼60,000 hits per 100,000 compared to ∼20,000 hits per 100,000). This may be related to the everbearing variety grown in G2 (Table 3.2), which remained in the greenhouse for multiple years, whereas G1 was left fallow until March. The greenhouse environment may also have provided conditions more conducive to infection than those in the field. Other potential factors contributing to differences in airborne inoculum between locations include variations in fungicide application and microclimatic conditions.

Spores of *Botrytis* are known to infect plants under moderate temperatures (15–25°C) and high humidity conditions [249]. These conditions were met in the field from June to September, and in the greenhouses from March/April to September, at these times *B. cinerea* DNA was also detected in the air (Figure 3.11). The environmental conditions alongside the airborne inoculum explain the increased disease scores seen between March and August to 2/3 in all the locations (Figure 3.11).

There are cases where both disease score and airborne inoculum drop to negligible levels, such as in both greenhouses in April. However, this pattern does not hold in the field, despite near-undetectable airborne inoculum in April, the disease score remains at 3. This highlights both the potential and the current limitations of AirSeq. On one hand, it can dynamically track shifts in airborne inoculum that often correspond closely to observed disease. On the other, discrepancies, such as the field result in April, demonstrate that inaccurate signals could mislead growers if not interpreted cautiously.

One possible explanation for this discrepancy is sampling error. Although replicate samples were taken, both may have been affected by technical issues such as poor processing or reduced sequencing. This suggests that two replicates may be insufficient for reliable detection, especially if samples are collected only monthly. Higher sampling frequency or improved technical replication may be necessary to ensure AirSeq results are robust enough to support disease control decision-making.

The use of fungicide also plays a large role in disease dynamics at the sampling location. Fungicide applications targeting *Botrytis* began in early March in the greenhouses and in May in the field, by which time *Botrytis* DNA had already been detected in the air at all locations but disease was only just detected on the plants. In the field, these applications had no clear effect on airborne inoculum levels or disease scores, which remained relatively stable, at 3, until July (Figure 3.11).

In G1 the first fungicide applied (Switch 19/03/22) does not appear to affect the disease score. However, Signum applied to both greenhouses on the 05/04/22 does reduce the disease score in both cases from 2 to 0. It is not possible in this case to determine whether fungicide application influenced the airborne inoculum, as air samples and fungicide treatments occurred on the same day, but the precise timing of application is unknown. Future studies could benefit from collecting air samples both before and after fungicide application to better assess its impact on airborne fungal spores.

Signum contains boscalid, a succinate dehydrogenase inhibitor (SDHI), and pyraclostrobin, a quinone outside inhibitor (QoI). The active ingredients are classified as medium - high and high resistance risk by FRAC, respectively, and their continued use without rotation increases the likelihood of resistance development. In this study Signum was only applied once in the field and G2 and twice in G1. The subsequent application of Signum on the 26/06/22 in G1 was not followed by a noticeable reduction in disease levels. Similarly, its only application in the field on 03/06/22 had no measurable impact on disease scores.

As Signum appeared effective in some instances but not in others, its efficacy in this context remains inconclusive. The variation in response could be attributed to evolving resistance in the pathogen population, interactions with other chemical treatments, or differing environmental conditions at the time of application. The persistence of disease at earlier time points, despite repeated applications of high-risk fungicides, supports the possibility that resistance may be emerging in *Botrytis* populations.

Interestingly, in July, a decline in disease scores and airborne inoculum was observed across all three locations. In the field, this decrease occurred without any fungicide applications during that month (last application 16/06/22), suggesting environmental factors may have contributed. Notably, across all the locations maximum temperatures in July exceeded 30°C, which is above the optimal range for *Botrytis* infection. In the greenhouses, multiple fungicides were applied during this time, making it difficult to determine whether the observed reductions were due to chemical control, suboptimal infection conditions, or

a combination of both.

Grey mould severity differed among locations, with the field reaching a higher maximum disease score than the greenhouses (three versus two). Despite this, airborne *Botrytis* inoculum was of similar magnitude across all three sites. The divergence between disease scores and airborne inoculum may reflect reduced fungicide use in the field and environmental differences, as field temperatures were generally lower than in the greenhouses. In Florida, strawberries grown in tunnels showed lower *Botrytis* incidence than those in fields [397]. The discrepancy between that study and the results presented here may arise because greenhouses do not match tunnels in temperature and humidity regimes, and due to broader climatic contrasts between Florida and England.

3.6.2.2 *Podosphaera* - powdery mildew

Spores of *Podosphaera* were detected in the air throughout the sampling season, with airborne abundance increasing from May to September, mirroring the rise in disease scores (Fig. 3.12). A clear correlation is evident in April, disease scores were zero across all sites, and airborne *Podosphaera* levels remained low (<100 hits per 100,000).

However, this relationship between airborne inoculum and disease progression was not always consistent. In some instances, such as the field in March and June, increases in airborne spores preceded visible disease. In contrast, G2 showed a decrease in spore counts following a decline in disease from April to May. In other cases, the two measures changed simultaneously, for example in G1, both disease and spore levels declined between June and July. These patterns highlight the complexity of the relationship between inoculum presence and disease symptoms.

For infection to occur, *P. aphanis* spores must be present in the environment under suitable climatic conditions. Specifically, temperatures between $15 - 25$°C and RH of 75% - 98% are required for successful germination and host penetration [16]. Once initial infections are established under these conditions, the pathogen rapidly proliferates through cyclic secondary infections within the plant canopy. This leads to further spore release, compounding airborne inoculum levels and intensifying disease pressure across the crop (Figure 3.2).

The environmental conditions required for *P. aphanis* infection were met consistently from June to September in the field, and from as early as March/April to September in the greenhouses. This seasonal window aligns closely with the observed onset and progression of disease, commencing in June in the field and March in the greenhouses. Further supporting the role of temperature and humidity in driving infection dynamics (Figure 3.12). An additional consideration when understanding the relationship between airborne inoculum measured by AirSeq and the disease score is the application of fungicides.

Despite the application of 19 fungicides targeting powdery mildew over the course of the season (Figure 3.14), there were few instances where either airborne inoculum or disease scores showed a marked decline. This raises questions about the timing of applications, potential fungicide resistance, or overall application efficacy. One occasion where fungicide use appeared effective occurred in April, when Signum was applied to both greenhouses on 05/04/22. This treatment coincided with the removal of visible disease symptoms in both greenhouses, which had been scored at 2 in the previous fortnight.

Signum, which also reduced grey mould caused by *B. cinerea*, contains two active

ingredients classified by FRAC as medium–high and high resistance risk. However, the effect was not sustained; by the following fortnight, disease scores in both greenhouses had risen to 3, the highest level recorded during the season. This suggests that while the fungicide may have temporarily suppressed visible symptoms, it did not prevent reinfection. One possible explanation for the quick reinfection is that the fungicide did not significantly reduce airborne inoculum levels. However, since spore sampling was conducted on the same day as the Signum application, it is difficult to determine the immediate impact of the treatment on *Podosphaera* abundance in the air.

Signum was applied again later in the season in G1 (26/06/22) and was also used once in the field (03/06/22). These applications did not noticeably reduce powdery mildew disease or airborne *Podosphaera* inoculum suggesting the reduction of disease in April may have been caused by more than just the fungicide or that the efficacy changes in different locations or with repeated applications.

In the absence of a clear, consistent reduction in disease following fungicide applications, it is important to also consider environmental conditions that play a role in shaping disease dynamics. Although the two greenhouses experienced broadly similar environmental conditions, disease progression differed between them, peaking at a score of 3 in May–June in G1, and from May to August in G2. These differences in disease scores may be attributable to subtle microclimatic variation, observer error, or other unmeasured factors.

For example, in June, both greenhouses experienced peaks in maximum temperatures above 30°C. In G1, this coincided with a reduction in disease, while no such decline was observed in G2. Similar temperature spikes occurred throughout July and August, yet reductions in disease were only apparent in G2 late in the season. Humidity patterns may also have contributed to the disease suppression seen in G1 in June, where there was a drop in relative humidity, which may have influenced pathogen development. However, a comparable drop in G2 did not lead to reduced disease, again highlighting the complexity of interactions involved.

Another reason for the differences in airborne inoculum and disease scores between the greenhouses may have been the Strawberry varieties (Table 3.2). The Favori variety in G1 are reported as strongly resistant to powdery mildew whilst those in G2 had intermediate resistance or were moderately susceptible to powdery mildew. This difference in varietal susceptibility may also account for the greater number of fungicide applications targetting *Podosphaera* in G2 compared to the other locations (Figure 3.14).

A comparable study conducted in Spain examined the airborne concentration of *P. aphanis* spores alongside disease incidence and additional environmental metadata. It reported a positive correlation between spore concentration and temperature, and negative correlations with both humidity and rainfall [43]. The study also found that increased airborne conidia levels were associated with higher disease incidence, which is consistent with the summer data presented here. In the current study, rainfall was not included as an environmental variable, but this may warrant consideration in future analyses, particularly for the field samples.

Another factor to consider is the potential interaction between co-infecting pathogens. Late-stage infection by *P. aphanis* has been shown to increase strawberry plant susceptibility to *B. cinerea* [66]. While the current data do not allow a causal relationship to be confirmed, similar disease patterns were observed. For instance, both pathogens were absent across all locations in late April, and both were absent in G2 in August. These

patterns may reflect either a biological interaction between the pathogens or a shared sensitivity to environmental conditions and fungicide applications. However, in the field, their trajectories diverged later in the season. By July and August, *Podosphaera* disease scores reached 4, while *Botrytis* decreased to 1–2. This contrasts with the literature that suggests *P. aphanis* infection increases susceptibility to *B. cinerea* [66], and highlights that such interactions may be context-dependent or influenced by other limiting factors.

3.6.2.3 *Phytophthora spp.* - crown & Leather Rot

The presence of *Phytophthora spp.* spores in the air was consistently low across all three locations (Fig. 3.13), which is expected given that this oomycete pathogen is primarily spread through water and infected plant material, rather than via airborne transmission [238, 239, 278]. However, in cases of severe infection, it is possible that airborne samplers may capture plant material containing the pathogen. This has been demonstrated in studies which were able to collect and identify *Phytophthora spp.* fungal spores from passive airborne collectors [101, 102, 132, 211].

Wet conditions are optimal for *Phytophthora* disease development [332], in this study there is not rainfall, irrigation or leaf wetness data. Therefore, the relationship between disease, airborne abundance and wetness cannot be discerned. Warm temperatures of 17 - 35°C are also required for disease development, with the optimal temperature for sporangium production at 20°C [25, 230].

Airborne *Phytophthora spp.* levels followed a broadly similar pattern across all locations, with peaks of ∼100 hits per 100,000 detected in June in the greenhouses and in July in the field (Figure 3.13). Despite this, no *Phytophthora* disease symptoms were observed in G2, and in the field, disease was reported only once in late April at a score of 2. Interestingly, this presence of visible disease in the field coincided with a sharp decline in airborne *Phytophthora* levels, suggesting a mismatch between measured airborne spore abundance and disease expression. However, *Phytophthora* abundance differed markedly between the two April field samples, raising the possibility that one replicate sequenced poorly and yielded unreliable data.

In contrast, G1 exhibited a relationship between the airborne spores and *Phytophthora* disease symptoms. With a disease score of 3 recorded in February, aligning with a modest increase in airborne inoculum followed by both disease scores and airborne spore counts declining from February to April. Later in the season, *Phytophthora* spore levels rose again in all three locations during June and July, yet disease was only observed in G1. This suggests a site-specific correlation between airborne inoculum and disease symptoms in G1, which was not evident in the other two locations.

One possibility for this discrepancy could be due to incorrect disease scoring, as at this time the plants were heavily infected with powdery mildew (score 3) and grey mould (score 2 - 3) (Figures 3.11 & 3.12) which may have impacted the agronomists ability to score the plants. Alternatively, airborne inoculum may have been present at all locations, but only plants in G1 developed infection, potentially because the planted cultivars were more susceptible or because microclimatic conditions favoured infection. This cannot be verified since *Phytophthora* susceptibility data for the strawberry cultivars is unavailable (Table 3.2).

Phytophthora had the lowest prevalence in this study and was not a primary target

of fungicide applications. As a predominantly soil borne oomycete, *Phytophthora* is often managed through soil fumigation; however, fumigation records for the sampling site were unavailable. Foliar fungicides with activity against *Phytophthora* are available, and in this study symptom scores improved following fungicide application.

There are two instances where *Phytophthora* disease symptoms decreased following fungicide applications. In the field, disease symptoms diminished after the first application of Switch on 05/05/22 and did not reappear for the remainder of the season. However, since the fungicide was applied only a few days before disease scoring, it is possible that the decline was already underway due to environmental factors.

In G1, disease severity declined from a score of 3 in February to 1 in March, prior to any fungicide treatment. Following the first Switch application, the disease remained at score 1, and subsequently decreased to 0 after an application of Signum. Notably, Signum was also effective against *Botrytis* and *Podosphaera* in this study. Disease reappeared in G1 in June, with a score of 2, but a cluster of fungicide applications during that month was followed by a complete absence of disease in July. Due to the number of treatments applied within a short window, it is difficult to identify which specific product, if any, was responsible for the reduction in disease.

These observations across the three considered diseases underscore the multifactorial nature of fungal epidemiology. With limited resolution from the airborne spore data and overlapping variables such as fungicide applications, temperature fluctuations, and humidity changes, it is difficult to draw definitive conclusions about the relative contribution of each factor. The complexity of these interactions suggests that both biological and environmental data must be considered together to fully understand disease progression. These findings inform future use of AirSeq and underscore the need to collect sufficient environmental metadata to contextualise results and assess disease risk, which depends on both airborne pathogen detections and environmental conditions.

3.6.3 *P. aphanis* tissue collected from leaves compared to airborne eDNA data

From the *P. aphanis* tissue collected from the infected plants, the sequencing inputs differed markedly across samples (Table 3.8). Sample 3 produced very few and very short reads, which is the most plausible explanation for its poor assembly and low recovery of *Podosphaera* signal in downstream analyses. By contrast, Samples 1, 2 and 4 yielded adequate read counts and lengths and assembled to appreciable sizes.

Genus level alignment data indicate that the tissue samples are dominated by *Podosphaera* reads (Figure 3.16), consistent with powdery mildew infection. When the AirSeq data was aligned to these assemblies and the reference genome, Samples 1, 2 and 4 each produced more passing alignments than the reference genome 813, 3,444 and 1,696 versus 397, with similar post filter identities and aligned lengths. This pattern indicates that the air sample reads have closer sequence identity to the plant collected strains than to the available reference, giving confidence that the organism captured in air is the same strain to that infecting the crop at the time of sampling.

Overall, these results suggest that air sampling does recover the locally infecting strain. The weak performance of Sample 3 likely reflects insufficient data rather than biological absence, and highlights the need for adequate sequencing depth to generate high quality

assemblies usable in downstream analyses.

3.6.4 Limitations of the study

As this was a preliminary study to test the efficacy of AirSeq in detecting the presence and abundance of plant pathogens, several constraints became apparent. These should be considered when interpreting the findings and provide lessons for future experiments, spanning the entire process from sample collection to laboratory work, bioinformatics and analysis.

3.6.4.1 Limitations with sample collection and processing

One limitation of the current study is the relatively low temporal resolution of airborne sampling, which was conducted on a monthly basis. In contrast, disease severity on plants was assessed fortnightly. Monthly sampling substantially reduces the ability to capture short-term fluctuations in airborne pathogen abundance, and consequently limits the opportunity to detect pathogens prior to the onset of visible infection. This restricts the potential of the method for early warning and timely intervention.

Higher-resolution sampling would not only improve the likelihood of detecting pathogens before symptoms appear but would also allow clearer insights into temporal trends in airborne pathogen dynamics. Such data could support the development of predictive models that integrate AirSeq data with environmental variables such as temperature, humidity and ventilation. These models could be used to guide proactive crop management strategies, including adjustments to greenhouse conditions, to create environments less favourable for pathogen establishment and spread.

Furthermore, the collections themselves were 30 minutes each and collected in succession. This short collection reduces the diversity of the sample as other studies have found that more species are identified with increased sample length [106] and could mean that important pathogenic species are not detected as they are prevalent at different times of the day. The shorter sample also limits the DNA yield requiring the use of WGA to have sufficient DNA for sequencing, but WGA can limit sequencing efficiency and introduce bias.

Additionally, by conducting back to back collections when there is a large difference in the abundance of one species between the two samples it is not possible to determine if this is an environmental difference caused by the sudden release of a pathogen or a technical error with the sampling. This large difference in the samples occurred relatively frequently such as with *Podosphaera* in G1 in March and June and in the field in June (Figure 3.12).

This experiment was conducted over a single growing season (December to August), which limits the extent to which the observed pathogen trends can be generalised, as they may reflect conditions unique to this sampling year. Replication across multiple years would have strengthened the ability to generalise the findings and provided more robust support for demonstrating AirSeq's capacity to detect airborne pathogens prior to the appearance of visible symptoms on plants. Furthermore, extending sampling across additional seasons or locations would have enabled evaluation under a broader range of environmental conditions and disease pressures. Nevertheless, as a preliminary assessment, a single year of data was sufficient to demonstrate the feasibility of the method and to identify areas requiring refinement.

The collection of *P. aphanis* tissue to assemble genomes and confirm that the strain corresponded to that identified in the airborne data represented an important step in further validating the AirSeq method. However, as these collections were undertaken the year following the initial sampling, they are not directly comparable with the majority of the data presented in this chapter.

Contamination is a major concern in airborne eDNA studies, particularly due to the low DNA yield and the focus on detecting all taxa within the airborne microbiome. While sterile procedures were followed during sample collection and processing, including replacing the neck and cone of the Coriolis µ sampler between locations, there are additional measures that could have been implemented to further reduce the risk of contamination.

For example, this study did not include negative controls, which would have allowed for the identification of potential contaminant taxa within the samples. However, a separate study using the same sampler and protocol included appropriate negative controls and found no evidence of contamination originating from the sampler itself [124].

Another potential limitation is the vast difference in the DNA yield between the samples both before and after WGA as shown in Figure 3.7. This difference was controlled by creating libraries with the same concentration of DNA where possible and normalising read counts against total read counts for each sample. But these differences could still potentially have affected the results. Perhaps in the future more duplicate samples could be collected to increase the reliability of the results.

Finally, there is only disease scoring data for three pathogens. Although these are the most prominent strawberry pathogens it could have been interesting to attempt to identify the presence of other diseases in the sampling locations either visually or through further sequencing of infected tissue.

3.6.4.2 Limitations with the bioinformatics and analysis

In addition to limitations with the sample collection and processing, bioinformatics adds another layer of considerations. As the way data is handled can affect the identified taxa, often there is a fine balance between minimising false positives and false negatives.

In this experiment the raw sequence data was filtered on quality using standard ONT parameters with only pass reads analysed. These reads were further filtered on length (>300 bp) and then aligned to reference genomes with *minimap2*, introducing a few potential sources of bias.

The length filtering step aimed to improve the accuracy of taxonomic assignments, since there can be difficulty in correctly assigning short reads as they often lack sufficient genomic variation or may be conserved regions. However, removing short reads may have led to the loss of potentially informative sequences. The 300 bp threshold was chosen based on preliminary assessments of read length distribution and the proportion of reads retained, but this decision may have introduced a bias by disproportionately filtering out shorter yet valid fragments (e.g. from easier to lyse organisms), possibly affecting downstream diversity estimates. Running the analysis with multiple thresholds could help address this bias; however, due to the time and computational resources required, this approach was not pursued in the current study.

The reference database used also has a large impact on the detected taxa as only species present in the database can be identified. Taxa not present in the database may

therefore be classified to closely related species whose reference genome is available. The limitations associated with reference databases used are described in more detail in the earlier literature review (Chapter 2).

An example of the challenges with read classification can be seen in this experiment, where some reads may have been misassigned despite the correct species being present in the database. Misclassification is particularly likely between closely related species with high genomic similarity. For instance, in the pathogen heatmaps (Figure 3.8), a high proportion of reads correctly align to *B. cinerea*, a strawberry pathgoen identified on the plants during sampling (Figure 3.11). However, a small number of reads also align to *B. elliptica*, a species known to infect plants in the *Lilium* genus and not expected in this context. Given their close relationship, it is plausible that these *B. elliptica* assignments are false positives arising from misaligned *B. cinerea* reads. Although the number of such reads is small and unlikely to influence overall results, similar misclassifications could be more problematic with other taxa or in studies targeting low-abundance species.

These issues are heavily influenced by the alignment parameters chosen during analysis, such as *minimap2* flags and the MQ filter. These parameters determine which reads are retained or discarded, and small changes can shift the balance between sensitivity and specificity. Without the use of mock communities or simulated datasets, it is difficult to evaluate the accuracy of these parameters in real-world samples. Furthermore, as the full diversity of airborne eDNA at the time of sampling cannot be directly observed or quantified, it remains uncertain whether unexpected reads represent true biological presence or result from taxonomic misclassification.

Reads were aligned using *minimap2* with the `map-ont` preset, which is tailored for ONT reads with an expected error rate of approximately 10% [217]. This setting optimises alignment sensitivity for long, noisy reads typical of nanopore sequencing.

minimap2 assigns a MQ score to each alignment to reflect confidence in its taxonomic assignment. This score considers both the strength of the primary alignment and its similarity to secondary alignments. Reads that align clearly to a single taxon receive high MQ scores, while those with multiple similar alignments receive lower scores, indicating greater uncertainty [217].

To improve the reliability of taxonomic classification in this study, alignments with an MQ score of <5 were excluded from the analysis but those with a MQ of 0 were retained to check the secondary alignments. These thresholds were chosen to strike a balance between retaining high-quality alignments and minimising the loss of informative reads. A more stringent threshold might reduce false positives but could also remove true matches, particularly for taxa with conserved genomic regions. Consequently, the selected threshold likely influenced the final community composition by shaping which reads were included.

Further, reads mapping equally well to multiple taxa were removed unless one alignment was clearly superior. Although an alternative approach, such as using a LCA algorithm could have retained these reads at higher taxonomic ranks, this was not applied in the current analysis. These ambiguous reads represented between 0% and 8.35% of total reads, suggesting a limited but non-negligible impact on the overall results. Future studies may benefit from using strategies like a LCA algorithm to include these ambiguous alignments.

An additional consideration with eDNA datasets is their compositional nature [126], meaning read counts represent relative not absolute abundances. Therefore, an increase in one taxon's reads can artificially lower the apparent abundance of others, even if their

actual quantities remain unchanged. This makes it difficult to determine whether observed differences between samples reflect true biological variation or shifts in community composition from differential sequencing depth. This impacts the interpretations of the results from this study as samples are being compared over time to monitor pathogen abundance changes.

The use of spike-ins has been successfully used to quantify the amount of fungal DNA within an airborne sample to overcome the compositionality concern [274]. Incorporating such controls into future studies would improve cross-sample comparisons and allow more accurate interpretation of airborne pathogen abundance in future studies.

These limitations highlight several areas where the AirSeq method could be refined and presents important considerations for interpreting the results. However, they do not undermine the findings of this preliminary study, which demonstrate the potential of AirSeq as a valuable tool for detecting airborne plant pathogens, potentially before visible disease symptoms appear.

3.7 Conclusion and Future Work

In conclusion, this study shows that it is possible to detect fungal plant pathogens from the air, and that the amount of airborne pathogen DNA correlates with disease incidence in strawberry plants.

This work has demonstrated that airborne sequencing can be used to monitor both the airborne pathogen community and specific diseases of interest. The ten most abundant pathogens were detectable at expected times of the year (Figure 3.10), and the seasonal patterns in airborne eDNA for the three closely monitored pathogens broadly reflected disease progression in strawberry crops (Figures 3.11, 3.12 & 3.13). Notably, disease onset occurred only when environmental conditions were favourable for each pathogen, suggesting that while the presence of airborne inoculum is necessary, it is not sufficient for infection. Suitable environmental conditions are also required. This distinction is important when interpreting future AirSeq data: growers may not need to act solely on high pathogen detection unless the environmental conditions are conducive to disease development. Additionally, the data shows that the strain detected in the air is more closely related to the strains infecting the plants than to the available reference genome.

These findings have important implications for future research. Validating AirSeq against pathogens that were simultaneously assessed through disease scoring increases confidence that the method captures genuine biological signals. This provides a strong foundation for deploying AirSeq in other settings where traditional disease monitoring is less feasible or not routinely carried out. Ultimately, the goal of AirSeq is to complement or even replace conventional disease scoring as a tool for managing crop health. Achieving this will depend on confirming that pathogen levels in the air reliably predict disease outbreaks. For instance, that detecting high levels of a pathogen signals a genuine risk of infection if no action is taken.

This study also highlights the broader value of AirSeq's WGS approach, which enables the simultaneous detection of a wide range of pathogens, not just those being actively targeted. Because all DNA in the air is sequenced, AirSeq can also flag emerging or unexpected pathogens that growers may not yet be monitoring. This feature is especially important as climate change alters the geographic distribution of plant diseases and introduces new

risks into existing growing environments.

If research at this strawberry farm were to continue, future work should aim to repeat the air sampling over several years and increase the sampling frequency from monthly to fortnightly or weekly. Although this was not feasible within the scope of this PhD, more frequent collections would enable better identification of repeatable seasonal patterns and improve understanding of how pathogen prevalence is influenced by environmental factors. Increased sampling intensity would also enhance the potential of AirSeq to provide early warning of pathogen presence before visible symptoms develop on plants.

A valuable future study would be to divide the collected airborne material for both sequencing and inoculation experiments. This would confirm the viability of the spores and provide further validation that the strains or species detected in the air are capable of causing infection in plants.

In this study, fungicide application had limited observable impact on either disease symptoms or levels of airborne inoculum, suggesting the pathogen populations had developed resistance to the chemicals used. To explore fungicide resistance, it may be useful to collect airborne spores and culture them in the lab, where their responses to different fungicides could be tested. Cultured isolates would also allow for deeper genome sequencing, potentially enabling the identification of resistance-conferring mutations.

Finally, a long-term goal is to guide fungicide application decisions based on AirSeq results. To evaluate this, a controlled trial could be designed with one greenhouse using AirSeq-informed management and another following standard practices. Comparing outcomes would help determine whether AirSeq reduces fungicide use while maintaining disease control.

The findings of this study highlight AirSeq as a promising tool for the early detection and surveillance of crop pathogens, with broad applications across cropping systems. While further refinement is needed to ensure the robustness and reliability of the method, the next chapter presents experiments aimed at improving the AirSeq pipeline and advancing it toward practical implementation.

Chapter 4

Protocol Refinement

4.1 Abstract

Establishing a reliable and unbiased protocol is essential in airborne eDNA studies to ensure that the identified microbiome communities are accurately captured and characterised. Each stage of the experimental process has the potential to introduce bias, making it crucial to understand and minimise the impact of methodological choices. In this chapter, a series of field and laboratory experiments were conducted to refine the AirSeq protocol for detecting airborne plant pathogens. Sampler comparisons identified the InnovaPrep Cub as the most suitable device due to its consistent taxonomic detection, cost-effectiveness and portability. Control experiments demonstrated the importance of including both negative and positive controls. A maize pollen dispersal study highlighted the influence of distance, wind direction and environmental variability on detection rates. Laboratory tests confirmed that short-term filter storage at –80°C is feasible but may reduce DNA yield, and that a 20-second bead beating step is sufficient for efficient cell lysis without introducing fragmentation bias. The findings from these experiments directly informed the design of long-term pathogen monitoring described in Chapter 5, where the optimised AirSeq protocol was applied under agricultural field conditions.

4.2 My contributions

I carried out all laboratory work for the filter storage, syringe filtering, bead beating duration and control experiments. The sampler comparison samples were collected by Dr Heavens and the distance from maize source samples were collected by the entire lab group. DNA extraction was carried out by Dr Heavens and myself, and sequencing was completed by Dr Heavens. I conducted all bioinformatic work independently, including data cleaning, preparation and analysis. Additionally, I used ChatGPT (GPT-5) [273] to refine my code and to enhance the grammar and flow of my writing.

4.3 Introduction

Understanding the composition of airborne microbial communities is crucial in fields such as plant pathology, allergen monitoring and public health surveillance. Accurate detection of fungal pathogens in air samples can support early warning systems and inform timely interventions.

In microbiome research, each stage of the workflow, including sample collection, processing, sequencing and analysis, can influence which taxa are detected. Consequently, several studies have proposed best practice guidelines to reduce bias and improve reproducibility across investigations [40, 92, 180]. For sample collection, this includes maintaining consistent protocols, such as using the same sampler, collecting samples at the same time of day, and recording relevant metadata. DNA extraction protocols should minimise contamination through sterilisation, apply bead beating carefully to preserve nucleic acid length, ensure uniform storage conditions, and incorporate both positive and negative controls.

While existing best-practice recommendations provide valuable guidance for airborne metagenomic studies, they are not specifically designed for the rapid identification of fungal plant pathogens. For instance, some DNA extraction protocols can take two to three days to complete [180], which is impractical in a plant pathology context where timely detection is critical. This work therefore builds on those broader guidelines while addressing the specific demands of detecting fungal taxa from airborne samples. The objective is to build on an existing protocol that has been shown to successfully identify airborne plant pathogens [124], in order to ensure it recovers a taxonomically diverse and unbiased microbial community, including rare taxa, and enables accurate species-level identification using WGS data.

Despite growing interest in airborne microbiome research, no studies were identified that systematically evaluate the entire workflow from air sampling through to sequencing and bioinformatic analysis. Existing work often focuses on isolated components of the process, such as comparing air samplers solely on the basis of DNA yield [250], without considering downstream sequencing quality or taxonomic resolution. Similarly, while some studies have examined the ability of passive traps to capture microorganisms and spores at varying distances from a biological source [135, 154], little attention has been paid to the detectability of DNA when using active air samplers.

Within the domain of DNA extraction of airborne samples, most research centres on comparing commercial kits or different solutions [92], with limited consideration of the other factors relevant to airborne material, such as filter handling, storage conditions, or

the influence of mechanical lysis speed and duration on recovery. This chapter therefore provides a comprehensive examination of the air microbiome pipeline, from optimising sampling methods, through to refining DNA extraction approaches from airborne samples.

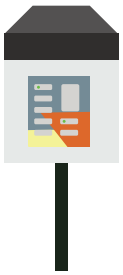
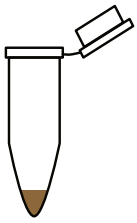
To compare air samples collected across different locations and times, it is essential to use a protocol that is both reliable and minimises bias. Reliability involves consistently capturing sufficient DNA for sequencing and ensuring that replicate samples produce comparable results. Minimising bias means avoiding preferential recovery, extraction or sequencing of particular taxonomic groups, species or cell types. An effective protocol should therefore capture, as closely as possible, the actual composition of airborne organisms present at the time of sampling. This allows researchers to interpret differences between samples as genuine ecological variation, rather than artefacts introduced during sample collection or processing.

While the aim is to develop a protocol that is as unbiased as possible, it is important to acknowledge that no method can completely eliminate bias. Both technical and biological factors contribute to this challenge. Technical factors include the characteristics of the sampler filter substrate, which may preferentially collect particles of certain sizes, and the limitations of current reference databases, which can prevent accurate taxonomic identification of sequenced reads. Biological factors involve differences in cell wall structure, which can affect DNA extraction efficiency, as well as variation in genome copy number, which may artificially inflate the apparent abundance of some species in the dataset.

4.3.1 Overview of the experimental design

To meet these aims, a series of experiments were conducted to refine each stage of the AirSeq protocol [124]. These experiments can be broadly divided into two categories: those carried out in the field and those undertaken in the laboratory. Table 4.1 summarises this division. This chapter presents the methods and results of these experiments and concludes with a set of proposed best-practice guidelines.

Table 4.1: Overview of where the different experiments sit within the pipeline

In the Field	**In the Laboratory**
	
• Sampler Comparison • Distance from maize source	• Filter Storage • Mechanical Lysis • Negative & Positive controls

4.3.2 Protocol refinement experiments based in the field

One set of the experiments were based around field collected samples, these included a sampler comparison and a distance-from-source experiment.

4.3.2.1 Differences between air samplers

As discussed in Chapter 2, air samplers vary in their particle size collection efficiency and have been shown to capture different airborne microbial community profiles, even when deployed simultaneously at the same location [155]. It was therefore necessary to evaluate a range of commercially available air samplers to determine the most suitable option for detecting airborne plant pathogens. Five different samplers (listed in Table 4.4) were tested in a city and agricultural location and assessed based on the diversity of species collected, and whether variation in the detected community was more strongly influenced by sampler type or sampling location. Based on these results, along with considerations of cost and portability, the InnovaPrep Cub was selected for use in all subsequent experiments.

4.3.2.2 Maize pollen dispersion

The distance-from-source experiment was designed to assess how proximity to a biological source - in this case, maize pollen - influences detection rates, and to examine the role of wind speed and direction in shaping airborne dispersal and sampler performance. Maize was selected as it is not a commonly grown crop in the UK and the pollen is wind-dispersed and therefore detectable with an air sampler.

Samplers were arranged in a V-shaped layout and at 10- and 100-metre grid spacing surrounding a flowering maize plot. The aim was to determine the proportion of maize DNA detectable at increasing distances under field conditions. Results showed that AirSeq can detect airborne maize DNA at distances up to 100 metres, with higher proportions detected at 10 metres. Detection rates were strongly affected by environmental factors, including wind speed, wind direction, and the abundance of source material, all of which influenced the proportion of maize DNA captured in each sample.

Maize plants are monoecious, containing both male and female reproductive structures. The male component, the tassel located at the top of the plant, produces pollen within the anthers, which is released, typically under dry conditions, from the tips of the anthers and has a typical volume-equivalent diameter of 90-100 μm [24]. The female part, the silk, captures pollen grains, which then travel down to fertilise the ovules and form kernels on the cob. Pollen release usually occurs over a five to eight day period, with peak shedding in the mid-morning [381]. Each tassel is estimated to produce approximately 25 million pollen grains [24], although environmental conditions, particularly temperature and humidity, strongly influence the timing and intensity of release [381].

Previous studies of maize pollen dispersal have primarily used passive samplers combined with microscopy [154, 174], or measured the proportion of successfully fertilised kernels resulting from cross-pollination between distinct maize varieties planted at fixed distances [24, 184]. These experiments consistently show that the majority of pollen is deposited close to the source. For instance, pollen abundance at 30 m has been reported as only 10 % of that at 1 m [184], and negligible deposition has been detected beyond 100 m [21]. Similarly, Jarosz et al. estimated that around 95 % of maize pollen settles within

10 m [174], while earlier work found that just 5 % of pollen remained airborne at 60 m compared with levels at 1 m [303]. Together, these findings highlight that most maize pollen is deposited within tens of metres, though a small fraction can persist and disperse over much longer distances.

The experiments presented here aimed to study these shorter dispersal distances ($<$ 100 m) using the AirSeq method, incorporating high-throughput short-interval sampling and WGS, to assess whether similar dispersal patterns can be detected using molecular approaches and to understand the sensitivity of the AirSeq approach.

4.3.3 Protocol refinement experiments based in the laboratory

The laboratory experiments described in this chapter were designed to evaluate key components of the airborne eDNA extraction and sequencing protocol. A range of strategies was employed to assess performance under different conditions. To investigate the effect of storage conditions on DNA recovery, the filter storage experiment was conducted using field-collected samples. These samples reflect the typical DNA concentrations and community compositions found in airborne environmental collections. In parallel, a mechanical lysis experiment was carried out using a mock community. This controlled approach enabled a high number of replicates and, due to the known taxonomic composition, allowed for accurate assessment of DNA recovery and relative abundance for each species. Additionally, negative controls were included at various stages of the sampling and DNA extraction process to identify potential sources of contamination, while positive controls were incorporated during sequencing to quantify the extent of barcode cross-talk.

4.3.3.1 Use of mock and field collected samples

While the mechanical lysis experiments used mock communities, all other experiments were based on field collected samples. Table 4.2 outlines the strengths and limitations of each approach. Mock communities provide a valuable benchmark, enabling direct comparisons between expected and observed taxonomic profiles, making them ideal for evaluating extraction efficiency. In contrast, field samples offer ecological realism and help uncover environmental and technical sources of variation. Together, these approaches provide complementary insights, contributing to the development of a robust and unbiased airborne DNA workflow.

4.3.3.2 Effect of filter storage on DNA yield

The filter storage experiment aimed to test whether storing filters at –80°C affects DNA yield, as appropriate storage conditions are frequently emphasised in microbiome best-practice guidelines. To address this, paired filters were compared: one processed immediately and the other stored at –80°C for three weeks before DNA extraction. Freezing was found to reduce overall DNA yield, although this effect was less pronounced when initial DNA quantities were higher.

4.3.3.3 DNA extraction with mechanical lysis

Mechanical lysis by bead beating requires balancing sufficient cell disruption with the risk of excessive DNA fragmentation, which can hinder taxonomic identification. To explore

Table 4.2: Comparison of mock community and field-collected samples for protocol refinement

Aspect	Mock Community	Field-Collected Sample
Complexity	Contains a known set of species (usually ~ 10), useful for evaluating protocol performance.	High complexity, includes thousands of unknown taxa at variable concentrations.
Reproducibility	Highly reproducible; allows for consistent comparisons between runs.	Sample composition varies by time, location and environmental conditions.
Cost and Access	Commercial options are expensive; homemade versions may lack consistency.	Collection can be performed at any time, with costs limited to consumables, though quality may be variable and yields low in some seasons.
Experimental Use	Best for testing specific protocol steps like DNA extraction or mechanical lysis.	Essential for testing collection methods and real-world sample conditions.
Limitations	Low diversity, limited ecological relevance.	Unknown true composition makes it hard to assess method accuracy.
Used In This Study	Mechanical lysis (bead beating).	Sampler comparison, distance from maize source, filter storage.

this balance, different bead beating durations were tested using a defined mock community. The results indicated that shorter bead beating times provided the best compromise, achieving effective lysis while maintaining longer DNA fragments suitable for downstream classification.

4.3.3.4 Importance and inclusion of control samples

A series of negative control experiments were conducted to identify potential sources of contamination across the AirSeq workflow (Figure 4.1). Because airborne microbial sampling involves low-biomass material, even minor contamination can be problematic. Controls were introduced at specific stages and carried through all subsequent steps to mimic standard sample handling. These experiments yielded very low DNA levels and showed no evidence of consistent contamination.

To assess potential barcode cross-talk during library preparation and sequencing, a control run was carried out using Lambda phage DNA alongside negative samples. Cross-talk was detected, with lambda reads appearing in the negative control. As a result, all subsequent experiments incorporated both a negative control (unused electrostatic filter) and a positive control (lambda DNA) to monitor the proportion of barcode cross-talk.

Together, these experiments informed the selection of DNA extraction conditions that maximise yield while preserving read quality and minimising taxonomic bias.

This chapter presents the methods and results of a series of protocol refinement experiments. Each experiment influenced the composition of the detected microbial community to some degree, underscoring the importance of careful optimisation at every stage. The overarching aim is to develop an air microbiome protocol that is both time- and cost-efficient, while ensuring the recovery of a taxonomically diverse community, including rare

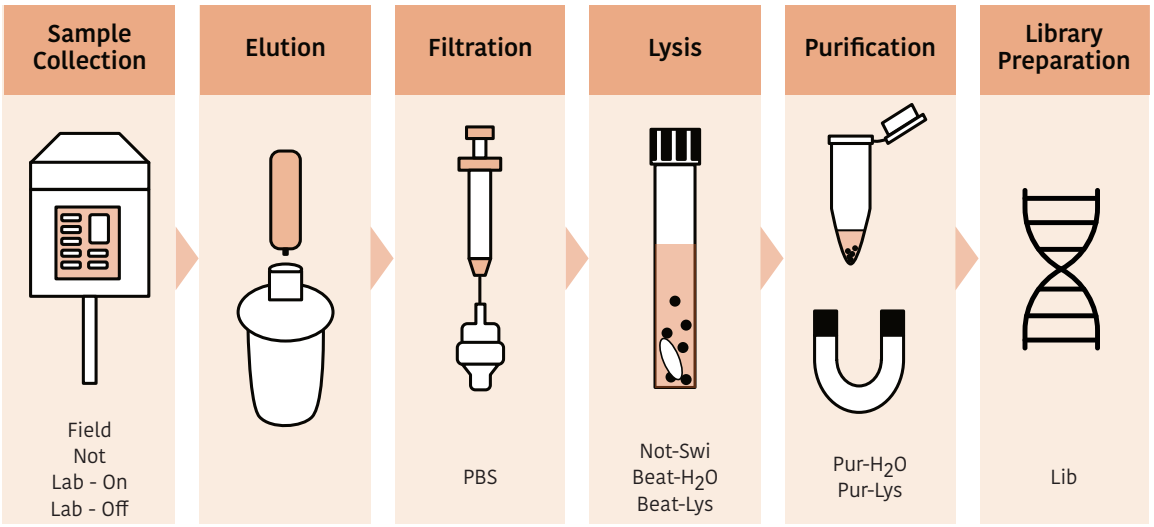


Figure 4.1: Schematic overview of the AirSeq protocol highlighting the points at which negative control samples were introduced. Each control was designed to simulate specific stages of the workflow, including sampling, elution, filtration, lysis, purification, and library preparation. The labels along the bottom correspond to the sample IDs.

taxa. The following sections detail the methods, outcomes and implications of each experiment, culminating in a proposed best-practice protocol for airborne microbial sampling and sequencing. The results from these experiments contribute to the broader aim of this thesis: to develop a reliable, field-deployable pipeline for airborne fungal pathogen detection.

4.4 Methods

4.4.1 AirSeq pipeline

All experiments followed a standardised AirSeq workflow encompassing sample collection, cell lysis, DNA extraction, isolation, and quantification, with minor adaptations for specific trials. An overview of the pipeline using the Cub sampler is shown in Figure 4.1, and experiment-specific modifications are described in the relevant subsections.

4.4.1.1 Sample collection

A range of air samplers were used across the experiments, with the Coriolis µ and InnovaPrep Cub being the most frequently employed.

For collections using the Coriolis µ (Bertin Technologies), 15 ml of sterile, nuclease-free water was added to a reusable collection cone, which was then attached to the sampler. Sampling was performed for 20 minutes at the maximum flow rate of 300 L/min. This collection time was selected to collect 6,000L's of air per collection. The collection fluid was immediately removed and syringe-filtered through a Swinny filter holder containing a 13 mm, 0.22 μm PVDF membrane (Durapore).

For collections using the Cub (InnovaPrep), the supplied electrostatic filter was fitted into the sampler. Sampling duration varied between experiments, but the default was a 60-minute collection at the maximum flow rate of 200 L/min. After sampling, particles collected onto the electrostatic filter were eluted using the manufacturer-supplied elution

cartridge, and the resulting fluid was passed through a 13 mm, 0.22 μm PVDF membrane using a Swinny filter holder and sterile syringe.

All PVDF membrane filters were placed on dry ice in 2 ml tubes for transport and transferred to a –80°C freezer upon arrival at the laboratory.

4.4.1.2 DNA extraction and isolation

In the laboratory, each membrane filter was placed into a 2 ml tube containing 250 mg of beads and 125 μl of a guanidinium-based lysis solution. Samples were homogenised using a SuperFastPrep-2™ beater (MP Biomedicals) at speed setting 20 for 20 seconds, followed by centrifugation.

DNA was isolated from the resulting lysate using a 1x magnetic bead clean-up. The lysate was transferred to a 1.5 ml Eppendorf tube containing magnetic beads, vortexed, and incubated at room temperature. The tubes were then placed on a magnetic particle concentrator to pellet the beads. The supernatant was discarded, and the pellet was washed twice with 70% ethanol. After removing residual ethanol by brief centrifugation and pipetting, DNA was eluted in sterile nuclease-free water at room temperature.

DNA yield was quantified using the Qubit High Sensitivity assay (Invitrogen, Thermo Fisher Scientific). 1 μl of DNA, 1 μl of dye, and 198 μl of buffer were combined in a 0.5 ml Qubit tube, incubated for two minutes, and placed on the Qubit fluorometer to obtain DNA concentration in μg/ml. For certain experiments, samples were subjected to WGA prior to sequencing, following the protocol outlined in Chapter 3.

4.4.1.3 Library preparation and sequencing

Library preparation was performed using either the Rapid Barcoding Kit (SQK-RBK110) or the Ligation Sequencing Kit (SQK-LSK114) from ONT, following the manufacturer's protocols.

Sequencing was carried out on ONT platforms using R9.4.1 or R10 flow cells (Flongle, MinION or PromethION). Flow cells were quality checked using MinKNOW software. If the number of active pores was sufficient, the flow cell was primed with the recommended volume of priming mix, incubated for 5 minutes, and topped up with an additional volume before sample loading.

Libraries were prepared with sequencing buffer, loading beads, nuclease-free water, and the final DNA library. Samples were loaded via the SpotON port (MinION/PromethION) or directly onto Flongle flow cells, as appropriate. Sequencing runs were executed using MinKNOW's real-time high accuracy basecalling mode, with a quality score filter set to 9.

4.4.1.4 Bioinformatic analysis

Following sequencing, the read data was copied to the HPC for bioinformatic analysis and MARTi [281] was used for initial taxonomic assignment. MARTi's BLAST-based analysis is initiated through the command line, then visualised through a web browser front-end. MARTi utilises the NCBI nt BLAST database and assigns reads to taxa through a LCA algorithm that places a read to the lowest taxonomic level consistent with the set of good BLAST hits.

Unless otherwise stated, MARTi was run against the NCBI nt database (nt_20240305) with the following parameters: `LCAMaxHits = 100`, which limits the maximum number of

BLAST hits considered in LCA assignment; `LCAMinIdentity = 85`, meaning only hits with at least 85% sequence identity were retained; `LCAMinLength = 150`, requiring a minimum alignment length of 150 bp; and `LCAMinReadLength = 200`, so reads shorter than 200 bp were excluded from analysis. After classification, results were downloaded from the front end with the LCA minimum support set to 0.1%. This threshold means that taxa representing less than 0.1% of the classified reads were not reported at their most specific assignment. Instead, such reads were reassigned higher up the taxonomic tree until they fell into a group exceeding the cutoff.

4.4.2 Experimental modifications

Table 4.3 outlines the broad differences between the different experimental protocols such as the sampler used and type of library preparation kit used.

Table 4.3: Overview of AirSeq protocol experiments, including the sampler and sequencing method used, and basecalling model used.

Experiment Name	Sampler Used	Controls	Library Prep	Flow Cell	Basecalling accuracy
Sampler comparison	All	None	Rapid	MinION	High
Distance from maize source	Cub	Yes	Ligation	Flongle, MinION & PromethION	SUP
Filter storage	Coriolis μ	N/A	N/A	N/A	N/A
Bead beating	None	No	Rapid	MinION	High
Negative controls	Cub	All	Ligation	Flongle	High
Positive control	None	Yes	Ligation	Flongle	High

4.4.2.1 Field experiments

Sampler comparison To assess the impact of sampler type on airborne microbial diversity, five air samplers were compared: the Coriolis µ, Coriolis Compact, InnovaPrep Bobcat, InnovaPrep Cub, and Smart Air Sampler System (SASS) 4100. Sampler specifications are provided in Table 4.4.

Each sampler was operated at its maximum flow rate for 25 and 50 minutes in two locations the garden at Natural History Museum (NHM) (London) and an agricultural field at Church Farm (Norwich). The sampling durations, flow rates, and total air volumes for each collection are detailed in Table 4.5. The NHM samples were collected on 9 June 2022, while the Church Farm samples were collected the following week. Images of both collection locations are shown in Figure 4.2. The experimental setup was otherwise identical to the general protocol.

Filters obtained from the sampler comparison experiment were processed following the standard protocol.

Library preparation was carried out using the Rapid Barcoding Kit (SQK-RBK110.96). Where possible, 5 ng of DNA was used per sample; if DNA yield was <5 ng, the entire

Table 4.4: Description of different air samplers

Sampler	**How it works**	**Particle collection size**	**Flow rate (L/min)**	**Run time**	**Cost (approx.)**
Coriolis Compact (Bertin)	Dry cyclone sampler; air particles are centrifuged and deposited on a dry surface.	500 nm – 10 µm	50	Up to 8 hours	$16,000
Coriolis µ (Bertin)	Wet cyclone sampler; air particles are aspirated and trapped in liquid via vortex.	Collection efficiency of 50% efficiency for 0.5 µm particles	100 – 300	Up to 6 hours	$18,000
ACD-200 Bobcat (InnovaPrep)	Dry filter sampler. Air particles are impacted onto an electret filter; particles are then eluted into a liquid using a specialised kit.	0.1 µm – 10 µm	100 or 200	Up to continuous (when plugged in)	$10,000, plus cost of disposable filters
AirPrep™ Cub Sampler (InnovaPrep)	Same as Bobcat: dry filter-based sampling with elution kit. Smaller and more portable form.	0.1 µm – 10 µm	50, 100, or 200	Continuous, 2 h, 1 h, or 30 min options	$1,500 – $2,500, plus cost of disposable filters
SASS 4100	High-volume, two-stage dry filter sampler; concentrates particles using centrifugal/impaction principles, and collects them on an electret filter	Collection efficiency of 50% for 0.5 µm particles, with bioaerosol filter	4000	Up to several days	$10,000, plus cost of disposable filters

Table 4.5: Collection duration, flow rate, and total air volume sampled for each air sampler and sample ID.

Air Sampler	**Collection duration (min)**	**Flow rate (L/min)**	**Total air sampled (L)**	**Sample ID**
Coriolis Compact	25	50	1,250	25_Compact_1, 25_Compact_2
	50	50	2,500	50_Compact_1, 50_Compact_2
Coriolis µ	25	300	7,500	25_µ_1, 25_µ_2
	50	300	15,000	50_µ_1, 50_µ_2
InnovaPrep Bobcat	25	200	5,000	25_Bobcat_1, 25_Bobcat_2
	50	200	10,000	50_Bobcat_1, 50_Bobcat_2
InnovaPrep Cub	25	200	5,000	25_Cub_1, 25_Cub_2
	50	200	10,000	50_Cub_1, 50_Cub_2
SASS 4100	25	4,000	100,000	25_Sass_1, 25_Sass_2
	50	4,000	200,000	50_Sass_1, 50_Sass_2

(a) NHM site (London) (b) Church Farm site (Norwich)

Figure 4.2: Pictures from the sampler comparison collection sites, taken on a different day to the sampler comparison data presented here but representative of the sites.

extract was used. Twelve samples had <5 ng DNA, including eight collected using the Coriolis Compact. Libraries were sequenced on two R9.4.1 flow cells (20 samples each), with samples grouped by location.

Once sequenced taxonomic assignment was carried out using MARTi (v0.9.26) with `LCAMinReadLength = 500`. The resulting taxon count data were exported from the front end as read counts per taxon and normalised by the total number of reads per barcoded sample to account for variation in sequencing depth. Ordination analyses at the genus level, including Principal Coordinate Analysis (PCoA) and t-SNE, were then performed using a custom Python script (`t-sne_samp_comp.py`) that applies the scikit-learn and scikit-bio libraries to compute Bray–Curtis dissimilarity, plot a PCoA of the first two principal coordinates, and generate a t-SNE projection.

In R, the packages `dplyr`, `ggplot2`, `phyloseq`, and others were used to generate a range of plots across several scripts using the read count data. These included DNA yield visualisations (`DNA_yield_graph.R`), stacked bar plots of taxonomic abundance (`Stacked_Phylum_Bars_Phyloseq.R`), the proportion of unique taxa per sampler (`taxa_proportion_unique.R`), and the top taxa per sample (`top_taxa_graphs.R`).

All scripts used in the analysis of the sampler comparison data are available at: `https://github.com/Mia-FGB/Sampler_Comparison_Scripts`.

Distance from maize source Sampling took place at Church Farm from 19 - 29 August 2024, images of the sampling site are shown in Figure 4.3. A 10 m × 10 m maize plot was established in an agricultural field that had been left fallow that year, resulting in the plot being surrounded by weeds or bare soil. To extend the pollination window, two maize varieties (*Jakleen* and *Prospect*) were sown in May. Sample collection began once the plants reached the flowering stage, with 1-hour samples collected using Cub samplers and processed according to the standard AirSeq protocol. Maize was selected for this experiment as it is not commonly grown in the UK or Norfolk and the pollen is known to be wind-dispersed and detectable with air sampling. During the course of this experiment an additional maize plot was identified 1.74 km away, shown in Figure 4.5.

Environmental data for the duration of the experiment were obtained from the Church Farm weather station, including 10-minute measurements of mean wind speed and direction, as well as maximum gust speed.

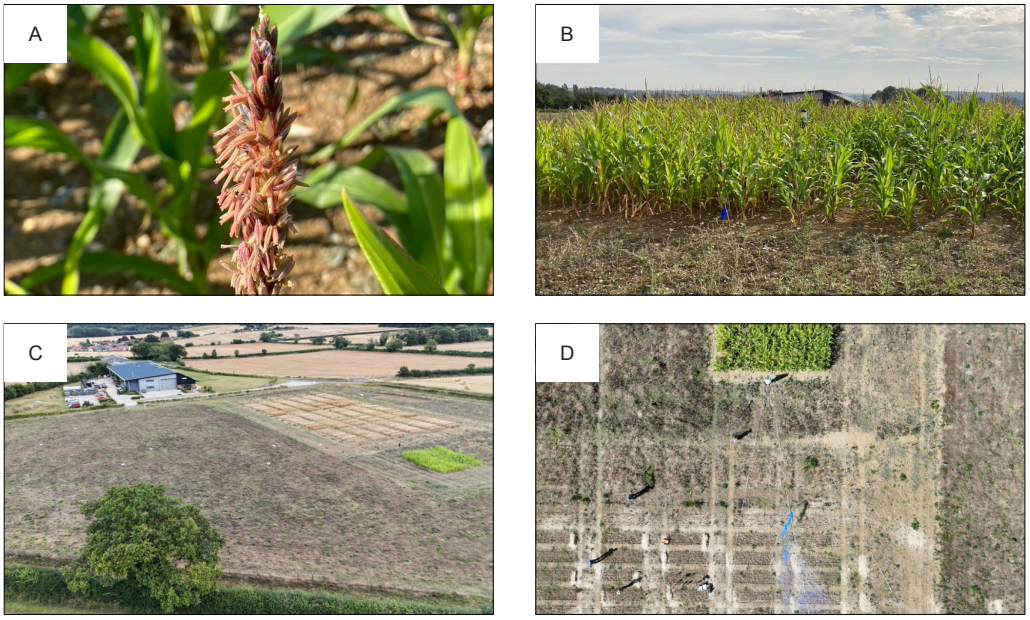


Figure 4.3: Representative images of the maize experimental plot: (A) pollinating maize tassel, (B) maize plot at ground level, (C) aerial view of trial site, (D) aerial view during smoke release.

To assess spatial distribution, nine Cub samplers were deployed simultaneously in one of three configurations: a 10 m grid, a 100 m grid, or a V-shape positioned downwind of the maize plot. Each configuration included a sampler placed at the centre of the plot. The layout of each sampling design is shown in Figure 4.4.

For the 10 m grid, a tape measure was used to ensure precise placement, with each sampler 10 metres from the edge of the maize plot. The 100 m grid was mapped using GPS coordinates.

For the V-shape configuration, a smoke grenade was used to determine wind direction, and the direction of the V was adjusted to match the direction of the smoke (and wind). Figure 4.3D shows an aerial view of the smoke release. The length of the V was determined by the field boundary, extending as far as possible before reaching a hedge or fence, with three rows of samplers distributed evenly along this length.

For all configurations, GPS coordinates and what3words data were recorded to document sampler locations. These coordinates were imported into Google Earth Pro to visualise the locations from a satellite view (Figure 4.5). Details on sample collection times, experimental design, and distances between samplers can be found in Table 4.6.

Table 4.6: Sample Collection Details

Layout	**Sample ID**	**Date**	**Start Time**	**Dist. from plot (m)**	**Layout**	**Sample ID**	**Date**	**Start Time**	**Dist. from plot (m)**
10M	190824_10_1	19/08/24	11:14	0	V	270824_V_1	27/08/24	12:21	0
10M	190824_10_2	19/08/24	11:14	10	V	270824_V_2	27/08/24	12:21	17
10M	190824_10_3	19/08/24	11:14	10	V	270824_V_3	27/08/24	12:21	17
10M	190824_10_4	19/08/24	11:14	10	V	270824_V_4	27/08/24	12:21	34
10M	190824_10_5	19/08/24	11:14	10	V	270824_V_5	27/08/24	12:21	34
10M	190824_10_6	19/08/24	11:14	10	V	270824_V_6	27/08/24	12:21	34

Layout	Sample ID	Date	Start Time	Dist. from plot (m)	Layout	Sample ID	Date	Start Time	Dist. from plot (m)
10M	190824_10_7	19/08/24	11:14	10	V	270824_V_7	27/08/24	12:21	50
10M	190824_10_8	19/08/24	11:14	10	V	270824_V_8	27/08/24	12:21	50
10M	190824_10_9	19/08/24	11:14	10	V	270824_V_9	27/08/24	12:21	50
100M	190824_100_1	19/08/24	13:08	0	10M	280824_10_1	28/08/24	09:00	0
100M	190824_100_2	19/08/24	13:08	100	10M	280824_10_2	28/08/24	09:00	10
100M	190824_100_3	19/08/24	13:08	100	10M	280824_10_3	28/08/24	09:00	10
100M	190824_100_4	19/08/24	13:08	100	10M	280824_10_4	28/08/24	09:00	10
100M	190824_100_5	19/08/24	13:08	100	10M	280824_10_5	28/08/24	09:00	10
100M	190824_100_6	19/08/24	13:08	100	10M	280824_10_6	28/08/24	09:00	10
100M	190824_100_7	19/08/24	13:08	100	10M	280824_10_7	28/08/24	09:00	10
100M	190824_100_8	19/08/24	13:08	100	10M	280824_10_8	28/08/24	09:00	10
100M	190824_100_9	19/08/24	13:08	100	10M	280824_10_9	28/08/24	09:00	10
V	200824_V1_1	20/08/24	10:46	0	100M	280824_100_1	28/08/24	10:16	0
V	200824_V1_2	20/08/24	10:46	25	100M	280824_100_2	28/08/24	10:16	100
V	200824_V1_3	20/08/24	10:46	25	100M	280824_100_3	28/08/24	10:16	100
V	200824_V1_4	20/08/24	10:46	55	100M	280824_100_4	28/08/24	10:16	100
V	200824_V1_5	20/08/24	10:46	55	100M	280824_100_5	28/08/24	10:16	100
V	200824_V1_6	20/08/24	10:46	55	100M	280824_100_6	28/08/24	10:16	100
V	200824_V1_7	20/08/24	10:46	85	100M	280824_100_7	28/08/24	10:16	100
V	200824_V1_8	20/08/24	10:46	85	100M	280824_100_8	28/08/24	10:16	100
V	200824_V1_9	20/08/24	10:46	85	100M	280824_100_9	28/08/24	10:16	100
V	200824_V2_1	20/08/24	11:59	0	V	280824_V_1	28/08/24	12:10	0
V	200824_V2_2	20/08/24	11:59	25	V	280824_V_2	28/08/24	12:10	12
V	200824_V2_3	20/08/24	11:59	25	V	280824_V_3	28/08/24	12:10	12
V	200824_V2_4	20/08/24	11:59	55	V	280824_V_4	28/08/24	12:10	24
V	200824_V2_5	20/08/24	11:59	55	V	280824_V_5	28/08/24	12:10	24
V	200824_V2_6	20/08/24	11:59	55	V	280824_V_6	28/08/24	12:10	24
V	200824_V2_7	20/08/24	11:59	85	V	280824_V_7	28/08/24	12:10	36
V	200824_V2_8	20/08/24	11:59	85	V	280824_V_8	28/08/24	12:10	36
V	200824_V2_9	20/08/24	11:59	85	V	280824_V_9	28/08/24	12:10	36
10M	210824_10_1	21/08/24	09:56	0	10M	290824_10_1	29/08/24	09:00	0
10M	210824_10_2	21/08/24	09:56	10	10M	290824_10_2	29/08/24	09:00	10
10M	210824_10_3	21/08/24	09:56	10	10M	290824_10_3	29/08/24	09:00	10
10M	210824_10_4	21/08/24	09:56	10	10M	290824_10_4	29/08/24	09:00	10
10M	210824_10_5	21/08/24	09:56	10	10M	290824_10_5	29/08/24	09:00	10
10M	210824_10_6	21/08/24	09:56	10	10M	290824_10_6	29/08/24	09:00	10
10M	210824_10_7	21/08/24	09:56	10	10M	290824_10_7	29/08/24	09:00	10
10M	210824_10_8	21/08/24	09:56	10	10M	290824_10_8	29/08/24	09:00	10
10M	210824_10_9	21/08/24	09:56	10	10M	290824_10_9	29/08/24	09:00	10
100M	210824_100_1	21/08/24	11:35	0	100M	290824_100_1	29/08/24	10:19	0
100M	210824_100_2	21/08/24	11:35	100	100M	290824_100_2	29/08/24	10:19	100
100M	210824_100_3	21/08/24	11:35	100	100M	290824_100_3	29/08/24	10:19	100
100M	210824_100_4	21/08/24	11:35	100	100M	290824_100_4	29/08/24	10:19	100
100M	210824_100_5	21/08/24	11:35	100	100M	290824_100_5	29/08/24	10:19	100
100M	210824_100_6	21/08/24	11:35	100	100M	290824_100_6	29/08/24	10:19	100
100M	210824_100_7	21/08/24	11:35	100	100M	290824_100_7	29/08/24	10:19	100
100M	210824_100_8	21/08/24	11:35	100	100M	290824_100_8	29/08/24	10:19	100
100M	210824_100_9	21/08/24	11:35	100	100M	290824_100_9	29/08/24	10:19	100
10M	220824_10_1	22/08/24	11:26	0	V	290824_V_1	29/08/24	11:58	0
10M	220824_10_2	22/08/24	11:26	10	V	290824_V_2	29/08/24	11:58	30
10M	220824_10_3	22/08/24	11:26	10	V	290824_V_3	29/08/24	11:58	30
10M	220824_10_4	22/08/24	11:26	10	V	290824_V_4	29/08/24	11:58	60
10M	220824_10_5	22/08/24	11:26	10	V	290824_V_5	29/08/24	11:58	60
10M	220824_10_6	22/08/24	11:26	10	V	290824_V_6	29/08/24	11:58	60
10M	220824_10_7	22/08/24	11:26	10	V	290824_V_7	29/08/24	11:58	90
10M	220824_10_8	22/08/24	11:26	10	V	290824_V_8	29/08/24	11:58	90
10M	220824_10_9	22/08/24	11:26	10	V	290824_V_9	29/08/24	11:58	90

Layout	Sample ID	Date	Start Time	Dist. from plot (m)	Layout	Sample ID	Date	Start Time	Dist. from plot (m)
100M	270824_100_1	27/08/24	10:41	0	10M	Bing_10m_3	27/08/24	09:03	0
100M	270824_100_2	27/08/24	10:41	100	10M	Col_10m_7	27/08/24	09:03	10
100M	270824_100_3	27/08/24	10:41	100	10M	Dar_10m_6	27/08/24	09:03	10
100M	270824_100_4	27/08/24	10:41	100	10M	Eli_10m_2	27/08/24	09:03	10
100M	270824_100_5	27/08/24	10:41	100	10M	Geo_10m_4	27/08/24	09:03	10
100M	270824_100_6	27/08/24	10:41	100	10M	Kit_10m_1	27/08/24	09:03	10
100M	270824_100_7	27/08/24	10:41	100	10M	lyd_10m_9	27/08/24	09:03	10
100M	270824_100_8	27/08/24	10:41	100	10M	Mar_10m_8	27/08/24	09:03	10
100M	270824_100_9	27/08/24	10:41	100	10M	Wic_10m_5	27/08/24	09:03	10

Following sample collection, filters were stored at ambient air temperature and transported back to the laboratory. Upon arrival, each filter was eluted and processed through a Swinny filter, following the procedure described previously. For each experimental set, an electret filter negative control was processed in parallel. In addition, a lambda DNA positive control was included for each experimental set during library preparation to validate sequencing performance.

Samples were prepared using the native barcoding ligation kit (SQK-NBD114.9) following the standard ONT protocol. Two of the 10 m experiments, collected on 19 and 27 August, were sequenced on separate MinION flow cells. Subsequent sample sets were sequenced on two PromethION flow cells. Additionally, the negative and lambda control samples from the second week were pooled and sequenced independently on a Flongle flow cell.

Lambda DNA and electret filter negative controls were included in sequencing to assess barcode cross-contamination and environmental contamination, respectively. Read counts aligning to *Escherichia coli* were monitored across all samples to evaluate potential cross-sample spread of lambda control material. PromethION sequenced controls confirmed low-level but detectable barcode cross-talk in some experimental runs.

Following sequencing, data from MinION, PromethION, and Flongle runs were rebasecalled using Dorado (v0.7.2) with the super-accuracy model.

To verify that conducting analysis on 100,000 randomly subsampled reads yields equivalent results to using the full set of sequenced reads, three samples were selected for comparison. For each of these samples, MARTi was run on both the total sequenced reads and a 100,000 read subsample. The absolute difference in the proportion of *Zea* genus reads per 100,000 analysed reads was then calculated. Subsampling was performed with the script `subsample_reads.sh`, available in the GitHub repository `https://github.com/Mia-FGB/hpc-scripts`. This script calls a Perl utility (`subsample_single.pl`), written by Dr Leggett and also available in the repository.

Following validation, all experimental samples were randomly subsampled to 100,000 reads per sample to ensure consistent comparisons across experiments and to reduce computational requirements. Where samples contained fewer than 100,000 reads, all available reads were retained for subsequent analyses; this primarily applied to the negative control samples.

Taxonomic assignment was performed using MARTi and the LCA cut-off was set to 0, the taxa count read table was used to evaluate the abundance of reads classified as *Zea* (maize genus) across time points and distances from the pollen source.

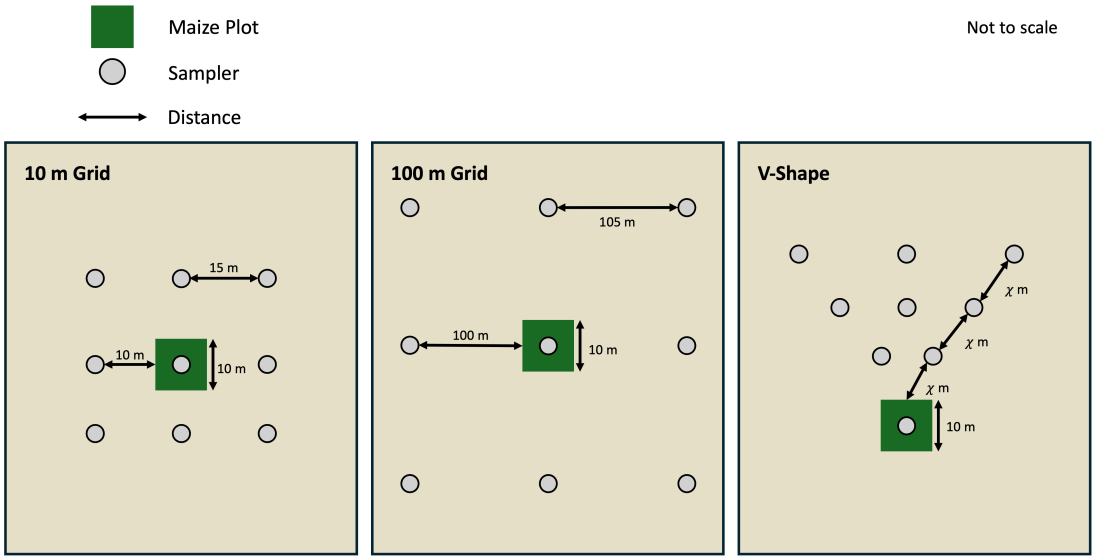


Figure 4.4: Figure showing the 3 different experimental sampler layouts

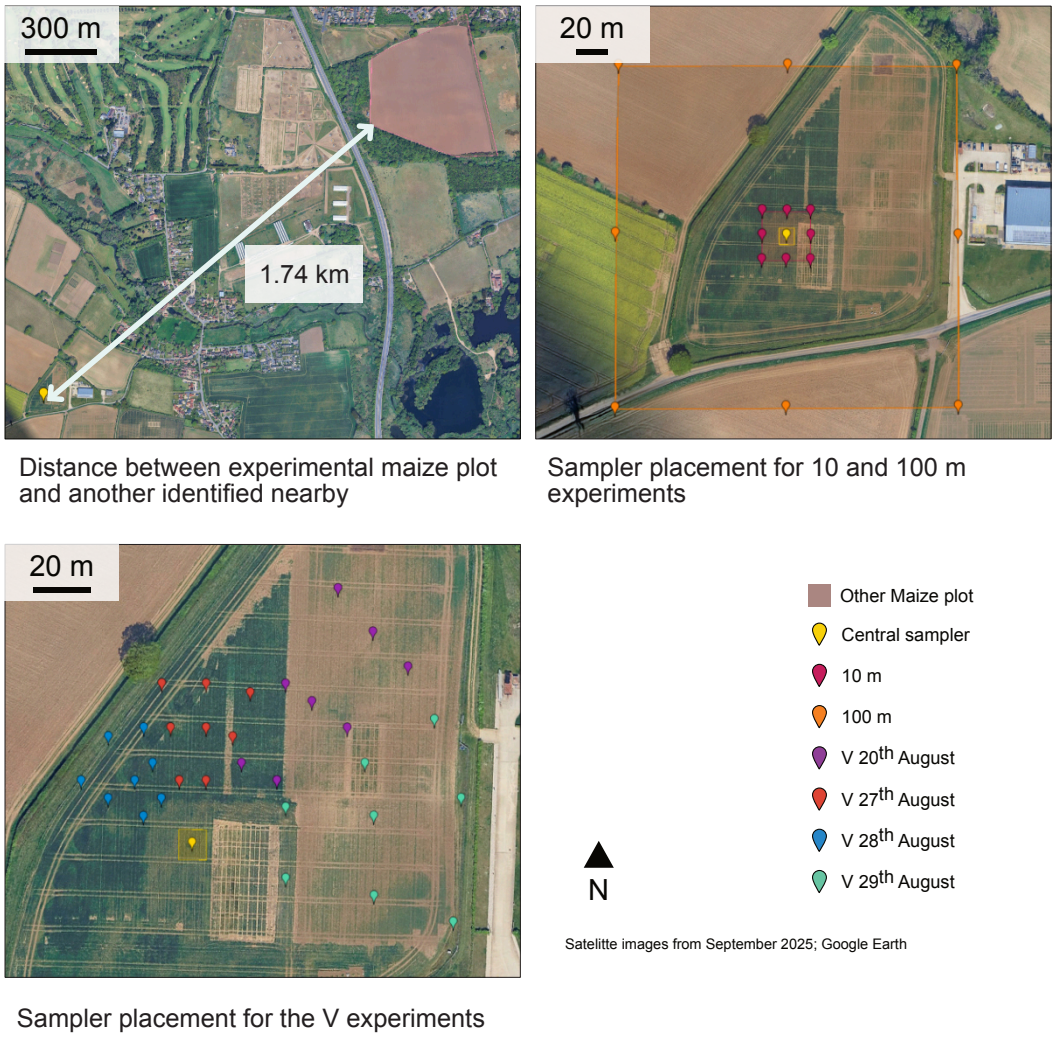


Figure 4.5: Satellite view of Church Farm showing the sampling locations for the maize experiments and a nearby maize plot. Image captured in September 2025 taken from Google Earth Pro. Locations marked and labelled by the author.

Additionally the relative abundance of the highest surrounding sampler compared with the central sampler was calculated for each experiment, using the following equation (4.1)

$$\text{Relative Abundance (\%)} = \left(\frac{\text{Highest surrounding sampler}_{ZeaHP100k}}{\text{Central sampler}_{ZeaHP100k}} \right) \times 100 \qquad (4.1)$$

All data analysis was performed in Python using the `pandas`, `seaborn`, and `matplotlib` libraries. The script `maize_analysis_figures.ipynb` was used to generate a range of visualisations, including bar charts showing how the proportion of *Zea*-aligned reads varied by time, date, and experimental condition; heatmaps displaying *Zea* read density (per 100k reads) across experimental layouts; and a scatter plot with a fitted curve examining the relationship between *Zea* read counts and distance from the maize source. Control data were visualised using the script `maize_control_analysis.ipynb`. Weather data were processed into wind vector plots illustrating average wind direction and speed during each collection period, using the script `weather_data_analysis.ipynb`. All scripts used in this study are available in the following GitHub repository: `https://github.com/Mia-FGB/maize_analysis`

4.4.2.2 Laboratory experiments

Filter storage To determine the effect of filter storage prior to processing, 4 samples (40 minutes each) were collected using the Coriolis µ outside the Earlham Institute (Norwich). The collection fluid from each sample was divided into two, and syringe filtered separately. From each sample one PVDF filter was processed immediately, while the other was stored at -80°C for 27 days. All samples were processed using the general extraction protocol, and DNA yield was assessed using the Qubit assay. The DNA yield data was then analysed in R to plot a bar chart comparing the fresh and frozen sample yields.

Length of mechanical lysis (bead beating) To evaluate the impact of bead-beating duration on DNA yield and read length, the ZymoBIOMICS Microbial Community Standard (Zymo Research) was used as a defined input. It comprises eight bacterial and two yeast species in known proportions, including both easy-to-lyse Gram-negative bacteria and tough-to-lyse Gram-positive bacteria and yeasts. The theoretical composition is shown in Table 4.7, the data is from [413].

Table 4.7: Composition of the ZymoBIOMICS Microbial Community Standard.

Species	**Group**	**Theoretical Proportion (%)**	**Genome Size (Mb)**
Bacillus subtilis	Gram-positive	12	4.045
Escherichia coli	Gram-negative	12	4.875
Enterococcus faecalis	Gram-positive	12	2.845
Lactobacillus fermentum	Gram-positive	12	1.905
Listeria monocytogenes	Gram-positive	12	2.992
Pseudomonas aeruginosa	Gram-negative	12	6.792
Staphylococcus aureus	Gram-positive	12	2.730
Salmonella enterica	Gram-negative	12	4.760
Cryptococcus neoformans	Yeast (fungal)	2	18.9
Saccharomyces cerevisiae	Yeast (fungal)	2	12.1

10 μl of ZymoBIOMICS Microbial Community Standard (whole cell) was added to a 2ml tube containing 250mg PowerSoil beads and 125 μl of lysis solution. Samples were then subjected to bead beating for 20, 40, 60, 120, 180, 360, or 600 seconds at approximately 2,300 Cycles per minute (CPM) (speed setting 20). A single replicate was performed for each duration, except for the 20-second condition, which was repeated twice. The rest of the DNA extraction and quantification followed the general protocol.

Samples from the bead-beating duration experiment were prepared for sequencing using the Rapid Barcoding Kit (SQK-RBK110.96) and run on a MinION flow cell following the standard ONT protocol. Reads were aligned to a reference database of the ten species in the mock community using *minimap2* (v2.0), and those with a MQ <5 were removed.

To evaluate the impact of bead-beating duration, two metrics were calculated: the number of reads mapping to each species and the N50 read length. N50, defined as the read length at which 50% of the total sequenced bases are contained in reads of that length or longer, is a key indicator of read length distribution and critical for accurate taxonomic assignment in metagenomic sequencing. Calculations were performed using a combination of shell and Python scripts.

Read count and N50 data were then exported into R, where they were visualised as line and bar graphs using the `ggplot2` and `dplyr` packages to assess the effect of bead-beating duration on DNA yield and read length.

Control experiments

Negative controls To assess potential sources of contamination, a series of negative control protocols were implemented (Table 4.8, see also workflow diagram in Figure 4.1). Two repeats were collected of each experiment.

Six negative control samples were collected using the InnovaPrep Cub sampler, as illustrated in the sample collection stage of Figure 4.1. Two of these samples ("Field") were collected outside the Earlham Institute (Norwich) following the standard field protocol. The remaining four samples were collected within the laboratory where DNA extraction, library preparation, and sequencing were routinely performed. Of these, two samplers were activated during collection ("Lab-On"), and two remained inactive throughout the sampling period ("Lab-Off"). There were also two electret filters that were not put into the sampler but taken through the standard protocol ("Not").

Filters were eluted and syringe-filtered according to the standard protocol. At this stage, PBS controls were introduced by passing sterile PBS through a syringe filter containing a PVDF membrane, in place of a sample. The resulting PVDF membranes were placed into bead-beating tubes containing 250mg of beads and either 125μl lysis solution or nuclease-free water.

At the next stage the Not-Swi, Beat-H_2O, and Beat-Lys control types were introduced. Not-Swi samples contained a sterile PVDF membrane that had never been placed in a Swinny filter, but were combined with PowerSoil beads and lysis solution. Beat-H_2O and Beat-Lys samples included only nuclease-free water or lysis solution with Power Soil beads, and no membrane filter.

All tubes were bead-beaten for 20 seconds at speed setting 20. From each, a 100 μl aliquot of the resulting lysate was purified using magnetic beads. At this purification stage,

the Pur-H_2O and Pur-Lys controls were introduced by adding magnetic beads to 100 μl of water or lysis solution alone, with no preceding bead beating.

All of the above negative controls were processed in parallel and quantified using the Qubit High Sensitivity assay.

Table 4.8: Description of how each negative control was generated. Tick (✓) indicates standard material used

Sample ID	**Sample Collected**	**Syringe Filtration**	**In Lysis Tube**	**In Purification Tube**	**In Library Prep**
Standard	Cub sampler	Elution liquid through PVDF membrane	PVDF filter, lysis solution and PowerSoil beads	Lysis eluate and magnetic beads	Purified DNA
Field	Collected in field	✓	✓	✓	✓
On	Collected in lab (sampler on)	✓	✓	✓	✓
Off	Collected in lab (sampler off)	✓	✓	✓	✓
Not	Cub filter never placed in sampler	✓	✓	✓	✓
PBS		PBS through PVDF membrane	✓	✓	✓
Not-Swi			✓	✓	✓
Beat-H_2O			Water and beads	✓	✓
Beat-Lys			Lysis Solution and beads	✓	✓
Pur-H_2O				Water and magnetic beads	✓
Pur-Lys				Lysis Solution and magnetic beads	✓
Lib					Sterile Water

Samples were prepared for sequencing using the Ligation Sequencing Kit (SQK-LSK114). For the DNA repair and end-prep step, 9 μl of sample was combined with 15 μl of water, except for the "Lib" samples, which contained 24 μl of water only. Adapter ligation followed the standard protocol. Each sample was loaded onto an individual Flongle flow cell and sequenced independently to eliminate barcode cross-talk.

Following sequencing, raw reads were processed in MARTi (v0.9.15) using standard parameters to generate taxonomic assignments and read counts for each control. Outputs were downloaded and further processed in a Python environment. Data manipulation was performed with `pandas`, and visualisations were created with `seaborn` to summarise DNA yield, read counts, the proportion of *Homo sapiens* reads, and the ratio of classified to unclassified reads.

Comparing classified and unclassified read lengths from negative controls To examine differences in read length distributions between sequences classified and unclassified by the MARTi pipeline, the negative control sequencing data were analysed. Classified reads were defined as those assigned a taxonomic identity by MARTi, whereas unclassified reads were sequences that passed quality filtering but remained unassigned.

Unclassified reads were isolated using a custom Python script, which compared read IDs present in the FASTA output of filtered reads with those listed in the MARTi per-read classification summary. Reads absent from the classification summary were extracted from the corresponding FASTA files for each sample.

Read length statistics, including N50, were then calculated separately for classified and unclassified sets using the `get_contig_stats.pl` script (provided by Dr Leggett) alongside additional parsing tools. The resulting summaries were converted to CSV format for further analysis in a Jupyter notebook, where contig statistics (total length, mean length, N50) were compared visually on both linear and logarithmic scales.

Positive control To assess potential barcode cross-talk during library preparation in barcoded sequencing runs containing low DNA inputs, a targeted control experiment was designed. A Flongle sequencing run was conducted using three barcoded samples: purified Lambda phage DNA, a water control, and a filter negative control. The filter negative control was prepared in the same way as the "Off" negative control (Table 4.8).

DNA was then extracted from the eluate of the filter negative control and 50 μL of lambda DNA using the standard bead beating and purification method described above. A volume of 9 μL from each sample was used in the library preparation, with the ligation kit (SQK-NBD114.24) following standard ONT protocol.

Following sequencing, raw reads were mapped only to the *Escherichia* phage Lambda reference genome (NCBI accession: J02459) using *minimap2* (v2.0) to quantify lambda-derived reads in each barcode. Alignments with a MQ <5 were removed from downstream analysis. The proportion of lambda reads in the sequence data was summed and presented as a Table.

4.5 Results

This section presents the results of the experimental work undertaken to test and optimise the AirSeq pipeline. The findings are structured according to two phases of the workflow: field-based and laboratory-based. Within each phase, several experiments were conducted, as outlined in Table 4.1.

4.5.1 Field experiments

This section describes the results of the sample collection experiments, including comparisons between samplers and the effect of sampler distance from source on maize pollen detection.

4.5.1.1 Sampler comparison

To evaluate the performance of different air samplers within the AirSeq pipeline, five devices (Coriolis µ, Coriolis Compact, InnovaPrep Bobcat, InnovaPrep Cub, and SASS 4100) were compared at two sites. The analysis focused on DNA yield, taxonomic diversity and community composition of the samples. Table 4.5 contains the collection duration, flow rate and total volume of air sampled in each collection.

Insect contamination The five samples which had visible insects on the filter following filtration are listed in Table 4.9. The Coriolis µ sampler had the highest number of samples contaminated, 3 out of 8. One of each of the InnovaPrep samplers contained a visible insect, but the SASS and Coriolis Compact had no contaminated filters. The location and collection duration do not appear to have a large influence on the chance of collecting an insect, but these samples do have considerably elevated DNA yields, particularly 50_µ_2 from Church Farm.

Table 4.9: Table of the samples that were shown to contain insect contamination.

Air Sampler	**Location**	**Sample ID**	**Collection duration (min)**	**DNA yield (ng)**
Coriolis µ	Church Farm	50_µ_2	50	380
Coriolis µ	NHM	25_µ_2	25	72
Coriolis µ	NHM	50_µ_1	50	45
InnovaPrep Bobcat	Church Farm	25_Bobcat_2	25	50
InnovaPrep Cub	NHM	50_Cub_1	50	20.5

DNA yield Across samplers, DNA yield generally increased with air volume (Figure 4.6). The SASS 4100 consistently produced the highest total yields, reflecting its substantially greater flow rate (4,000 L/min). At 25 minutes it recovered an average of 30.1 ng, rising to 52.3 ng at 50 minutes. The Coriolis µ also performed well, with yields of 27.5 ng and 113.4 ng at 25 and 50 minutes, respectively, although several unusually high values were associated with insect contamination.

The InnovaPrep samplers (Bobcat and Cub) gave intermediate results. At 25 minutes, the Bobcat initially appeared to exceed the Cub (17.5 ng vs 3.5 ng), but this difference was largely driven by an insect-contaminated replicate; once excluded, the Bobcat mean was reduced to 6.7 ng. At 50 minutes, yields were more comparable (7.0 ng for Bobcat vs 11.4 ng for Cub), again with Cub values influenced by a contaminated replicate.

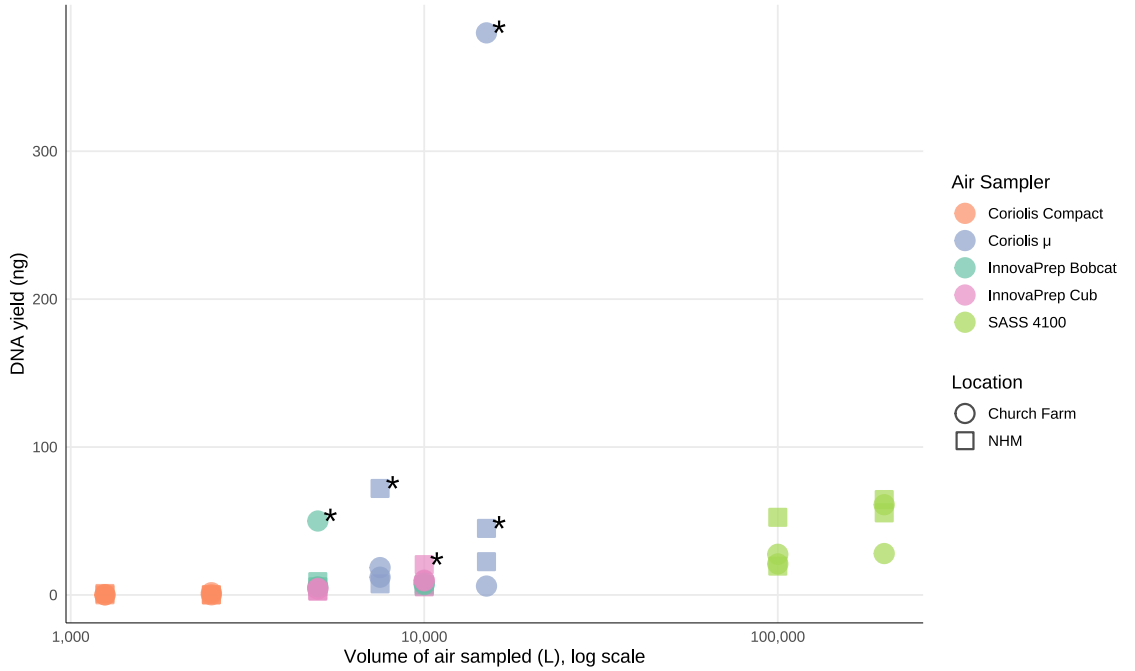


Figure 4.6: Scatter plots showing total DNA yield (ng) against volume of air sampled (L), coloured by air sampler and shaped by location (Church Farm: circle, NHM: square), samples with visible insect contamination on the PVDF filter are marked with an asterisk (*).

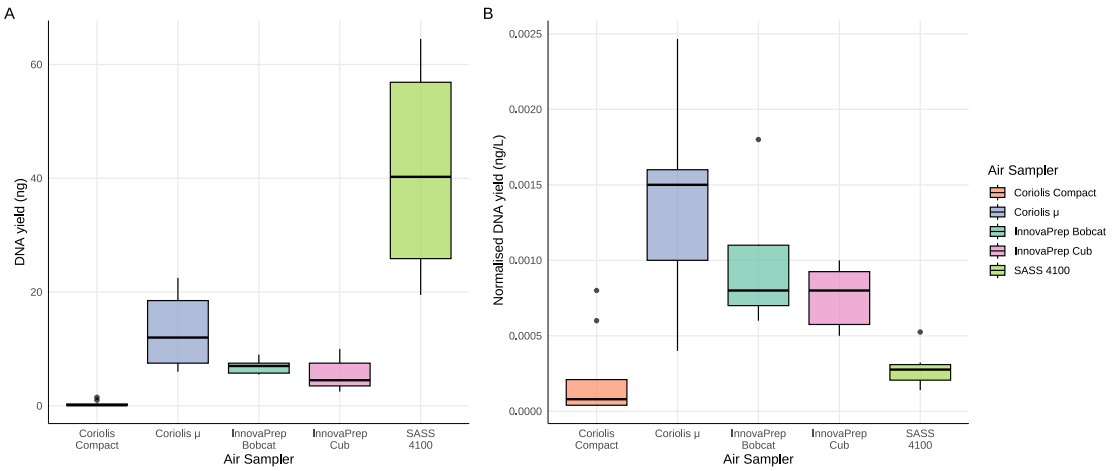


Figure 4.7: Boxplots comparing DNA yield across five air samplers. A) Total DNA yield (ng), B) DNA yield normalised by the volume of air sampled (ng/L). Each box represents the interquartile range (IQR), with the horizontal line indicating the median. Outliers are shown as individual points. Samples containing visible insect material were excluded prior to analysis.

The Coriolis Compact consistently underperformed, yielding negligible amounts (<1 ng) at both durations, regardless of site. This pattern suggests that the air sampler design and flow rate, rather than sampling location, were the dominant factors shaping yield.

With visibly contaminated samples excluded (Figure 4.7A), the SASS samples produced the highest DNA yields, followed by the Coriolis µ, while the Bobcat and Cub yielded lower amounts but exhibited less variability between replicates. The Coriolis Compact consistently yielded below 0.5 ng.

When yields were normalised by air volume (Figure 4.7B), differences between samplers narrowed. The Coriolis µ showed the highest recovery efficiency, followed by the Bobcat and Cub. The SASS, despite its high absolute yield, had the lowest normalised recovery, indicating that its superior performance was largely a consequence of its high flow rate rather than greater efficiency.

Total reads per sampler and duration The number of passing-filter reads varied substantially by sampler and by collection duration (Figure 4.8), despite all libraries being normalised to 5 ng prior to sequencing. Across all samplers, the 50-minute collections produced on average more reads than the 25-minute collections (23,826 vs 18,856), although the effect was not consistent across individual sampler types.

The 50-minute Coriolis µ samples produced the highest read numbers, with a mean of 62,012, compared to 22,332 from the shorter 25-minute samples. The SASS also yielded relatively high read counts (25-min: 26,465 and 50-min: 33,243). In the 25-minute samples the Bobcat had the highest mean read count (35,054), substantially higher than the Cub (5,847). But this was reversed in the 50-minute samples where the Bobcat had the second lowest number of reads (7,261). The Coriolis Compact generated the lowest read counts, averaging fewer than 5,000 reads per sample.

Analysis of taxonomic composition of samples MARTi taxonomic assignments were used to perform several analyses, including the generation of stacked bar charts at the phylum level, assessment of taxonomic richness across samples, identification of unique taxa recovered by each sampler, and determination of the five most abundant species per sampler. In addition, ordination plots (PCoA and t-SNE) were generated to visualise relationships in community composition between samples.

Phylum level analysis Across all samples, common phyla included Streptophyta, Pseudomonadota, Actinomycetota, Basidiomycota, and Ascomycota (Figure 4.9). Reads assigned to Arthropoda and Chordata varied by sampler and collection duration, and were generally more abundant in samples with visible insect contamination. The Coriolis µ sampler had the most visibly contaminated filters (Table 4.9), and many of its samples were dominated by Arthropoda reads. Some discrepancies were observed: for example, neither of the 25-minute Coriolis µ samples collected at Church Farm showed visible contamination, yet both were dominated by Arthropoda reads. Conversely, the 50-minute µ_2 sample collected at NHM was visibly contaminated but contained no Arthropoda reads. Furthermore, neither of the visibly contaminated Bobcat and Cub samples (Church Farm 25-minute Bobcat_2 and NHM 25-minute Cub_1) contained Arthropoda reads.

Church Farm samples showed more evenly distributed communities overall than NHM samples. Streptophyta was dominant in the majority of samples collected at NHM, but

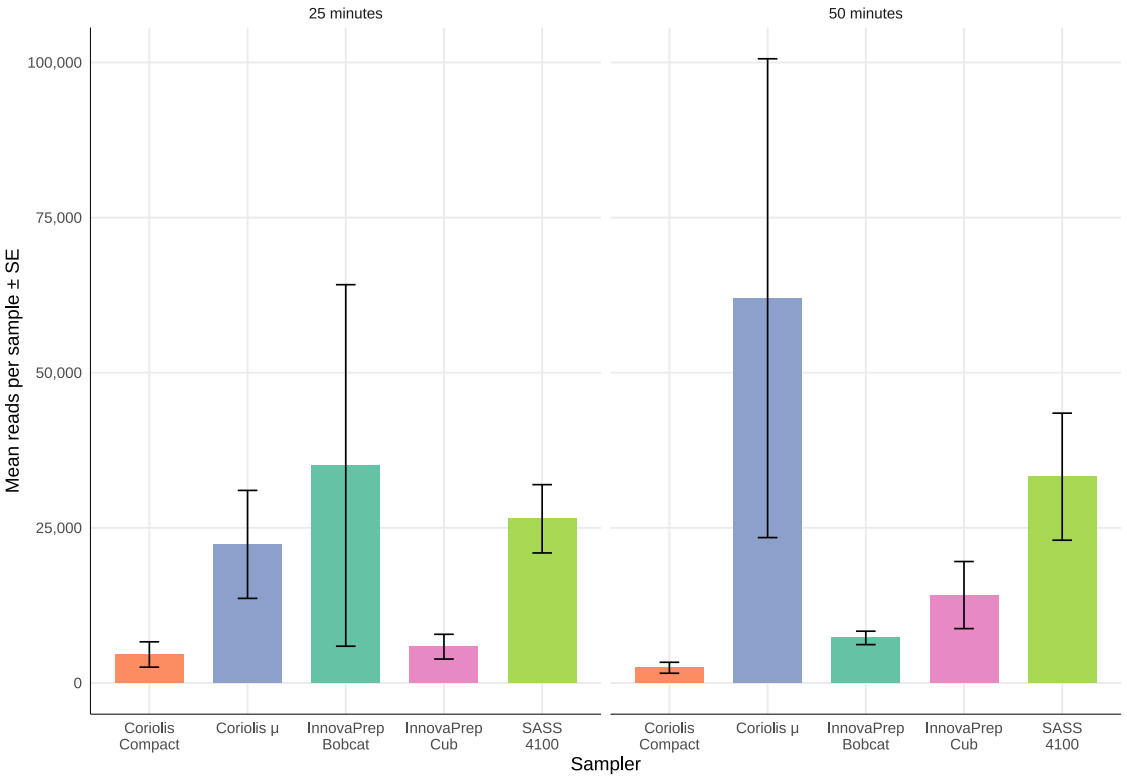


Figure 4.8: Mean number of reads sequenced (± SE) for each air sampler and collection duration.

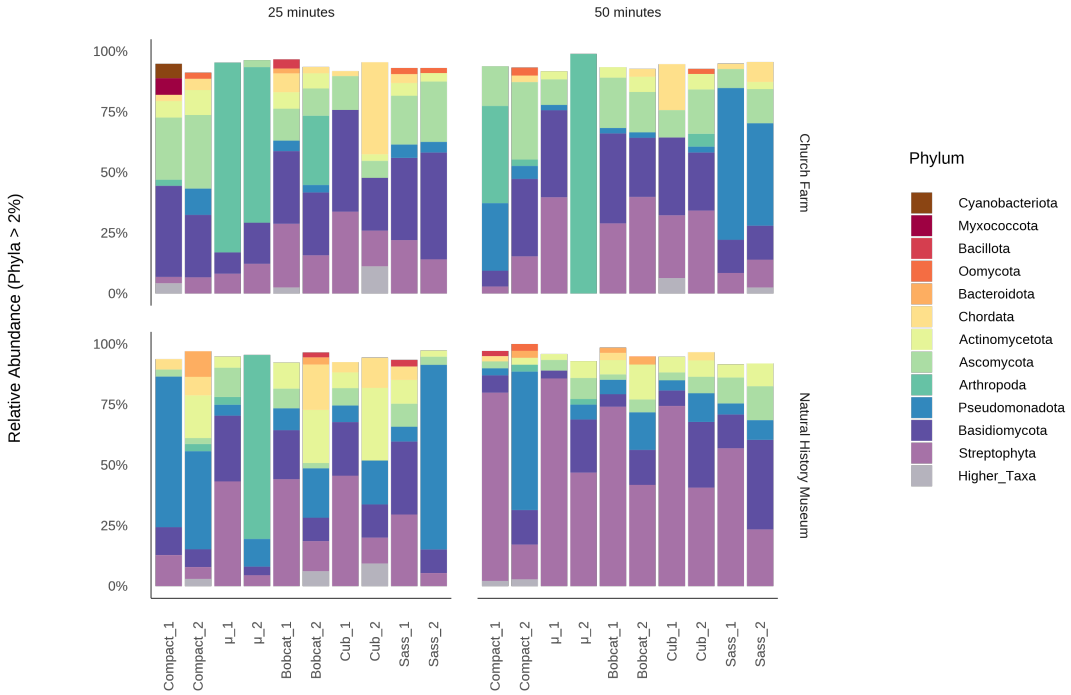


Figure 4.9: Stacked bar plot showing the relative abundance of phyla (each >2% within-sample abundance) across replicate air samples. Each bar represents a single sample replicate, grouped by air sampler type and faceted by sampling location and duration. Read counts were agglomerated at the phylum level and normalised within each replicate to reflect relative abundance. "Higher Taxa" reads represent sequences classified only to taxonomic ranks higher than Phyla level.

other taxa also featured prominently. Chordata was observed in several Church Farm samples, with highest values in one replicate of each Cub sample duration and moderate levels in SASS and Coriolis Compact.

At NHM the 25-minute samples showed greater variability, than the longer samples. Streptophyta is the predominate phyla in almost all of the initial repeat samples, aside from the Coriolis Compact. This demonstrates both the consistency between most of the air samplers, and the need to simultaneously collect samples when comparing air samplers. The majority of the second repeats instead contain elevated levels of Actinomycetota, alongside many other phyla. Both of the Coriolis Compact 25-minute samples are predominantly Pseudomonadota, which is consistent with the second SASS repeat but no other samples from this collection.

The longer, 50-minute samples collected at NHM were consistently dominated by plant DNA, with Streptophyta accounting for 62–84% of the relative abundance depending on the sampler. The Bobcat and Coriolis µ samples had the highest mean Streptophyta proportions. Only two of these samples were less than half Streptophyta (Compact_2 and SASS_2), which instead contain higher abundance of Pseudomonadota or Basidiomycota.

Additionally the Coriolis µ samples show consistently high Arthropoda levels, in 4 out of the 8 samples, likely due to the known insect contamination in the filters. Other notable differences in the air samplers include Cyanobacteria and Myxococcota only being detected by a single collection (Church Farm 25-minute Compact_1).

Across all of the samples Bacillota, Oomycota and Bacteroidota are only detected in a handful and often at very low abundance. Generally, they are only identified in one of the 2 replicates which may be due to differences in sampling time or there rarity. One example, is in the 25-minute Church Farm samples where Oomycota is present in Compact_2, and both of the SASS collections.

These results demonstrate that taxonomic composition differs not only by location and sampling time, but also by the air sampler used, even at higher taxonomic ranks. The observed variability, particularly in Streptophyta, Arthropoda and Chordata, underscores the need to explore the data at finer resolution. The next section examines species-level patterns, focusing on the most abundant species and the overlap between samplers.

Taxonomic richness and unique species To evaluate each sampler's ability to detect a diverse range of taxa, taxonomic richness (Figure 4.10) and the proportion of unique and shared species detected by each sampler (Figure 4.11) was assessed. For both analyses, the MARTi alignment data were normalised to 100,000 reads (HP100k) and filtered to include only taxa with a HP100k ≥ 1 and read count > 5, in order to exclude potentially erroneous alignments. In cases where samples contained fewer than 100,000 reads, normalisation was extrapolated.

Taxonomic richness, shared and unique species counts were calculated using different methods, resulting in differences in magnitude. Richness values represent the mean number of distinct taxa detected per sample (normalised to 100,000 reads), thereby reflecting the consistency of detection across replicates. In contrast, shared and unique species counts represent the total number of distinct species detected per sampler, with each species counted only once, regardless of the number of samples in which it was found. Consequently, richness values are higher as they account for repeated detections, whereas the shared and unique analysis combines data from each sampler.

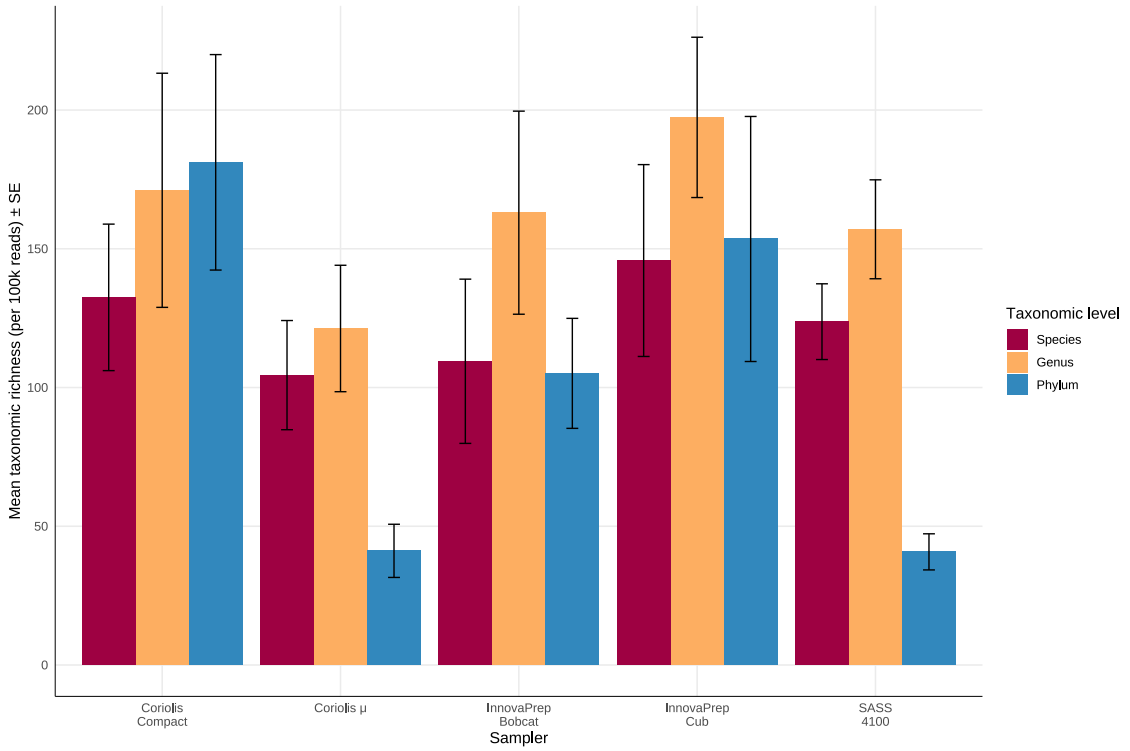


Figure 4.10: Mean number of taxa detected (± SE) for each air sampler. Richness is shown at species, genus, and phylum levels. Data were filtered to include only taxa with HP100k ≥ 1 and read count > 5. Values are normalised to 100,000 reads.

Across the samplers species and genus richness are closely correlated, with genus richness consistently higher (Figure 4.10). This suggests some reads could only be confidently assigned to genus level, lacking corresponding species-level identification. Phylum-level richness is generally the lowest across samplers, except for the Coriolis Compact, which shows elevated richness possibly due to the detection of two phyla not identified by any other sampler (Figure 4.9).

At the species and genus levels, the Cub sampler exhibited the highest average richness (Species: 145.8 ± 34.6; Genus: 197.4 ± 28.9), while the Coriolis µ recorded the lowest (Species: 104.5 ± 19.7; Genus: 121.3 ± 22.8). The Coriolis Compact had the second highest richness at these levels, with the Bobcat recording the second lowest richness at species level and the SASS the second lowest richness at genus level. At the phylum level, the Coriolis Compact showed the highest richness (181.2 ± 38.8), while the SASS recorded the lowest (40.8 ± 6.5).

These findings indicate relatively minor differences in the number of species detected across samplers, with greater variation attributable to differences in sampling duration and location. Consequently, it is important to assess whether the same species are being consistently detected by different samplers when deployed under comparable conditions.

When species are counted once per sampler, the Coriolis Compact, Cub, and SASS identified the highest total number of species per 100k, while the Bobcat detected the fewest (Figure 4.11A). The InnovaPrep samplers (Cub and Bobcat) reported the lowest number of unique species (Cub: 3.75; Bobcat: 3.54 per 100k), potentially due to their similar collection mechanisms causing overlapping detections. Notably, the Cub detected substantially more shared species than the Bobcat (Cub: 40.0; Bobcat: 10.1 per 100k). This discrepancy is likely driven by normalisation effects, since the Bobcat produced a

Table 4.10: Top five unique species per sampler, ranked by total HP100k. Where a sampler has fewer than five unique species, all identified unique species are shown. The table lists the species, category, total HP100k, and the number of samples in which each species was detected.

Sampler	**Species**	**Category**	**Total HP100k**	**No. of samples**
Coriolis Compact	*Arsenophonus nasoniae*	Bacteria (Isolated from insect)	753.38	1
	Hymenobacter sediminicola	Bacteria	265.05	1
	Arsenophonus endosymbiont of *Aphis craccivora*	Bacteria (Insect symbiont)	221.58	1
Coriolis µ	*Aphis gossypii*	Insect	5,426.58	3
	Rhopalosiphum padi	Insect	4,723.88	1
	Melangyna quadrimaculata	Insect	3,114.88	3
	Syrphus vitripennis	Insect	1,865.40	3
	Buchnera aphidicola	Bacteria (Insect symbiont)	1,638.90	1
Innovaprep Bobcat	*Lestremiinae sp.* MAB-2008	Insect	121.75	1
	Janibacter melonis	Bacteria	98.75	1
	Pseudoduganella plicata	Bacteria	65.83	1
	Bibio marci	Insect	17.16	1
	Dilophus febrilis	Insect	14.71	1
Innovaprep Cub	*Corynebacterium propinquum*	Bacteria	105.51	2
	Talaromyces rugulosus	Fungus	67.83	1
	Nocardioides sp. S5	Bacteria	26.16	1
SASS 4100	*Ralstonia pickettii*	Bacteria	22,739.31	5
	Cupriavidus basilensis	Bacteria	2,018.8	2
	Ralstonia wenshanensis	Bacteria	1,836.14	3
	Ralstonia insidiosa	Bacteria	817.67	3
	Streptomyces laculatispora	Bacteria	269.14	1

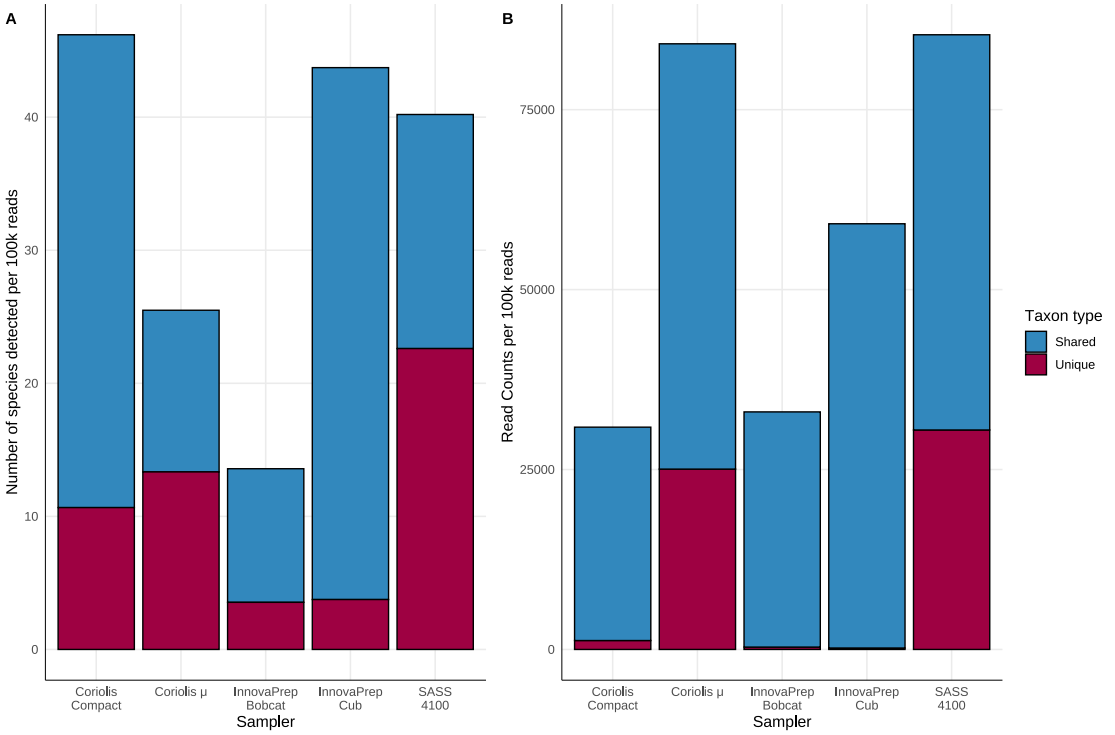


Figure 4.11: Unique and shared species by sampler. A) Number of Species per 100,000 reads for each sampler, split into shared and unique species. B) Sum of read counts per 100,000 reads for shared and unique species. Unique species are those detected by only one sampler across all samples. Data were filtered to retain species with HP100k ≥ 1 and raw read count ≥ 5.

higher number of passing-filter reads than the Cub (Figure 4.8), which reduces species per 100k when the additional reads are concentrated in a small set of taxa.

A comparison of normalised read counts for unique and shared taxa highlights that the contribution of unique species was generally low (Figure 4.11B). Exceptions were the Coriolis µ and SASS, where unique taxa accounted for a more noticeable fraction of total reads.

Examination of the taxa driving these patterns (Table 4.10) shows that most Coriolis µ unique detections were insects, whereas the SASS captured additional bacteria, particularly *Ralstonia spp.* Across all samplers, the unique taxa were rarely fungal, only the Cub detected a unique fungal species. As a result, the unique fraction does not appear to represent missing fungal pathogens, which are the primary focus of AirSeq.

Top 5 most abundant species per sampler The stacked bar chart in Figure 4.12 illustrates the five most abundant species detected per sampler, organised by location and sampling duration. After applying filtering thresholds (HP100k ≥ 1 and read count > 5), a total of 39 unique species were retained.

Some species were exclusively detected by specific samplers, such as *Bradyrhizobium sp. PSBBO68* in Coriolis Compact samples and *Ralstonia pickettii* in SASS samples. Others were consistently identified across multiple samplers and locations, including *Ustilago hordei* (a barley pathogen), *Urtica urens* (nettle), and *Lolium perenne* (ryegrass).

Marked differences are observed in the Coriolis Compact collections between locations and sample durations. Repeated collections also show variability in species composition.

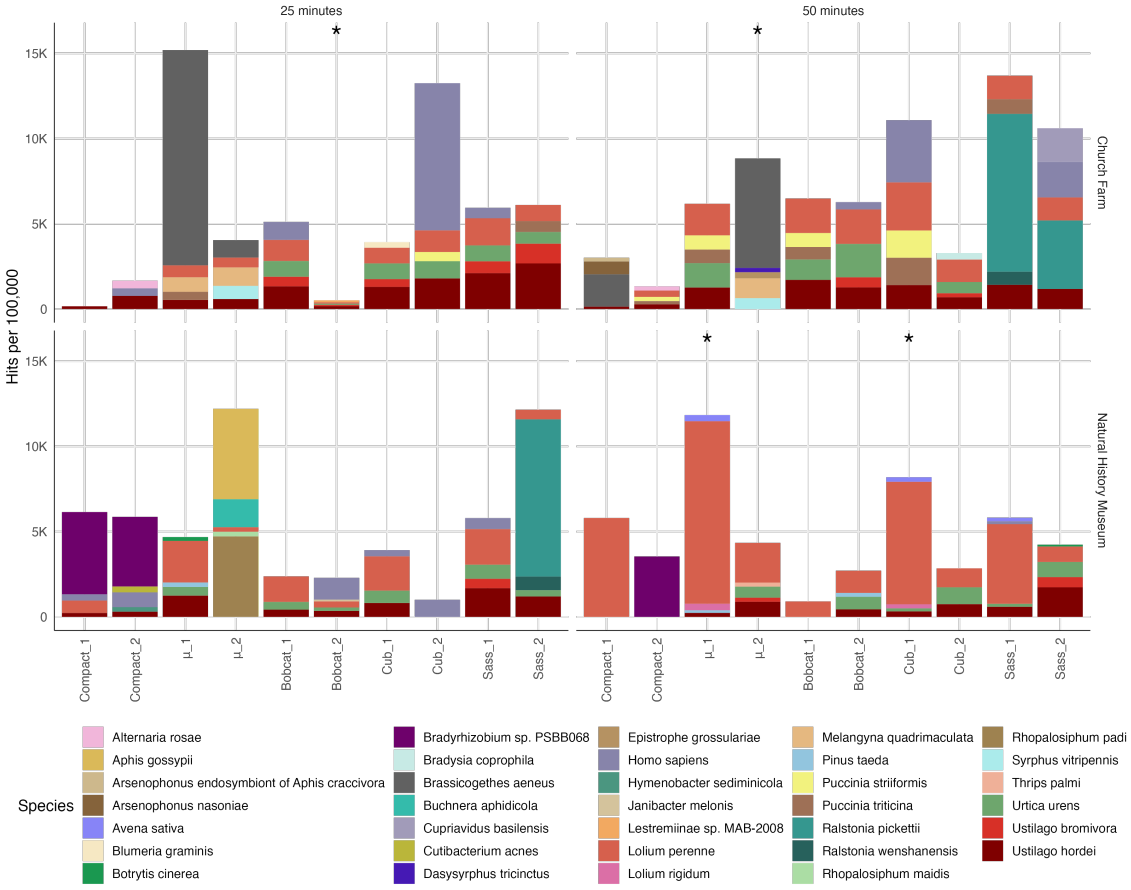


Figure 4.12: Hits per 100,000 of the top five most abundant species per sampler. Rows represent sampler location and columns represent sample duration. Data were filtered to include taxa with HP100k ≥ 1 and read count > 5. An asterisk (*) above a bar indicates that the sample contained a visibly trapped insect.

Three of the four NHM samples collected with the Coriolis Compact are predominantly composed of *Bradyrhizobium sp. PSBBO68*, a soil-associated bacterium not detected by other samplers. The remaining NHM sample (50-minute Compact_1) consists entirely of *L. perenne*, detected in many other samples. In contrast, Coriolis Compact samples from Church Farm contained fewer reads, especially the 25-minute samples, and exhibited greater species diversity. The first replicate of the 50-minute Church Farm sample was dominated by *Brassicogethes aeneus*, a beetle also found in high proportions in the 25-minute Micro_1 and 50-minute Micro_2 Church Farm samples. Although insects were not visually noted on these filters (except for 50-minute Micro_2), the presence of insect reads suggests that trapped insects may have gone unnoticed. Additionally, nearly all Coriolis Compact samples contained reads assigned to *Homo sapiens*.

As noted above, many of the Coriolis µ samples collected at Church Farm are dominated by insect DNA, particularly *B. aeneus*, with smaller proportions of reads from hoverfly species such as *Melangyna quadrimaculata* and *Syrphus vitripennis*. An exception is the 50-minute µ_1 sample, which aligns more closely with the species composition seen in other Church Farm samples, particularly those from the Bobcat and Cub samplers. This sample includes common taxa such as *U. hordei*, *U. urens*, *L. perenne*, and two fungal pathogens, *Puccinia triticina* and *Puccinia striiformis*. The NHM 25-minute µ_2 sample contained visible insects and contained high proportions of reads from aphid species, including *Rhopalosiphum padi* and *Aphis gossypii*, as well as their obligate endosymbiont *Buchnera aphidicola*. The other Coriolis µ samples from NHM contained species commonly detected across other samples.

Samples collected using the Bobcat and Cub samplers shared many of the same species found in other samples, including *U. hordei*, *U. urens*, and *L. perenne*. These samples also included reads for *H. sapiens*, *Avena sativa* (oat), and fungal pathogens such as *Ustilago bromivora*, *P. triticina*, and *P. striiformis*, all of which were frequently observed across the dataset. Notably, the Bobcat and Cub samplers exhibited the highest proportions of human DNA among all samplers, especially the Cub samples one of which is composed entirely of *H. sapien* reads in this figure (Church Farm 25-minute Cub_1).

The SASS samples also contained species detected by other samplers, along with some unique detections. For example, three SASS samples (two 50-minute Church Farm samples and 25-minute SASS_2 from Church Farm) showed high proportions of *R. pickettii*, a soil-associated bacterium.

Overall, the Coriolis Compact sampler detected a distinct set of abundant species compared to the other samplers, with the exception of *U. hordei*, which was identified across multiple devices. The Bobcat, Cub and SASS samplers detected a broadly similar and more diverse set of abundant species. Whilst the Coriolis µ samples were often dominated by insect DNA and their associated symbionts. However, when such contamination was absent, the species profiles of Coriolis µ samples aligned more closely with those of the other samplers.

Ordination plots The HP100k-normalised species-level MARTi assignment data were visualised using ordination plots (Figure 4.13). Two approaches were employed: t-SNE based on presence-absence data (Figure 4.13A) and PCoA using Bray–Curtis distance (Figure 4.13B).

Bray–Curtis distance was selected over Jaccard distance for the PCoA as it incorporates

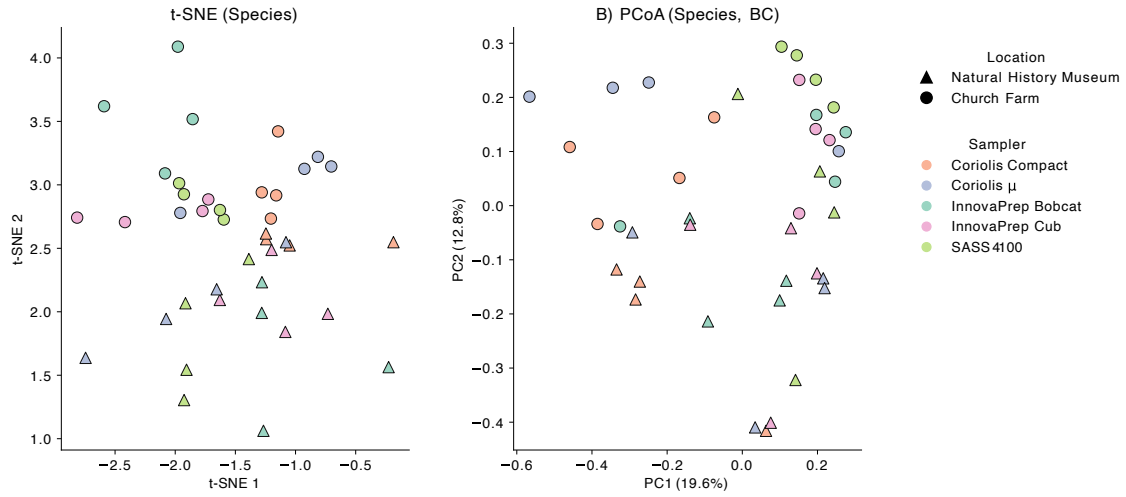


Figure 4.13: Ordination plots generated from a presence–absence matrix of species-level assignments derived from MARTi data. A) t-SNE projection; B) PCoA using Bray–Curtis distances (axes 1 and 2). Points are coloured by according to the sampler used and and shaped by location (triangle: NHM, circle: Church Farm).

taxon abundance in addition to presence-absence, offering greater sensitivity to variations in dominant species [33].

Both ordinations reveal some clustering based on collection site. In the t-SNE plot (Figure 4.13A), Church Farm samples show sampler-specific clustering, particularly among the Coriolis Compact and Coriolis µ samples. The Bobcat samples lie at the periphery of the Church Farm cluster. In contrast, NHM samples are more widely dispersed, indicating greater heterogeneity. Notably, Coriolis Compact samples from both sites tend to cluster centrally, suggesting some consistency across locations for this sampler.

In the PCoA plot (Figure 4.13B), a distinct cluster forms among Church Farm samples, excluding those collected with the Coriolis µ and Coriolis Compact samplers. This pattern indicates an influence of both site and sampler on community structure. NHM samples, by comparison, are more scattered, suggesting a more variable or less structured airborne microbial community at that location.

Overall, both ordination methods reveal consistent trends, showing that both sampling location and sampler type influence community composition. However, location appears to be the strongest predictor of species composition.

In conclusion the results from the sampler comparison experiment show the composition of airborne samples varied according to sampler type, location, and sampling duration. The InnovaPrep samplers (Cub and Bobcat) produced consistent community profiles across conditions, while the SASS yielded relatively high DNA concentrations. In contrast, the Coriolis Compact consistently produced the lowest DNA yields, and the Coriolis µ frequently recovered insect DNA, which may have influenced species composition results.

4.5.1.2 Distance from Maize Source

The following section presents results from the distance-from-source maize experiments, including subsampling validation, control performance, temporal patterns in airborne *Zea mays* pollen, and spatial distribution showing the decline in pollen abundance with distance from the source plot. Most results are expressed as *Zea* counts per 100,000 reads (*Zea* HP100k).

Table 4.11: Comparison of *Zea* read counts in 100k subsampled and full datasets.

Sample	**Analysed Reads**	**Total Sample (*Zea* Count)**[1]	**Total Sample (*Zea* HP100k)**[2]	**100k Subsample (*Zea* Count)** [3]	**Absolute Difference**
290824_10_3	414,687	72	17.36	23	-5.64
280824_10_1	169,126	57,788	34,168	33,687	481.61
270824_100_2	280,381	139	49.58	50	-0.42

[1] *Zea* reads in the full dataset before any filtering.
[2] *Zea* hits per 100k reads after MARTi quality and length filtering.
[3] *Zea* reads identified in a 100k subsample of total reads.

Subsampling validation The results of the subsampling validation (Table 4.11), indicate that subsampling has a negligible impact on the normalised *Zea* HP100k values. Across all three samples, the absolute differences between the subsample and full dataset ranged from –5.64 to +481.61 hits per 100,000 reads. Although the latter value appears numerically large, it represents only a 1.4% relative difference, highlighting that subsampling yields results consistent with those obtained from the full dataset.

Control sample performance Negative (water) and positive (lambda) controls were included during sequencing to assess the impact of barcode cross-talk. Each control sample was individually barcoded and then either pooled with all experimental samples for sequencing, or pooled separately with other controls only. This allowed comparison of potential barcode cross-talk under both sequencing conditions.

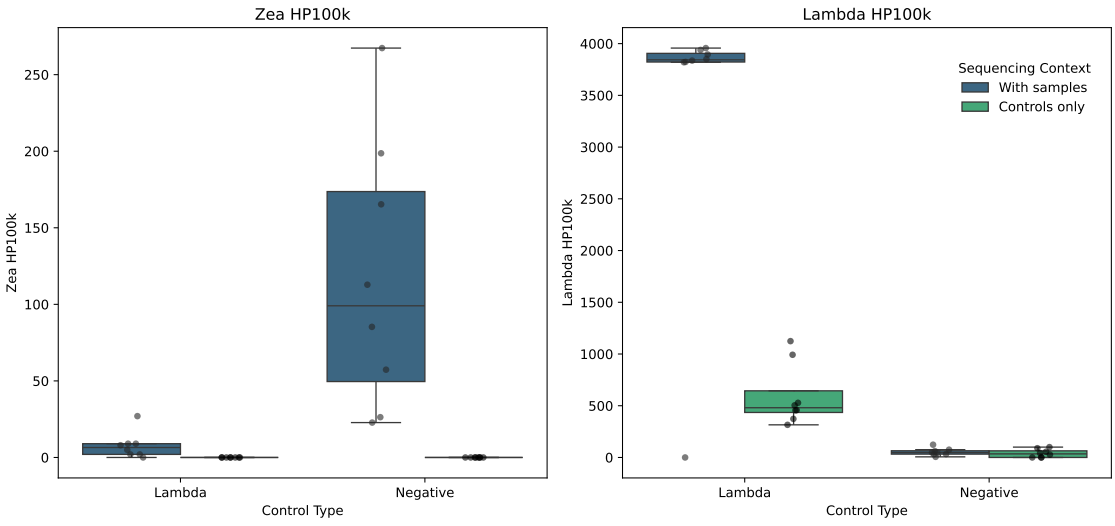


Figure 4.14: Boxplots showing the HP100k of *Zea* (left) and Lambda (right) aligned reads for the negative (water) and lambda control samples. Library preparation and sequencing were either performed alongside experimental samples ("With samples") or only with other controls ("Controls only").

Figure 4.14 shows the proportion of *Zea* and lambda reads in different control types, depending on whether they were sequenced alongside experimental samples or independently. *Zea* reads were only detected in controls that were prepared with experimental samples, with notably higher abundance in the Negative controls (mean: 116.98 HP100k) compared

to the lambda controls (mean: 7.75 HP100k). The highest lambda HP100k value was observed in the lambda control prepared with the experimental samples (3390.12 HP100k), followed by the lambda control prepared independently (593.83 HP100k). Lambda reads were also present in the Negative controls both when sequenced with experimental samples (53.08 HP100k) and when prepared alone (39.31 HP100k).

Temporal patterns Throughout the maize trial, a sampler was consistently positioned at the centre of the plot. The proportion of *Zea* reads detected in this central sampler provides insight into how pollen release varied both over the course of each day and across the sampling fortnight.

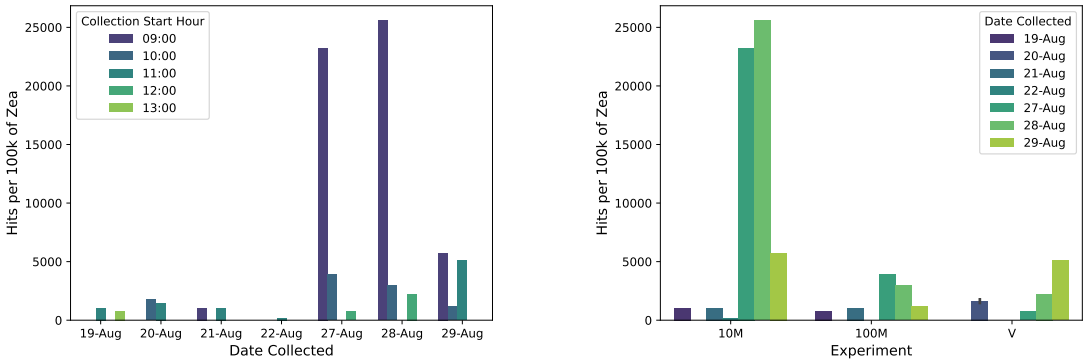


Figure 4.15: Bar plots showing the proportion of *Zea* aligned reads (Zea HP100k) detected in samples collected from the central sampler within the maize plot. The left plot shows the abundance by date and collection start, and the right plot shows abundance by date and experimental layout.

Detectable airborne *Zea* abundance peaks during the 09:00 and 10:00 collections on 27 and 28 August, with fluctuations observed across the other sampling dates (Figure 4.15, left). When the central sampler data are grouped by the experiment conducted at the time of collection (Figure 4.15, right), it becomes clear that airborne maize pollen abundance varies considerably between experiments. The 10 m experiments conducted on 27 and 28 August show an approximately 25-fold increase in *Zea* HP100k compared to those from 19 and 21 August.

Maize abundance normalisation A comparison of *Zea* abundance (HP100k) between the central sampler and the highest-value surrounding sampler revealed substantial differences in detection across distances (Table 4.12). For the 10 m configuration, excluding a single outlier on 28 August, surrounding samplers captured an average of 35% of the central sampler's reads (range: 10–74%). In contrast, for the 100 m configuration, surrounding samplers consistently recorded between 1% and 3% of the central sampler's value, with one exception on 21 August (13%). In the V-shaped layout, where the distance between the central and surrounding samplers varied, surrounding samplers collected between 1% and 23% of the central sampler's reads.

Wind speeds during sampling ranged from 2.4 to 9.2 m/s (Table 4.12). Within this range, no consistent relationship was observed between wind speed and the proportion of reads detected at surrounding samplers. In the 10 m experiments, for example, the highest average wind speed (9.2 m/s 22 August) coincided with one of the largest percentage dif-

ferences (66%), but this collection also yielded unusually low counts at the central sampler. Additionally, the lowest wind speed (2.4 m/s 28 August) was associated with the smallest difference (0.9%), yet other dates with similar wind speeds produced larger differences. At greater distances (100 m and V configurations), the highest percentage differences were not linked to the highest wind speeds. Overall, distance between samplers, together with the absolute abundance of *Zea* at the central sampler, appear to be the dominant factors shaping relative detection.

Table 4.12: Percentage difference in *Zea* abundance (HP100k) between the central sampler and the highest-value surrounding sampler across sampler layouts.

Layout	**Date**	**Wind speed (average m/s)**	**Central sampler (Zea HP100k)**	**Highest sampler (Zea HP100k)**	**Difference (%)**
10 m	19/08/24	5.13	1033	122	11.81
10 m	21/08/24	4.51	1,042	772	74.09
10 m	22/08/24	9.24	197	130	65.99
10 m	27/08/24	5.62	23,247	3,065	13.18
10 m	28/08/24	2.44	25,566	225	0.88
10 m	29/08/24	2.98	5753	618	10.74
100 m	19/08/24	5.66	786	26	3.31
100 m	21/08/24	5.09	1057	137	12.96
100 m	27/08/24	6.14	3923	51	1.30
100 m	28/08/24	3.79	2976	46	1.55
100 m	29/08/24	3.05	1190	14	1.18
V	20/08/24 (1)	4.66	1781	410	23.02
V	20/08/24 (2)	6.04	1469	133	9.05
V	27/08/24	5.72	761	102	13.40
V	28/08/24	4.86	2226	172	7.73
V	29/08/24	4.05	5113	90	1.76

Spatial patterns of maize distribution For each experiment, a heatmap showing *Zea* HP100k values across sampling points is presented alongside a wind vector plot representing conditions during the corresponding sampling period (Figures 4.17,4.18,4.19). A schematic of the sampler layout and numbering system is provided in Figure 4.16 for reference.

10 m grid experiment The results of the 10 m grid experiments are shown in Figure 4.17, *Zea* HP100k values in non-central samplers ranged from 13 (sampler 6, 29 August) to 3,065 (sampler 2, 27 August). Addiotionally, the relationship between wind speed, central sampler abundance, and maize detection at 10 m is not consistent.

On 19 August, surrounding samplers showed similar *Zea* levels (47–122 HP100k), with the highest in sampler 3, broadly aligning with the NE wind direction. In contrast, the highest wind speed across all 10 m experiments was recorded on 22 August (9.24 m/s), again directed towards the NE. Despite this, surrounding sampler values remained low and closely grouped (17–130 HP100k). This may reflect the relatively low abundance in the central sampler that day (197 HP100k), with similar values in downwind samplers 2 and 3 (129 and 130 HP100k), suggesting that limited airborne pollen was largely directed downwind, while upwind samples recorded minimal presence.

A stronger contrast was observed on 21 August, where sampler 5 recorded 772 HP100k, consistent with an E wind, whilst the surrounding samplers contained 29 - 112 HP100k. Despite a lower average wind speed (4.41 m/s compared to 9.24 m/s on 22 August) and a broader wind direction spread. Additionally, there was a higher proportion of *Zea* detected in the C sampler on the 21 August (1,042 HP100K). Suggesting the abundance of airborne maize pollen in the C sampler has a larger impact on detection by surrounding samplers than wind speed.

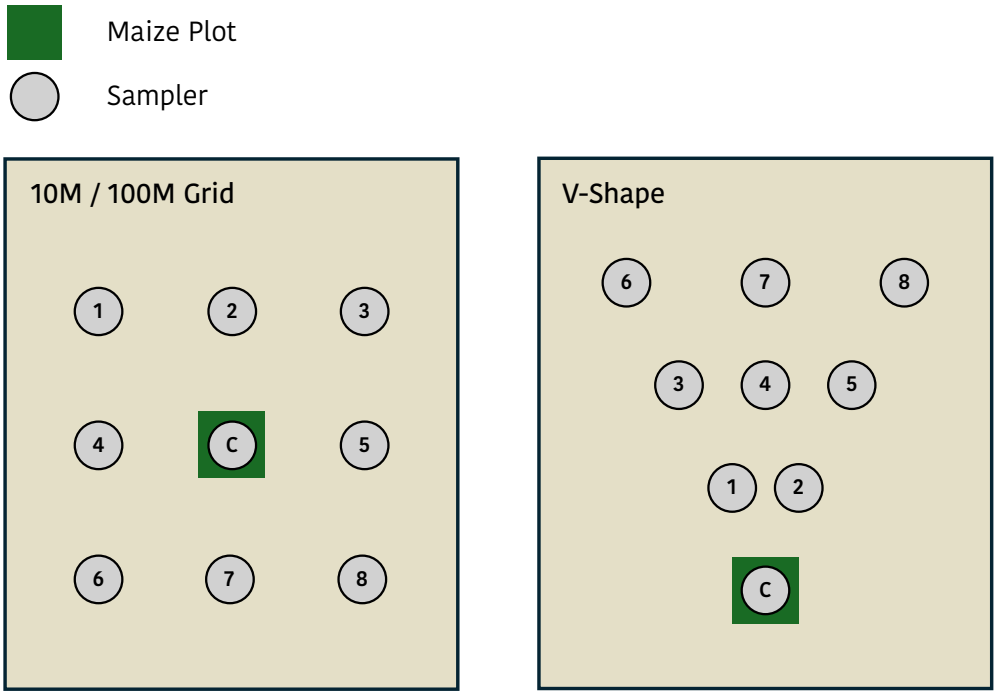
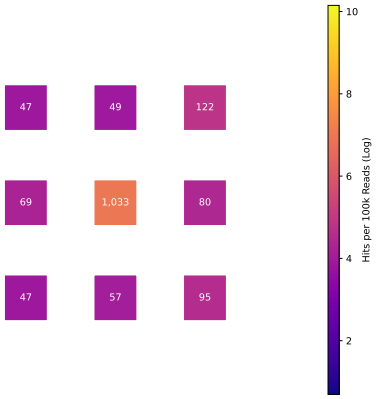


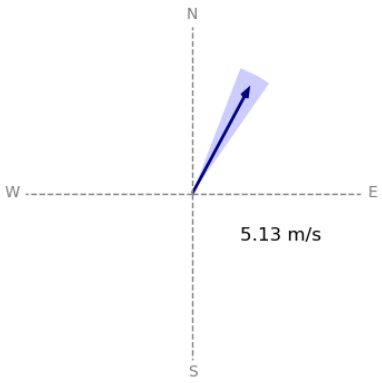
Figure 4.16: Diagram showing the sampler numbering scheme used to identify sampling locations across the experimental plots.

On 27 and 28 August, C sampler *Zea* HP100k values peaked above 23,000, with differing wind conditions influencing the spatial distribution. On 27 August (5.62 m/s average wind speed), sampler 2, directly downwind of the maize, recorded the highest non-central value (3,065 HP100k), and other upwind samplers (1 and 3) also exceeded 300 HP100k. In contrast, on 28 August the average wind speed dropped to 2.44 m/s, and the upwind sampler (1) contained just 162 HP100k. Interestingly, sampler 6, not directly downwind, recorded a higher value (225 HP100k), while sampler 4, located between 1 and 6, had only 77 HP100k. This suggests that under lighter winds, pollen dispersal becomes less directionally constrained.

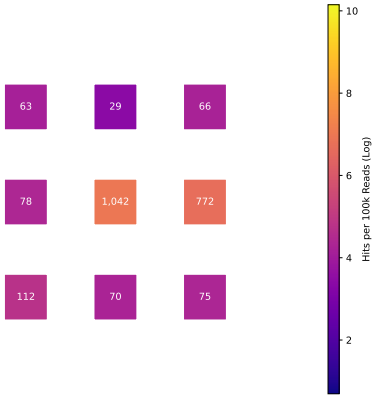
The final 10 m experiment, conducted on 29 August, had the third-highest central *Zea* abundance (5,753 HP100k) with a moderate wind speed (2.98 m/s). Two upwind samplers (3 and 5) recorded 152 and 618 HP100k respectively, with sampler 5 lying in the more direct wind path through the plot. All other surrounding samplers contained fewer than 50 HP100k, further reinforcing the influence of directional wind flow in distributing pollen.



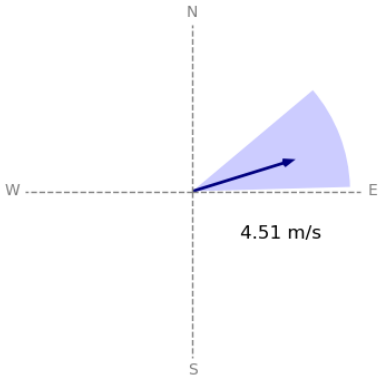
(a) 10 m Heatmap - 19/08



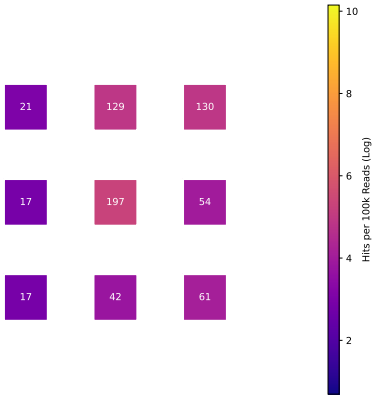
(b) 10 m Wind - 19/08



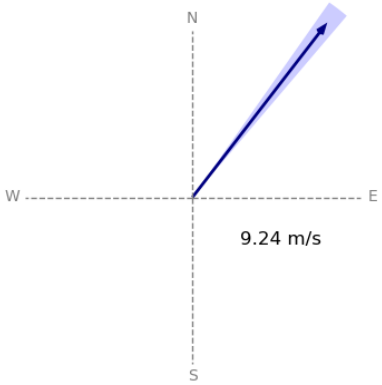
(c) 10 m Heatmap - 21/08



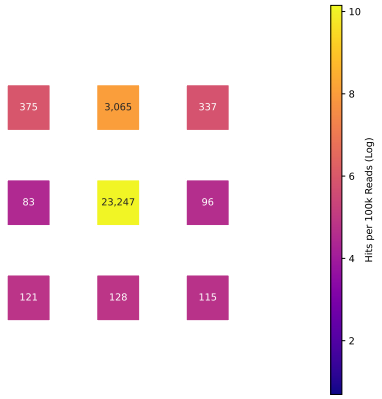
(d) 10 m Wind - 21/08



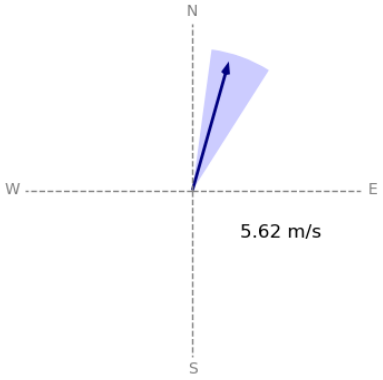
(e) 10 m Heatmap - 22/08



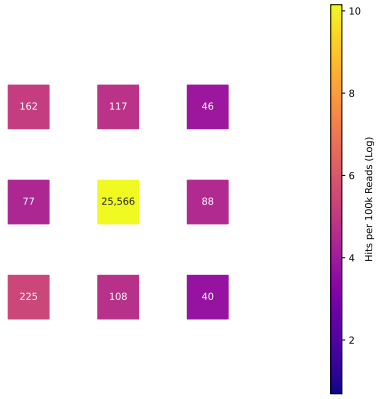
(f) 10 m Wind - 22/08



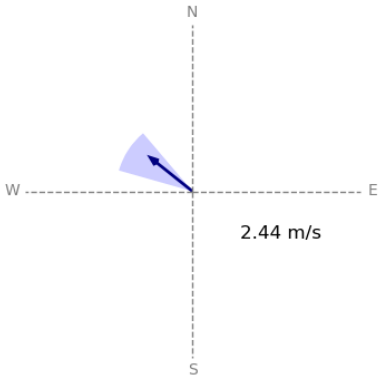
(g) 10 m Heatmap - 27/08



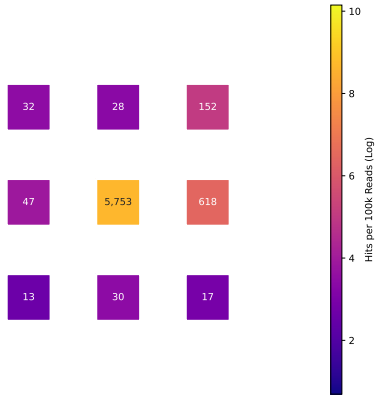
(h) 10 m Wind - 27/08



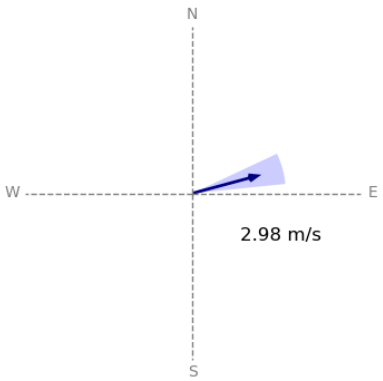
(i) 10 m Heatmap - 28/08



(j) 10 m Wind - 28/08



(k) 10 m Heatmap - 29/08



(l) 10 m Wind - 29/08

Figure 4.17: Spatial distribution of *Zea* HP100k and wind conditions for 10 m experiments. Each row shows a heatmap of *Zea* read abundance (HP100k) across sampling points, with values labelled in each grid cell. A logarithmic colour scale represents variation in abundance. Accompanying each heatmap is a wind vector plot for the corresponding sampling period. Arrow direction indicates wind flow, length reflects mean wind speed (annotated), and the light blue arc shows the range of wind directions, with radius indicating maximum speed.

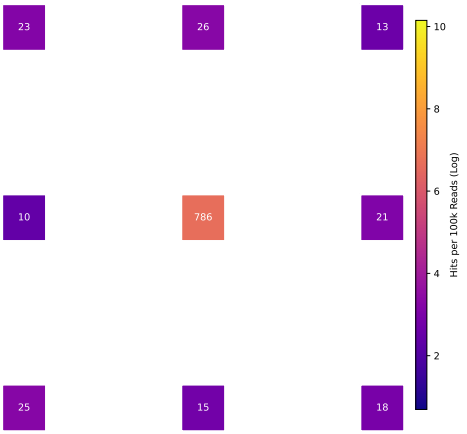
100 m grid experiment Over the same period as the 10 m experiments, a 100 m grid experiment was conducted (Figure 4.18). *Zea* HP100k values in non-central samplers were generally lower than in the 10 m experiments, ranging from 1 (sampler 3, 27 August) to 137 (sampler 8, 21 August). This reduction may be attributed to the increased sampling distance from the maize plot or to lower airborne *Zea* abundance, with central sampler (C) values ranging from 786 (19 August) to 3,931 HP100k (27 August).

The experiments on 19 and 21 August had similar conditions, with central sampler *Zea* HP100k values of 786 and 1,057, and NE winds averaging 5.66 and 5.09 m/s, respectively. On 19 August, all non-central samplers recorded less than 26 HP100k, suggesting minimal dispersal regardless of wind direction. In contrast, on 21 August, samplers 6 and 8 detected the highest *Zea* abundances across the 100 m experiments (>100 HP100k), despite being positioned upwind of the maize plot. All other samplers on that day recorded less than 21 HP100k.

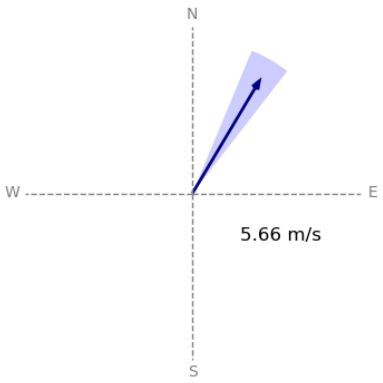
The highest wind speed and central *Zea* abundance were recorded on 27 August (6.14 m/s and 3,931 HP100k). On this day, all surrounding samplers showed elevated *Zea* levels compared to lower wind speed days, ranging from 8 to 51 HP100k. However, the directly downwind sampler (3) had one of the lowest readings (15 HP100k), suggesting that *Zea* distribution did not clearly correlate with wind direction on this occasion.

The lowest wind speeds were recorded on 28 and 29 August (3.79 and 3.05 m/s), with corresponding central *Zea* abundances of 2,976 and 1,190 HP100k. These days yielded some of the lowest surrounding sampler values, particularly on 29 August, where all but one sampler detected fewer than 10 HP100k.

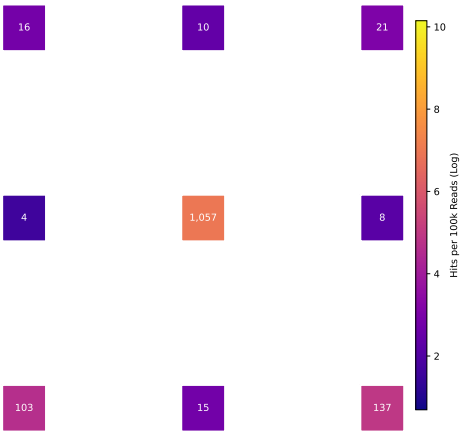
Overall, these results suggest that wind speed may have a greater influence than wind direction on *Zea* detection at 100 m from the source. On high wind speed days, increased detection may not reflect local release alone but could also indicate long-range transport of pollen from external sources. As shown in Figure 4.5, there was another maize plot located 1.74 km away from the central sampler.



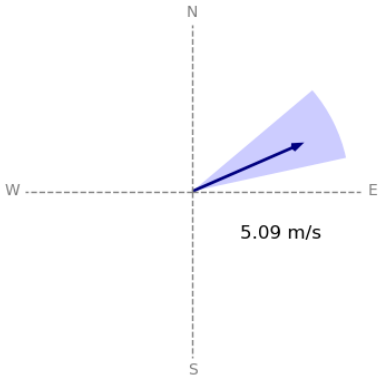
(a) 100 m Heatmap - 19/08



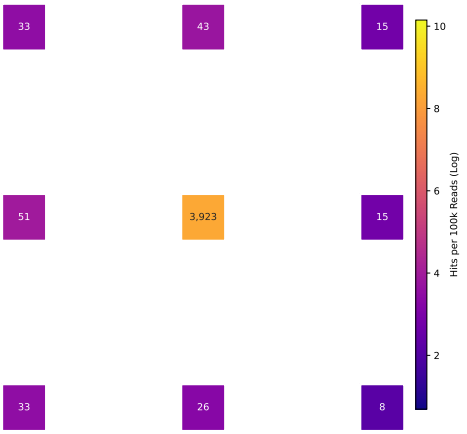
(b) 100 m Wind - 19/08



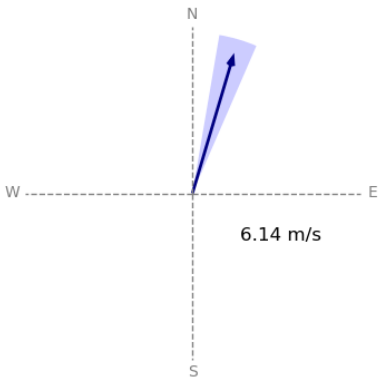
(c) 100 m Heatmap - 21/08



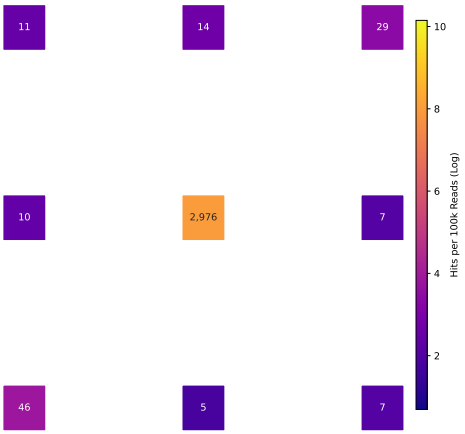
(d) 100 m Wind - 21/08



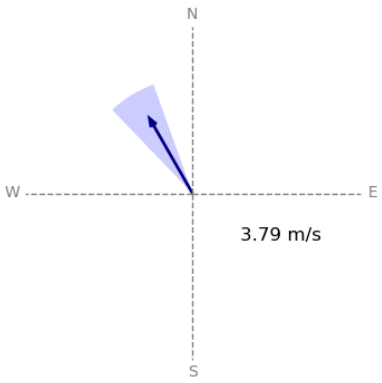
(e) 100 m Heatmap - 27/08



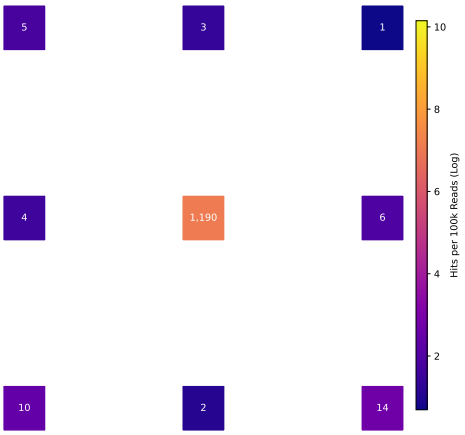
(f) 100 m Wind - 27/08



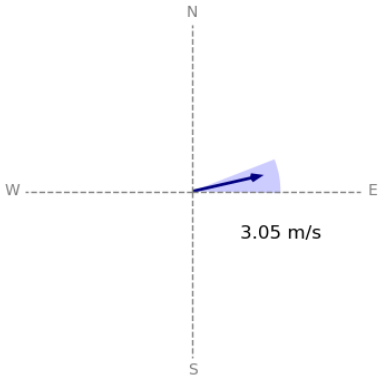
(g) 100 m Heatmap - 28/08



(h) 100 m Wind - 28/08



(i) 100 m Heatmap - 29/08



(j) 100 m Wind - 29/08

Figure 4.18: Spatial distribution of *Zea* HP100k and wind conditions for 100 m experiments. Figure legend as described in Figure 4.17

V-Shape experiment The data from the V-shape experiments are shown in Figure 4.19. In these heatmaps, sampler positions are plotted according to GPS data collected during each experiment and are therefore to scale (i.e. larger squares indicate samplers placed closer together during collection). The arrangement of samplers was based on wind direction at the start of sampling, so the distances between them vary depending on wind conditions and field constraints. These distances are provided in Table 4.13. The central sampler always remained at the centre of the maize plot. The numbering scheme is illustrated in the schematic in Figure 4.16.

Table 4.13: Sampling distances from the maize plot edge on each experimental date.

Date	**Distance from maize plot edge (m)**
20/08	25, 55, 85
27/08	17, 34, 50
28/08	12, 24, 36
29/08	30, 60, 90

Zea HP100k values in non-central samplers ranged from 9 (sampler 3, 28 August) to 410 (sampler 2, 10:46 on 20 August), while C sampler values ranged from 761 (27 August) to 5,113 HP100k (29 August).

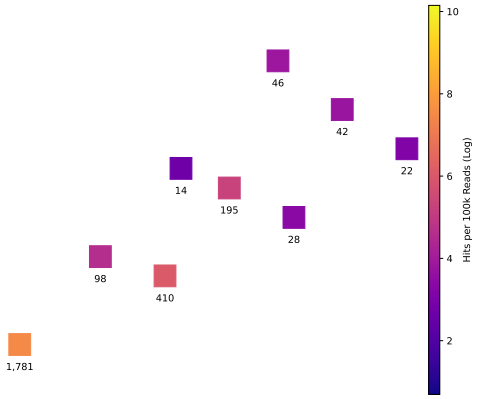
On 20 August, two consecutive collections were made at 10:46 and 11:59 using the same sampler positions. During this time, *Zea* abundance in the central sampler decreased slightly from 1,781 to 1,469 HP100k, while wind speed increased from 4.66 to 6.04 m/s, maintaining a NE direction. In both collections, sampler 2 recorded the highest non-central *Zea* abundance (410 and 105 HP100k), while sampler 1, placed at the same distance (25 m), recorded substantially less (36 and 15 HP100k). The lower values in sampler 1 during the second collection may be due to a narrower wind direction range, positioning it outside the dispersal path.

Samplers placed 55 m from the plot also recorded high *Zea* levels: sampler 4 in the first collection (195 HP100k) and sampler 3 in the second (133 HP100k), exceeding that of the nearer sampler 2 in the latter case. The furthest samplers (85 m) recorded values comparable to some nearer but off-path samplers, reinforcing that wind direction, rather than distance alone, was the dominant factor influencing detection.

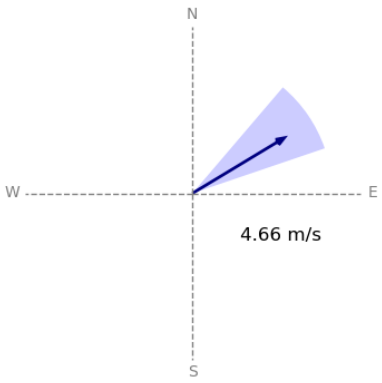
The experiments on 27 and 28 August had the shortest sampling distances (50 m and 36 m), with similar wind speeds (5.72 and 4.86 m/s) but differing wind directions (N–NE and N–NW). Central *Zea* abundances were 761 and 2,226 HP100k, respectively. On both days, sampler 2 recorded the highest non-central values (>100 HP100k), while all others remained below 40 HP100k. These distributions align with wind direction and sampler placement rather than distance from the maize plot.

The final V-shape collection on 29 August spanned the greatest distance (90 m), with an average wind speed of 4.05 m/s and a NE–E wind direction. The C sampler recorded the highest *Zea* abundance of all V-shape experiments (5,113 HP100k). The closest samplers recorded the highest non-central values (90 and 68 HP100k), while samplers 5 and 6 (42 and 50 HP100k) also showed elevated levels. Notably, sampler 6, at 90 m, detected similar *Zea* abundance to a sampler at 30 m, highlighting again the influence of wind direction.

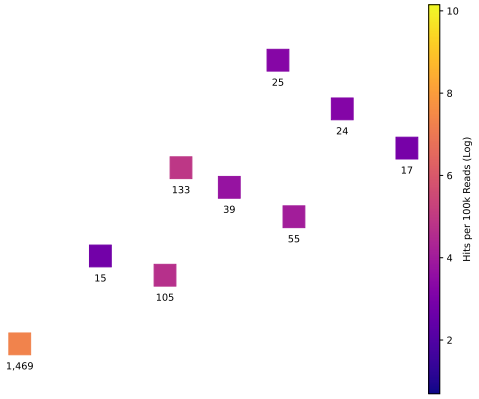
Overall, the V-shape experiments demonstrate that wind direction had the strongest influence on *Zea* detection in the surrounding area, exceeding the effects of both distance from the plot and central sampler abundance.



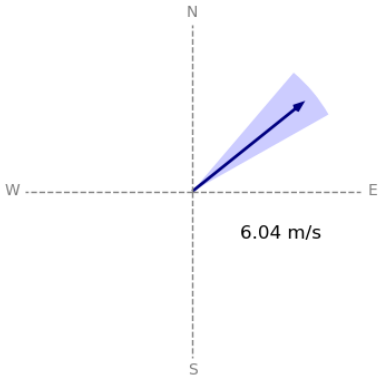
(a) V-shape Heatmap - 10:46 20/08



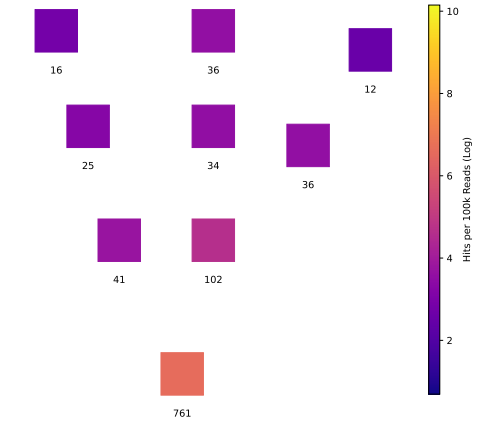
(b) V-shape Wind - 10:46 20/08



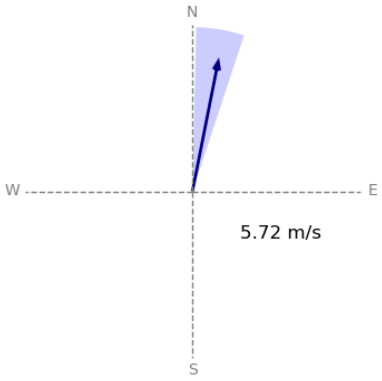
(c) V-shape Heatmap - 11:59 20/08



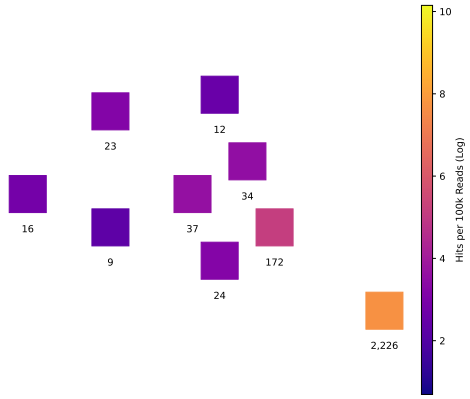
(d) V-shape Wind - 11:59 20/08



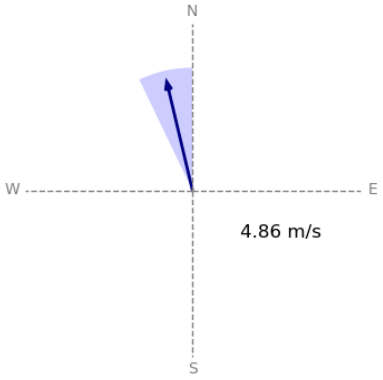
(e) V-shape Heatmap - 27/08



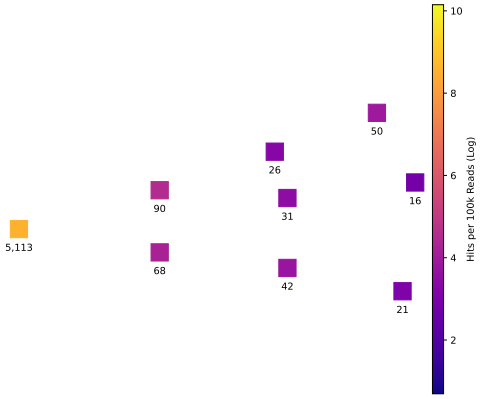
(f) V-shape Wind - 27/08



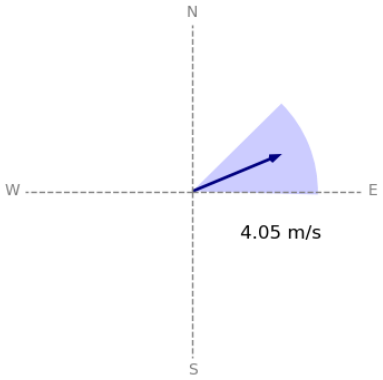
(g) V-shape Heatmap - 28/08



(h) V-shape Wind - 28/08



(i) V-shape Heatmap - 29/08



(j) V-shape Wind - 29/08

Figure 4.19: Spatial distribution of *Zea* HP100k and wind conditions for V-shape experiments. Figure legend as described in Figure 4.17

These results highlight the complex interplay between wind direction, airborne maize abundance, and sampling position. Some experiments show clear directional patterns, while others appear more strongly influenced by low pollen release or variable wind conditions. The 10 m experiments indicate that maize can be detected at short distances regardless of wind direction or speed. At 100 m, wind speed appears to be the dominant factor influencing detection, although the lack of a consistent directional effect suggests that some detected pollen may originate from external sources rather than the central plot. Finally, the V-shape experiments, which spanned distances from 12 to 90 m, provide strong evidence that wind direction is the primary factor affecting maize detection. While surrounding samplers consistently detected maize, abundances were highest in those positioned directly downwind of the plot.

Distance dependent detection of maize There is a clear decline in *Zea* abundance with increasing distance from the maize plot (Figure 4.20), within the V-shape experiments. The exponential curve indicates a steep drop between the central sampler and the closest

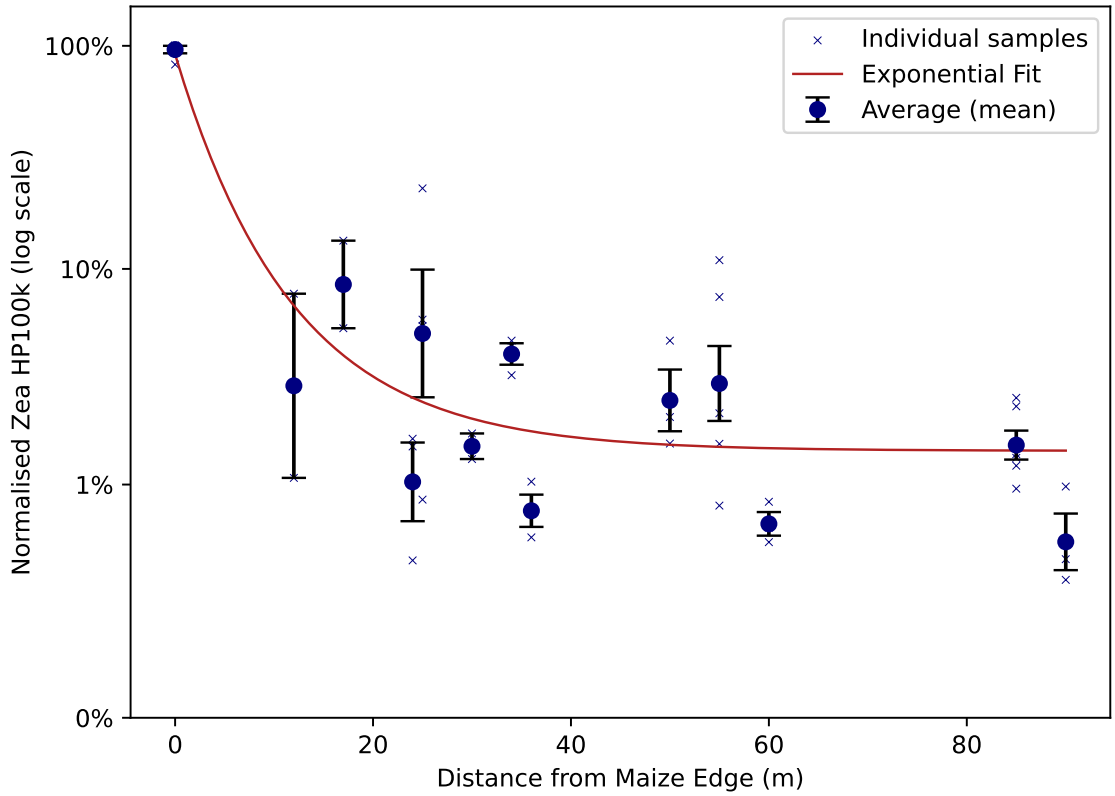


Figure 4.20: Normalised *Zea* HP100k (% of central sampler per experiment) plotted against distance from the maize edge on a logarithmic scale. Crosses show individual samples; dark blue circles represent the mean and error bars show standard error. The red line indicates an exponential decay fit. Data from the V-shape experiments.

samplers (12 m), where values range from 1–10% of the central HP100k. The decline continues across the 10–25 m range, which also shows substantial variation in normalised values. Beyond 40 m, the curve begins to flatten, suggesting smaller relative changes in *Zea* abundance at greater distances.

The variation in values at similar distances likely reflects differences in sampler placement relative to wind direction and the maize plot, as illustrated in the earlier heatmaps (Figure 4.19).

Overall, the distance-based experiments demonstrate that AirSeq is capable of detecting airborne maize DNA at distances up to 100 m from the source. However, there is a marked decline in *Zea* abundance with increasing distance from the plot. Detection is also strongly influenced by environmental factors such as wind speed, wind direction, and the quantity of source material present, all of which affect the proportion of maize DNA captured in each sample.

4.5.2 Laboratory experiments

A series of laboratory experiments were carried out to optimise DNA extraction protocols for air samples, with the aim of minimising taxonomic bias. These experiments tested the effects of filter storage conditions (immediate processing vs freezing), the duration of mechanical lysis, and the inclusion of negative and positive controls during both processing and sequencing.

4.5.2.1 Effect of filter storage

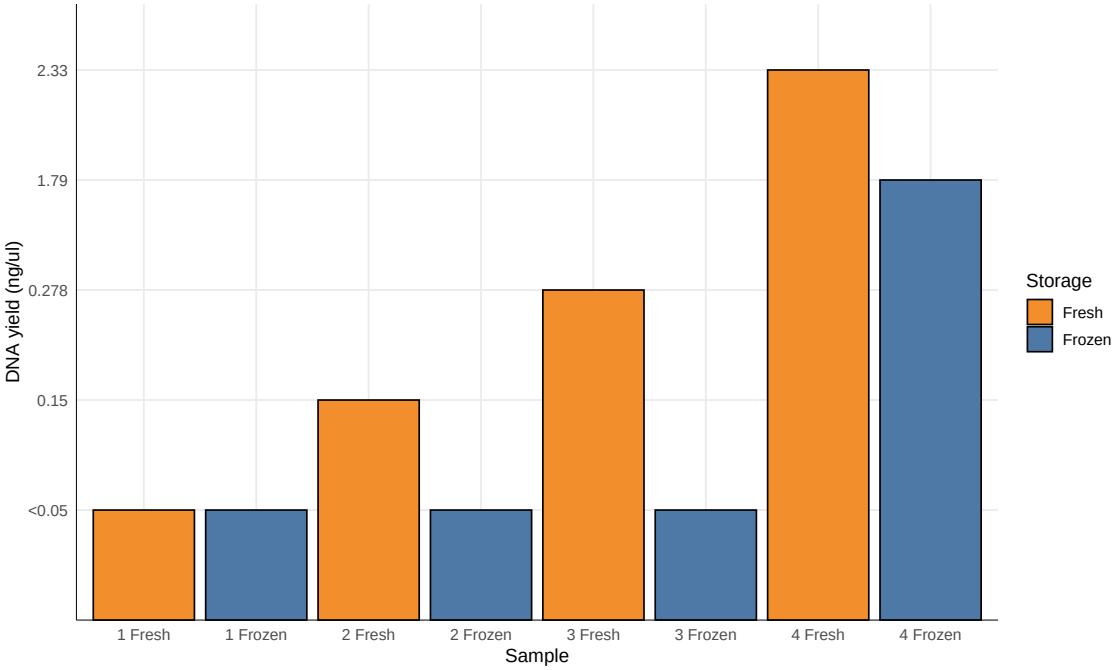


Figure 4.21: Bar chart showing the effect of storing filters in -80°C freezer for 27 days on DNA yield (ng/μL).

The filter storage experiment tested whether freezing filters at –80°C prior to extraction affects DNA yield. Coriolis µ samples were split and each half was filtered, with one processed immediately and the other frozen for one month. DNA yields were consistently lower in frozen samples, indicating that immediate processing is preferable to ensure sufficient concentrations for sequencing without amplification (Figure 4.21).

Because this experiment was conducted with field-collected samples, overall DNA yields were low. Sample 1 yielded DNA below the detection threshold when processed fresh ($<$0.05 ng/μL) and remained undetectable after freezer storage. Samples 2 and 3 produced low initial yields ($<$0.3 ng/μL), which fell below the detection threshold following storage. In contrast, Sample 4, which had the highest initial yield (2.33 ng/μL), showed a reduction after freezing (1.79 ng/μL) but still exceeded the yields of all other fresh samples.

These results suggest that -80°C storage reduces DNA yield, particularly in low-yield samples. However, if initial concentrations are sufficient, post-storage yields may remain within usable limits for downstream processing.

4.5.2.2 Mechanical lysis (bead beating)

The impact of bead beating duration on DNA extraction was tested using a mock microbial community lysed for 20–600 seconds, with sequencing used to compare read length (N50) and species composition across treatments.

Across the tested durations, N50 generally decreased as bead-beating time increased (Figure 4.22A), consistent with increased mechanical lysis leading to greater DNA fragmentation. An exception was observed at 40 seconds, where N50 values were unexpectedly lower than at 60 seconds.

Bead-beating duration influenced the relative abundance of species detected in the mock community (Figure 4.22B). Observed proportions were compared to the theoretical

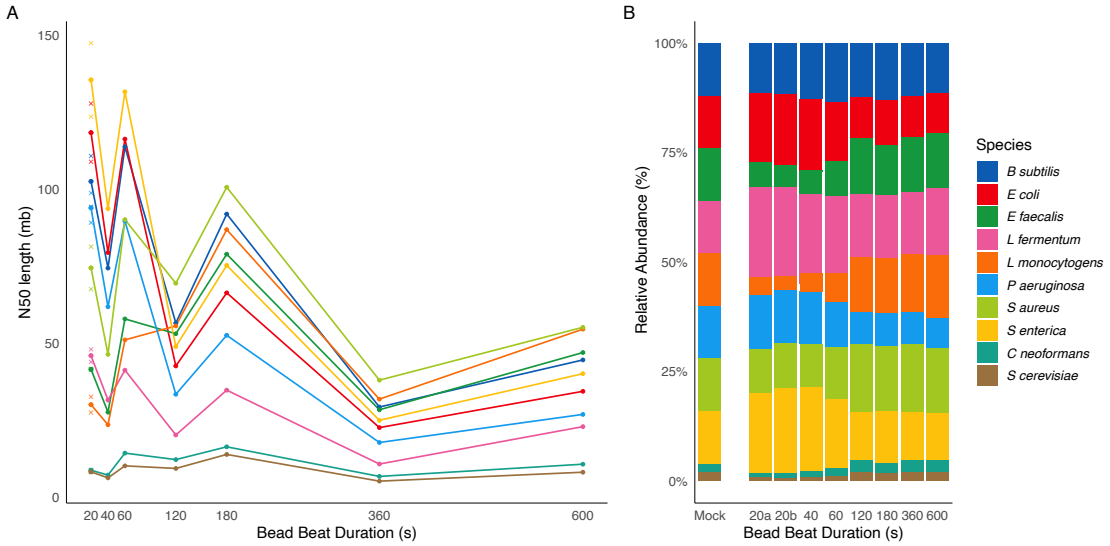


Figure 4.22: Effect of bead-beating length on mock community. A) N50 read length per species across bead-beating durations. The 20-second time point is plotted as an average, with each replicate shown as a cross. B) Corresponding relative species composition across bead-beating durations. Mock composition represents expected proportions in the control mixture.

composition shown in Table 4.7, where each bacterial species is expected to comprise 12% of the community and each yeast species 2%.

Following the standard 20-second bead-beating step, all species were detected; however, *L. fermentum* and *S. enterica* were overrepresented relative to their expected proportions (20.6% and 18.7%, respectively). In contrast, *E. faecalis* (5.2%) and *L. monocytogenes* (3.6%), as well as the yeasts *S. cerevisiae* (0.9%) and *C. neoformans* (1.0%), were underrepresented.

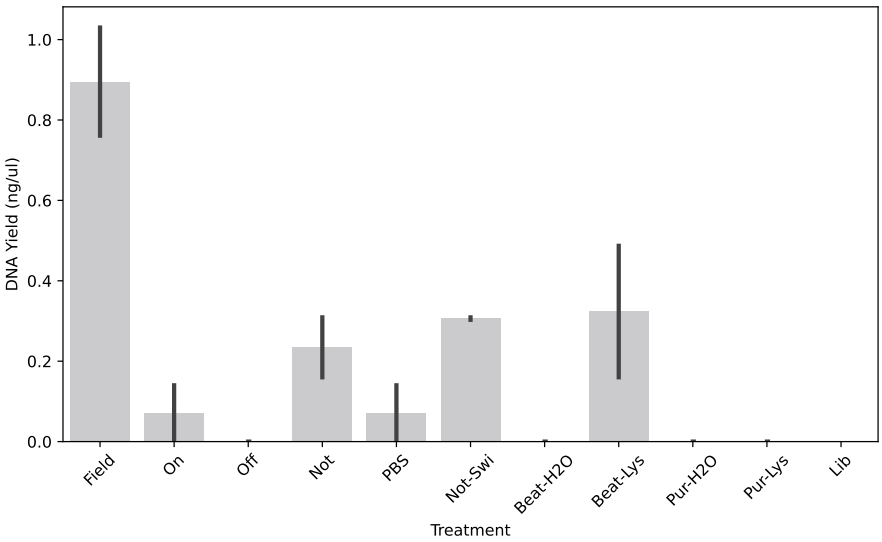
Longer bead-beating durations (≥120 s) produced more consistent profiles, with bacterial species ranging from 7.3%-15.4% and the yeasts between 2.0% and 2.8%. Despite the improved balance at extended lysis durations, some discrepancies persisted. At 600 seconds, *E. coli* and *P. aeruginosa* were detected at lower-than-expected levels (9.1% and 7.0%, respectively), while *L. fermentum* remained overrepresented (15%).

4.5.2.3 Control experiments

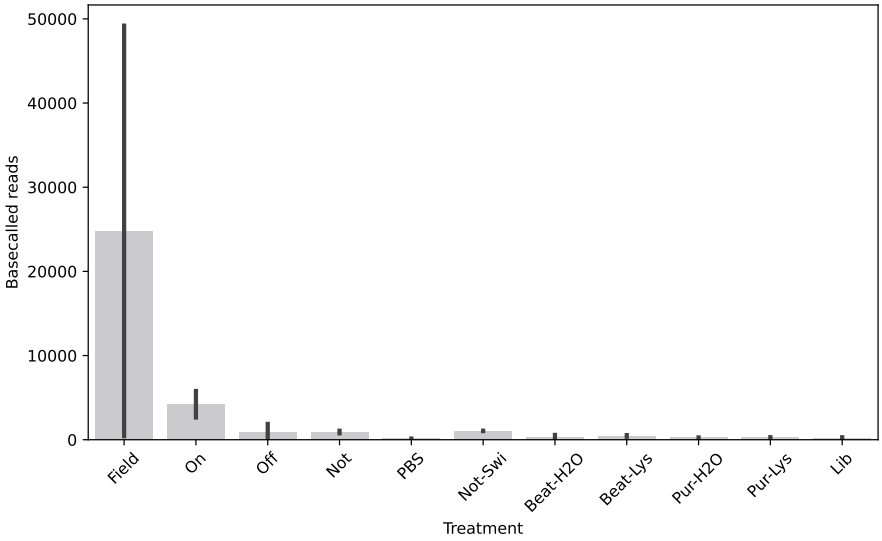
Negative controls The negative control experiment tested 11 stages of the AirSeq workflow (Figure 4.1), assessing potential contamination by comparing DNA yield, read counts, the proportion of *Homo sapiens* reads, and the ratio of classified to unclassified reads.

There is generally a proportional relationship between the number of basecalled reads and passed filter reads (Figure 4.23). However, DNA yield does not consistently correlate with either basecalled or passed filter reads. Across all of the negative controls, there was very little sequencing data generated.

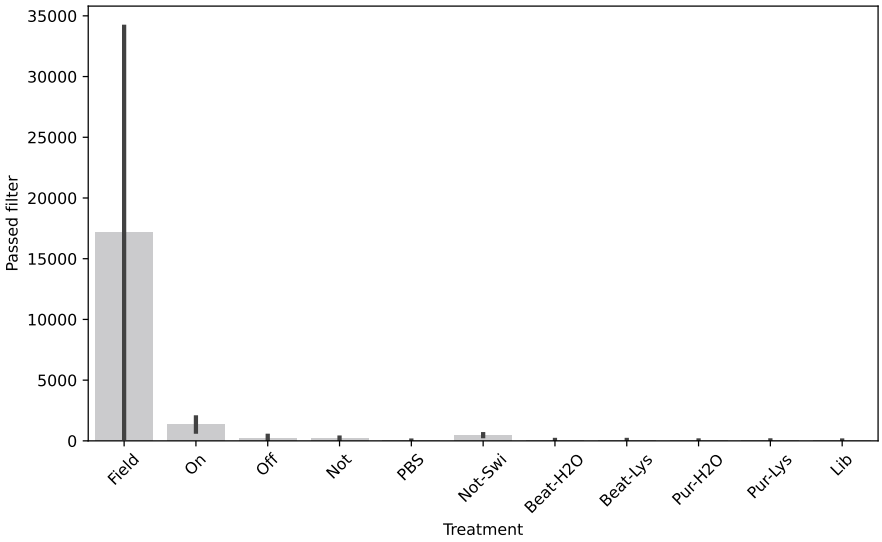
Field samples, which served as positive controls, produced the highest DNA yields (0.76 and 1.03 ng/μl), as well as the highest numbers of basecalled and passed filter reads. One replicate sequenced substantially better than the other (49,191 vs. 430 reads), in both cases, less than half of the reads were removed after filtering on read quality in MARTi (30% and 43%). As positive controls, the high yields and sequencing output from the Field samples are reassuring.



(a) DNA yield



(b) Basecalled reads



(c) Passed filter reads

Figure 4.23: Bar charts showing the average DNA yield, number of basecalled reads, and number of passed filter reads for each method, based on two replicates. Error bars represent standard deviation and are shown in black.

The "On" samples (sampler switched on in the lab) had the next highest sequencing success, with 5,786 and 2,637 basecalled reads and 1,931 and 759 passed filter reads. Despite low DNA concentrations (0.14 ng/μl and undetectable), these samples sequenced well, suggesting that high-quality DNA can still be captured from air even at low concentrations, although there was a lower proportion of reads passing filter. Interestingly, the "Not", "Not-Swi", and "Beat-Lys" samples contained more DNA, but their sequencing results were less successful, indicating that the quality of DNA may be more important than quantity in this context.

The "Off" samples had undetectable DNA concentrations but showed contrasting sequencing results: one replicate produced 1,880 basecalled reads (416 passed filter), while the other produced just 17 reads, only one of which passed filter. The "Not" samples (filter eluted without being placed in a sampler) had moderate DNA yields (0.16 and 0.309 ng/μl) but relatively low sequencing success, with few basecalled reads (1,067 and 761) and high read failure rates (74.7% and 91.85%). Among "PBS" samples (PBS passed through PVDF membrane), only one had measurable DNA (0.14 ng/μl). Both replicates yielded few basecalled reads (138 and 132), with high failure rates during filtering (82.6% and 97.7%).

"Not-Swi" (filter in tube with lysis buffer and beads) samples had moderate DNA yields (0.2 ng/μl) and low read counts (1,075 and 1,029), with approximately half passing filter. The "Beat-H_2O" samples (beads and water only) had no measurable DNA. One replicate yielded 569 reads and the other only 55, with over 85% of reads failing to pass filtering. Both "Beat-Lys" samples (beads and lysis buffer) contained measurable DNA (0.5 and 0.16 ng/μl), but produced few basecalled reads (549 and 353) and high failure rates (91.6% and 78.8%).

"Pur-H_2O" and "Pur-Lys" samples, which included magnetic bead purification steps, had no detectable DNA. They generated very few reads (229 – 301 basecalled) with high loss rates, only 24 – 34 reads passing filter in "Pur-H_2O", and 87 – 91.5% failure in "Pur-Lys". "Lib" samples, where water replaced DNA in library preparation, yielded low numbers of basecalled reads (289 and 49). Most of these did not pass filter (90% and 88% failure).

***Homo sapiens* contamination** The proportion of *Homo sapiens* reads can serve as a proxy for known contamination. While a small amount of human DNA is expected in airborne samples, high proportions may indicate contamination. Figure 4.24 shows the average proportion of human reads across samples.

The "Field" samples, which serve as positive controls, contained no detectable human reads. This is reassuring, as it suggests low levels of contamination under field conditions. In contrast, all other samples showed relatively high proportions of human reads. However, these samples also had far fewer total reads passing filter, so even a small absolute number of human reads can represent a large proportion of the total. For instance, 10 human reads would be negligible in a sample with thousands of passed reads but would dominate a sample with only a few dozen.

The "Off" samples collected in the laboratory had the highest proportion of *Homo sapiens* reads, likely due to exposure in an enclosed workspace where airborne human DNA is more prevalent. However, the "On" samples, which were collected in the same laboratory, did not exhibit similarly high levels of human reads.

Overall, the elevated proportions of human reads in all samples except the "Field"

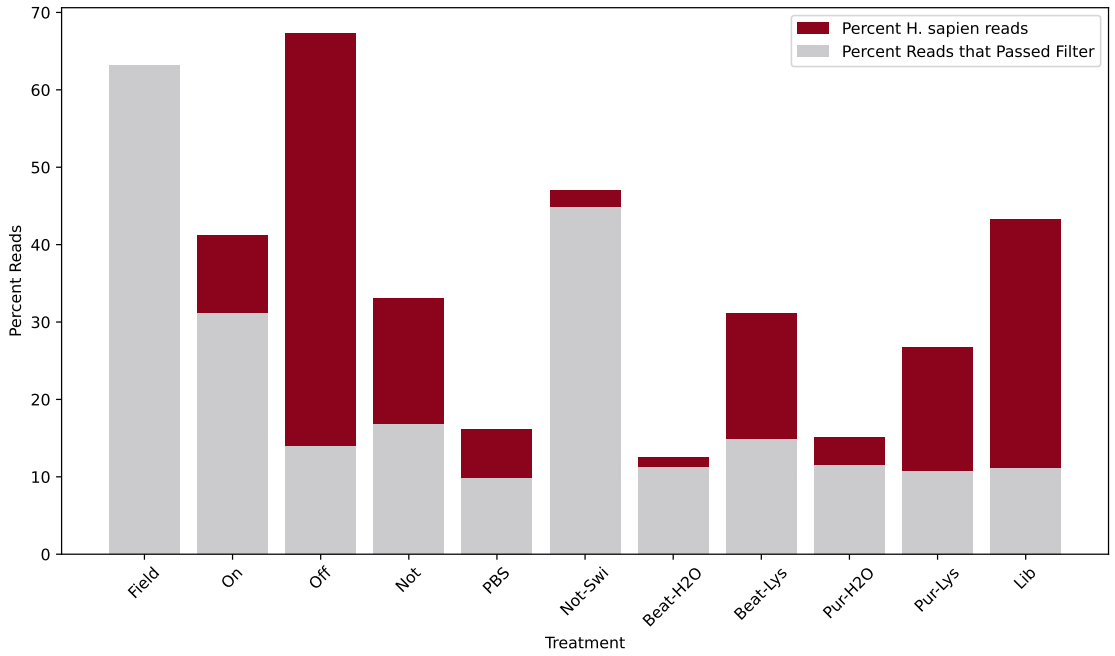


Figure 4.24: Stacked bar chart showing the percentage of reads that passed filter and the percentage of reads that were *H. sapiens*.

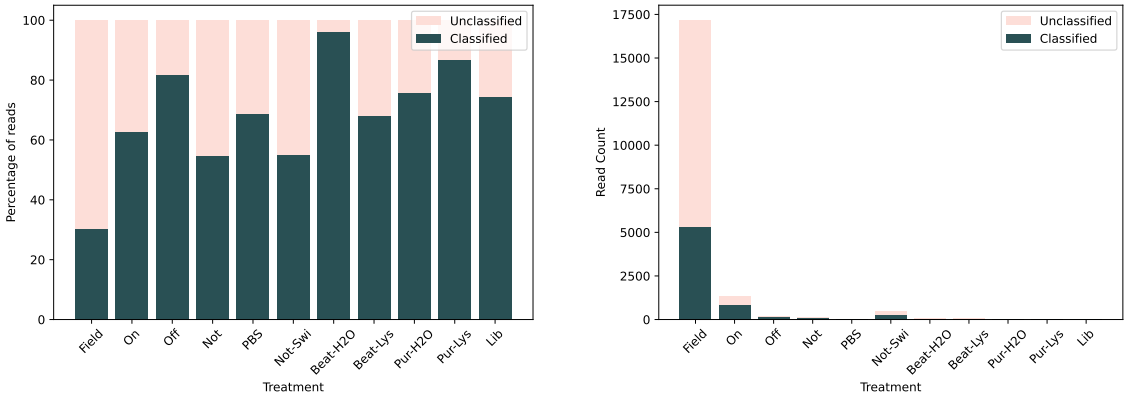


Figure 4.25: Bar charts showing the proportion of Classified and Unclassified Reads as percentage (left) and count (right) data.

control suggest that when other DNA sources are minimal, human contamination becomes more apparent. However, this appears to be predictable laboratory contamination and is unlikely to interfere with downstream analyses focused on plant pathogens.

Proportion of classified & unclassified reads Figure 4.25 shows the number of classified and unclassified reads across all samples. The "Field" samples have the highest proportion of unclassified reads, which is expected, as outdoor environmental samples often contain a broad diversity of sequences that are difficult to classify. In contrast, the "Beat-H_2O" samples show the highest proportion of classified reads, although this is based on a small number of total pass filter reads (5 and 76).

All of the control samples contain far fewer reads than the "Field" samples. As a result, the proportions of classified or unclassified reads in these low-read samples are less meaningful, as small differences in read numbers can greatly distort percentage-based comparisons.

Table 4.14: Results from Lambda experiment using *minimap2* alignments

Sample	**Passed Reads**	**Lambda Alignments (MQ > 5)**	**Unique Lambda Alignments (MQ > 5)**	**% Passed Reads**
Lambda	419,657	426,585	402,330	96
Water	225	182	165	73
Negative	91	88	81	89

Positive control The number of pass reads in the water and negative control samples is substantially lower than in the lambda sample, which is reassuring and suggests minimal cross-contamination between samples (Table 4.14).

The number of lambda alignments exceeds the number of passed reads used in the mapping (Table 4.14). This is because some reads align to multiple regions of the lambda genome from different positions, resulting in multiple alignments per read. To address this, unique read identifiers were used to count only one alignment per read. This approach produces a more accurate estimate of the number of distinct reads mapping to the lambda genome, which is appropriately lower than the total number of passed reads.

The majority of reads in the water and negative sample were assigned to the lambda genome, providing clear evidence of barcode cross-talk when samples were pooled. As lambda DNA is not expected to occur in either of the controls, its presence in these samples is likely due to cross-contamination from the lambda sample during barcoded sequencing.

To prevent barcode cross-talk from influencing results in the future, negative controls will be sequenced separately on a Flongle flow cell.

4.6 Discussion

The results presented in this chapter demonstrate that each component of the AirSeq pipeline can significantly influence the detected airborne taxa within a sample, underscoring the importance of experimentally validating protocol steps. Key variables include the choice of air sampler, proximity to the biological source, subsequent sample storage and processing and inclusion of negative and positive controls.

While other stages of the protocol could benefit from further optimisation, the foundational AirSeq method has already been validated for plant pathogen detection [124], as well as in the previous chapter of this thesis (Chapter 3). The experiments presented here were therefore designed to build on this foundation, refining the method to improve cost-effectiveness and enhance the detection of fungal taxa whilst retaining unbiased samples. Additionally, a better understanding of external factors such as distance from source was necessary to inform future sampling campaigns and ensure robust experimental design.

The first set of experiments focused on optimising the sample collection stage of the AirSeq pipeline. This included comparing different air samplers and determining the detection range from a known biological source. Following on from this there were experiments conducted on the filter storage, DNA extraction and sequencing stages of the protocol.

4.6.1 Field experiments

4.6.1.1 Sampler comparison

The sampler comparison experiment aimed to identify an affordable, portable air sampler capable of collecting sufficient DNA from 25- or 50-minute collections. By comparing the taxa detected across different samplers at the same time and place, the goal was to select an instrument that captured airborne communities without introducing undue bias.

In this experiment five air samplers were compared, Coriolis Compact, Coriolis µ, InnovaPrep Bobcat and Cub and SASS 4100, operated at their maximum flow rate for both 25 and 50 minute sampling durations at two sites (Church Farm and NHM). The comparison included assessments of DNA yield, as well as taxonomic community comparisons based on phylum-level abundance, taxonomic richness, detection of unique species, and the top five most abundant species per sample. Ordination analyses of airborne communities at the species level were also performed. These analyses demonstrated differences between samplers, with no single device consistently emerging as the most effective.

The observed differences in DNA yield between air samplers largely reflect variation in flow rate and, consequently, the volume of air sampled. Across all devices tested, there was a clear trend of increasing DNA yield with greater air volume (Figure 4.6), consistent with previous studies on bioaerosol sampling and DNA recovery [106, 226]. Furthermore, the reduced variation in DNA yield between samplers after normalising for air volume is consistent with earlier findings [250].

Despite operating at a much lower flow rate (300 Litres per Minute (L/min)), the Coriolis µ produced DNA yields comparable to the SASS 4100 (4,000 L/min) (Figure 4.6A). The highest-yielding Coriolis µ samples, however, were contaminated with insects, and once these were excluded the SASS clearly outperformed the Coriolis µ (Figure 4.6B). In contrast, after both air-volume normalisation and removal of contaminated samples, the Coriolis µ achieved the highest yields of all samplers (Figure 4.7B). These results show that while flow rate is a major driver of DNA recovery, collection mechanism and sample contamination also strongly influence performance. Notably, on a per-litre basis the SASS yielded less DNA than the Coriolis µ, Bobcat, or Cub, indicating that the high DNA yields are primarily due to the large volume of air processed rather than a different collection mechanism.

The InnovaPrep Bobcat and Cub both produced DNA yields sufficient for downstream metagenomic sequencing when run at their highest flow rate of 200 L/min for 25 minutes. Longer collection times led to higher yields, suggesting that extended sampling can help offset the effects of lower flow rates. Nevertheless, in dynamic airborne environments, the ability to recover high absolute quantities of DNA within short timeframes remains important. In this context, the SASS's high throughput allows for effective DNA capture in time-sensitive applications. However, its large size and limited portability constrain its usefulness in field-based studies. The Cub sampler offers a practical alternative, being lightweight and portable, with a battery life of approximately 4 hours at 200 L/min, which facilitates deployment in outdoor environments. Although absolute DNA yields may be lower, the consistent taxonomic detection observed suggests it is well suited to applications where portability is prioritised over short collection times, such as within the AirSeq pipeline.

Furthermore, the Coriolis Compact, which operates at 50 L/min, consistently produced

negligible DNA yields, even when sampling for 50 minutes. While this sampler employs dry cyclone technology and has been used successfully in previous studies alongside the SASS and Coriolis µ [95, 121, 339], those studies either used PCR amplification [95, 339] or cultured the samples before analysis [121]. These approaches require significantly less DNA input than WGS, which is the basis of the AirSeq pipeline. Consequently, the Coriolis Compact's low flow rate remains a fundamental limitation. One study demonstrated that a one-hour Coriolis Compact collection yielded fewer fungal particles and lower species richness than a 10-minute Coriolis µ sample [121]. Although the Coriolis Compact is compact and compatible with some airborne detection protocols, its limited DNA recovery renders it unsuitable for use with AirSeq.

In addition to assessing DNA yield, it is crucial to compare the taxonomic profiles recovered by each air sampler to determine whether specific instruments introduce bias, particularly if they fail to detect taxa identified by others at the same location and time. One challenge in interpreting this comparison is distinguishing between taxa uniquely identified by a single sampler due to genuine biological rarity versus misidentification. This issue is evident at the phylum level, where Cyanobacteria and Myxococcota reads were only detected in a Coriolis Compact sample (Figure 4.9). Furthermore, examination of the top five most abundant species in that same sample (Figure 4.12) reveals poor correlation with the other samples collected at the same site. Overall, Coriolis Compact samples consistently showed the least overlap in the top five species compared to those detected by other samplers, further supporting the conclusion that the identified taxa were not abundant and the Coriolis Compact is not suitable for AirSeq applications.

In contrast, the Bobcat, Cub and SASS samplers demonstrated strong consistency at both the phylum level (Figure 4.9) and in the top five most abundant species (Figure 4.12). Approximately half of the samples collected using the Coriolis µ were contaminated with insects, which skewed the phylum-level composition toward Arthropoda and may have reduced the amount of sequence data from fungal pathogens. Nonetheless, Coriolis µ samples without insect contamination showed comparable dominant species profiles to those of the other samplers. This suggests that, if a method were developed to remove or filter insects prior to DNA extraction, without biasing the airborne microbial community, the Coriolis µ could be a viable option for future AirSeq studies. However, the instrument design also imposes limitations. The Coriolis µ has a unidirectional inlet that must be oriented into the wind, but the small opening can be difficult to align, particularly under changeable field conditions. In addition, its wet cyclone mechanism can be compromised on hot days when the collection liquid evaporates, reducing the achievable sampling duration.

Taxonomic richness (i.e. the number of unique taxa per 100,000 reads) was comparable between samplers at both genus and species levels, indicating that despite differences in sampled air volume, a similar diversity of taxa was recovered (Figure 4.10). However, richness at the phylum level varied more noticeably between samplers, largely due to the absence of Bacillota, Oomycota and Bacteroidota in most Coriolis µ and SASS samples (Figure 4.9).

Although unique species were detected by all samplers, they generally accounted for only a small proportion of total reads (Figure 4.11B). This suggests that their contribution to overall taxonomic profiles was limited, with the exception of the Coriolis µ and SASS, where unique species represented a somewhat larger fraction. Closer inspection of the taxa responsible (Table 4.10) shows that these were mostly non-fungal, consisting primarily of

insects, for the Coriolis µ, and bacteria, for the SASS. The Cub was the only sampler to detect a unique fungal species. Collectively, these results indicate that the unique detections are unlikely to represent fungal pathogens that were missed by other samplers, which is the main focus of AirSeq.

The apparent differences in unique species counts between samplers should not be over-interpreted as evidence of greater sensitivity or taxonomic breadth. The Bobcat and Cub reported very few unique species, yet both consistently detected the dominant taxa also captured by other samplers, indicating they remain effective despite low unique counts. The relatively high number of unique species recorded by the SASS likely reflects its larger air volume, which in this experiment primarily increased the detection of bacterial taxa rather than fungal spores. Without a ground truth, it is difficult to determine whether these differences represent real biological variation or sampler-specific biases, but the overall pattern suggests that unique species contribute little to the fungal pathogen signal of interest.

Given that unique species contributed little to the overall signal and were rarely fungal, it is also important to examine the dominant taxa identified by the different samplers. To this end, the five most abundant species detected by each sampler were identified (Figure 4.12).

The comparison of abundant species detected across samplers revealed notable variation in taxonomic composition. The Coriolis Compact sampler recovered a distinct set of taxa, including some uniquely detected species such as *Bradyrhizobium sp. PSBBO68*, and consistently showed low overlap with other samplers, except for *Ustilago hordei*. The divergence observed in the species detected by the Coriolis Compact sampler may be attributable to the specific characteristics of dry cyclone sampling technology. Its lower air intake and potentially different particle capture profile may bias species recovery compared to higher-flow rate or filter-based methods.

In contrast, the Bobcat, Cub and SASS samplers detected a more diverse and overlapping suite of species, including common environmental taxa such as *U. hordei*, *Urtica urens*, and *Lolium perenne*, suggesting greater consistency and sensitivity in capturing biologically relevant signals. The similarity in species detected by the SASS and the InnovaPrep samplers (Cub and Bobcat) further supports the view that these devices are capturing meaningful biological signals, rather than reflecting overlap solely from shared sampling technology. Nonetheless, all three employ dry filter collection, with the SASS using an electret filter similar to those of the InnovaPrep samplers (Table 4.4). Thus, some of the observed similarity may be attributable to filter design, and testing a wider range of sampler types would be a useful next step.

The Coriolis µ samples were the most contaminated with insects (Table 4.9), suggesting that the sampler design, particularly the wet cyclone mechanism, make it more prone to incidental insect capture. Consequently, many samples were dominated by insect DNA and their symbionts, especially at the Church Farm site. When such contamination was absent, however, the species profiles were more consistent with those obtained from the other samplers, indicating that the Coriolis µ can recover representative taxa when not confounded by insect capture.

Ordination analysis suggests that sampling location had a greater influence on the detected community composition than the choice of air sampler (Figure 4.13). This contrasts with findings from a previous comparative study [155], which concluded that sampler type

exerted a stronger influence than collection location. However, that study was conducted indoors using samplers with more divergent collection mechanisms, which may have had a greater effect on the community profiles than the relatively similar samplers compared here.

There are limitations to the sampler comparison study. For each sampling duration and location, only two replicates were collected per sampler, and these were not taken simultaneously. This limits the ability to detect consistent patterns between samplers, as temporal variation may influence results. In addition, the different flow rates of the samplers led to substantial variation in the volume of air sampled, which may influence observed taxonomic diversity. A potential follow-up study could standardise air volume by varying the sampling duration, although this would introduce a separate variable, exposure time, which may also affect the communities detected.

In conclusion, based on the comparison of five air samplers, the InnovaPrep Cub is recommended for future AirSeq experiments. The Cub consistently yielded sufficient DNA for WGS, and its samples did not suffer from insect contamination, unlike those collected with the Coriolis µ. Taxonomic profiles at the phylum level were comparable to those of other samplers, and the top five most abundant species generally aligned with those from other devices, with the exception of a single outlier. Although the SASS and Bobcat samplers also performed well, the Cub is significantly more affordable, priced between $1,500 and $2,500 compared to over $10,000 for the others, and is highly portable making it well suited to field deployment.

4.6.1.2 Distance from maize source

The aim of the distance-from-source maize experiment was to assess the sensitivity of the AirSeq method, to compare its performance with previous studies that primarily employed passive traps and microscopy for pollen quantification, and to evaluate the influence of wind speed and direction on maize pollen dispersal distance. Samples were collected within and surrounding a 10 × 10 m maize plot, spanning distances from 10 to 100 m. A control sequencing experiment was conducted to assess barcode cross-talk. Temporal and diurnal variation in maize pollen release was monitored using a sampler positioned within the plot. Data from the 10 m, 100 m and downwind V-shaped transects were analysed in relation to concurrent wind conditions and the abundance of maize pollen released at the source during collection.

Barcode cross-talk was evident when negative control samples were sequenced alongside experimental samples (Figure 4.14). Negative controls exhibited an unexpectedly high proportion of maize reads, suggesting that DNA from other samples had been incorrectly assigned to the controls. Cross-talk was more pronounced in the negative controls than in the lambda positive controls.

This difference may be related to the DNA pooling process during ONT library preparation. Barcodes and adaptors are added to each sample individually and incubated before all samples are pooled. It is possible that, during pooling, the negative samples, containing negligible DNA, still carried free adaptors that subsequently bound to unbound DNA from other samples. In contrast, the lambda controls likely had sufficient DNA to bind adaptors prior to pooling, reducing the chance of cross-sample contamination.

Therefore, it can be concluded that the experimental samples, which contained DNA

volumes comparable to the lambda controls, were minimally affected by barcode cross-talk. As such, cross-talk is unlikely to have impacted the validity of the results.

Maize pollen abundance within the source plot showed substantial variation both across and within sampling days (Figure 4.15), consistent with findings from previous microscopy-based studies using passive samplers [157, 174, 381]. Although the exact timing of peak pollen release can vary between years and individual days, it is typically reported to occur between 09:30 and 10:30, followed by a sharp decline in the afternoon and no release overnight. Pollen shedding requires dry tassels and is therefore highly dependent on relative humidity, wind speed and temperature [381].

In this study, daily samples were collected over a fortnight between approximately 09:00 and 13:00. Peak pollen release was observed at or before 09:00 on several days, with considerable variability from day to day.

Fluctuations in maize pollen release during the sampling period have the potential to introduce bias into the experiment, particularly as the 10 m samples were typically collected first each day, coinciding with peak pollen release (Figure 4.15). To minimise this source of bias in future experiments, the order of sampler deployment could be alternated each day, thereby balancing the environmental conditions under which each sampling distance is assessed.

To account for this variability, normalising the data based on daily pollen abundance is appropriate (Table 4.12). This approach aligns with previous studies, which have normalised pollen counts either by the total pollen collected on a given day or by the yield from a canopy-level sampler [157, 381].

When the data are normalised to the central sampler, the results indicate a strong spatial decline in maize signal with increasing distance from the source (Table 4.12). The relatively wide range observed at 10 m suggests localised variability in airborne particle dispersion, potentially influenced by micro-environmental conditions or turbulence. In contrast, the consistently low values at 100 m reflect a steep decline in maize pollen detection with distance, likely due to the limited dispersal range of maize pollen.

There was a substantial difference in the abundance of maize pollen detected between 10 and 100 m. At 10 m, all surrounding samplers consistently detected maize pollen, with downwind samplers generally recording higher levels. In contrast, samplers positioned at 100 m detected very low levels of maize, likely due to the greater distance from the source. At this distance, wind direction appeared to have little effect on which samplers detected maize, suggesting that detections may have been influenced by external sources rather than the experimental plot. The V-shaped sampling transects indicated that, beyond 10 m, wind direction had a greater influence on pollen detection than distance alone.

The observed decline in maize pollen detection at distances beyond 10 m from the source plot is consistent with previous research on maize pollen dispersal. One study reported a decrease in pollen abundance with increasing distance from the source, with low levels detectable up to 4.45 km away [154]. Another study, using a dense array of sticky traps, found little to no pollen deposition beyond 100 m from the source field [21]. Similarly, research by Jarosz et al. estimated that approximately 95% of maize pollen is deposited within 10 m of its source [174], while earlier work found that only 5% of pollen remained airborne at 60 m compared to levels measured at 1 m [303]. Collectively, these studies support the conclusion that the majority of maize pollen is deposited within 10-100 m of the source. However, they also indicate the presence of a long dispersal tail, with a

small proportion of pollen capable of travelling much greater distances [24].

The distance-from-source experiment had certain limitations. Despite sampling over multiple days, substantial variation in environmental conditions and pollen release limited the comparability of samples. Other studies have addressed this challenge by employing season-long passive monitoring to account for fluctuations in pollen dispersal throughout the flowering period, offering more robust characterisation of dispersal patterns [154]. The short sampling intervals used in this study may therefore be less reliable for capturing representative dispersal dynamics. Furthermore, local geography, including hedgerows and topography, likely influenced wind flow and may have affected pollen transport [231]. Finally, while there were no other maize fields on the experimental farm, the possibility remains that pollen from external sources, such as other maize fields identified in the wider region (Figure 4.5), may have been detected, particularly given that long-distance dispersal of several kilometres has been reported [154].

In this experiment, the maize plot measured 10 × 10 m, which is considerably smaller than a typical agricultural field and may therefore limit the generalisability of the findings. Smaller plots have been criticised for potentially misrepresenting pollen dispersal dynamics; one study argues that they are more likely to suggest an exponential decline in pollen concentration with distance, rather than a power-law distribution that accounts for a long dispersal tail [24]. Supporting this, a study that sampled around full-scale agricultural fields detected maize pollen at distances up to 4.5 km and consistently observed deposition beyond 100 m from the source [154].

The comparatively low detection at 100 m in the present study may reflect the limited size of the source plot, but it could also be attributed to the short sampling duration, as only 1-hour active samples were collected. In contrast, the greatest detection distances in previous research were achieved using season-long passive sampling [154]. These extended collection periods, combined with the use of larger source plots, likely increased the likelihood of detecting low-abundance pollen at greater distances. The use of a small plot in this study was primarily determined by practical and financial constraints. Nevertheless, future work involving larger, real-world maize fields could provide more representative dispersal data, though such studies would face logistical challenges, particularly in gaining access to surrounding land to enable systematic sampling at defined distances.

Overall, this experiment demonstrates the potential of AirSeq to detect pollen at varying distances from a defined source and highlights the strong influence of wind direction and temporal variation in pollen release on dispersal patterns. The results align with established trends in the literature, including the rapid decline in pollen abundance beyond 10–100 m. This experiment provides valuable insight into both the capabilities and the practical considerations involved in applying AirSeq for pollen monitoring.

Future studies incorporating larger source plots, extended sampling periods and refined experimental designs will be essential for fully characterising maize pollen movement under real-world agricultural conditions.

4.6.2 Laboratory experiments

The final set of protocol refinement experiments focused on laboratory-based procedures, specifically the storage of sample filters, the mechanical lysis of cells and the inclusion of control samples. These experiments aimed to evaluate whether the existing protocols were

suitable for the unbiased detection of airborne pathogens.

The filter storage experiment revealed that storing filters at -80°C for three weeks led to a reduction in DNA yield; however, this effect was less pronounced when the starting material contained a higher volume of DNA. In the mechanical lysis (bead beating) experiment, longer lysis durations were associated with a reduction in the N50 of detected species, suggesting possible DNA fragmentation. Importantly, a 20-second bead beating duration was sufficient to recover all known species present in the mock community, indicating this to be an effective and balanced lysis time.

The negative controls did not reveal any consistent source of contamination in the laboratory protocol but emphasised the importance of including a negative control in all experiments. The positive controls confirmed the occurrence of barcode cross-talk during library preparation and sequencing.

4.6.2.1 Filter storage

Findings from the filter storage experiment align with previous studies; however, results across the literature vary. For example, one study reported no significant loss in DNA yield after five days of storage at –20 °C compared to immediate processing [226]. In contrast, another study observed a marked decline in yield beginning within the first week, with continued losses over a 12-week period depending on storage conditions [127].

In the present study, only a single time point was tested, comparing 0 days to 27 days of storage at –80 °C, and a noticeable decline in DNA yield was observed across all samples (Figure 4.21).

These differing outcomes highlight the complexity of DNA preservation and suggest that yield may be influenced by factors such as filter type, initial DNA concentration, storage duration and temperature, as well as handling protocols. Future work could include a time-course assay to identify the point at which DNA degradation begins, enabling clearer guidance on maximum freezer storage durations. Additionally, sequencing of the extracted DNA would be valuable to determine whether freezing introduces taxonomic bias, as certain species are known to degrade more rapidly under suboptimal storage conditions [127].

Including mock communities alongside field-collected samples would enhance repeatability and help distinguish true biological shifts from methodological artefacts. It is also worth noting that this experiment used Coriolis µ samplers rather than Cub samplers. Repeating the experiment with Cub samplers would help ensure the results are directly applicable to the latest AirSeq protocol.

4.6.2.2 Mechanical lysis (bead beating)

The bead-beating experiment showed that longer lysis steps reduced the N50 values of mock community taxa (Figure 4.22), indicating that extended lysis produces shorter DNA fragments, which may compromise read length and taxonomic resolution. This aligns with findings from a similar study [30], which demonstrated that both the duration and intensity of mechanical lysis have a significant inverse effect on DNA fragment length. In that study, shorter lysis times and reduced speed resulted in longer sequencing reads. The authors reported that 1900 CPM for 20–30 seconds using the SuperFastPrep-2^{TM} maximised fragment length without compromising DNA yield.

Although homogenisation speed was not assessed in the present study, the findings similarly support a shorter lysis duration, with 20 seconds at 2,300 CPM (speed setting 20) using the same device identified as optimal. This duration appears to provide a balance between effective cell lysis and minimal DNA shearing.

The anomalously low N50 observed after 40 seconds of bead beating, compared to 60 seconds, likely reflects experimental noise, potentially influenced by uncontrolled variables. This is particularly plausible given the absence of replicates, which limits the ability to distinguish genuine effects from random variation.

In addition to sequence length, the bead beating experiment indicated that extended lysis times affect taxonomic representation (Figure 4.22B). From the shortest 20 second duration all of the species are identified with an overrepresentation of *L. fermentum* (gram-positive) and *S. enterica* (gram-negative). Additionally, other Gram-positive bacteria are underrepresented, alongside the yeasts. This underrepresentation of Gram-positive species and yeasts is likely due to the cell wall structures that are harder to lyse [22, 291, 350, 353]. This is similar to other comparative DNA extraction studies which identified differential lysis efficiency between microbial groups based on cell wall composition [144].

As lysis duration increased, the observed species abundances more closely aligned with the expected proportions of the mock community. However, *L. fermentum* remained consistently overrepresented even after 600 seconds, while the Gram-negative species, *E. coli* and *P. aeruginosa*, were underrepresented. One possible explanation for differences in relative abundance is genome size, with larger genomes potentially yielding more sequencing reads. However, the mock community is standardised by genomic DNA input, and *L. fermentum* has the smallest genome of all taxa in the community (1.905 Mb), whereas *E. coli* (4.875 Mb) and *P. aeruginosa* (6.792 Mb) have comparatively larger genomes (Table 4.7). The discrepancies observed after 600 seconds are therefore more likely due to differences in lysis efficiency and DNA fragmentation. Notably, *L. fermentum* displayed one of the lowest N50 values among the bacteria (Figure 4.22A), indicating highly fragmented DNA. This fragmentation could lead to an inflated read count, as the greater number of shorter fragments increases the likelihood of sequencing relative to species with higher N50 values.

Following 20 seconds of bead beating, all species in the mock community, including the two yeasts, were detectable, with acceptable N50 values and relative abundances reasonably consistent with the known composition. Therefore, the mechanical lysis step in the AirSeq protocol will use a 20-second bead-beating duration. This approach acknowledges that while sufficient for species recovery, the relative abundances of detected taxa may not fully reflect true biological variation.

It should be noted that these experiments were conducted using mock communities in solution rather than samples collected on filters, as the latter introduced high levels of variability. Consequently, the results may not fully reflect conditions encountered in airborne sample processing. Future optimisation could involve repeating the experiment on filter-bound samples, to more accurately replicate laboratory workflows and assess the robustness of the extraction protocol.

Previous studies have shown that reliable DNA extraction from fungal samples typically requires extended enzymatic digestion, freeze–thaw cycles or mechanical disruption using glass beads [200]. This raises the possibility that a 20-second bead beating step, while sufficient for bacteria and yeasts, may not be adequate for consistent lysis of all fungal taxa encountered in airborne samples. However, field samples processed using the current

AirSeq protocol have successfully identified a range of fungal pathogens [124], and another study found that bead beating combined with lysis buffer was the most effective and efficient method for extracting fungal DNA [134]. While these findings suggest that the existing extraction method is broadly effective for fungal spores, further testing using a defined mock community of fungal spores would be beneficial to confirm consistent lysis across taxa and to rule out any extraction bias. This was beyond the scope of the present study due to the lack of affordable fungal mock communities suitable for repeated use.

4.6.2.3 Control experiments

Negative controls The inclusion of negative controls indicated potential contamination in the samples arising from laboratory processes (Figure 4.23a). However, DNA yields in these controls were considerably lower than those observed in field-collected samples. Moreover, when sequenced, these controls produced very few pass-filter reads (Figure 4.23) and a low proportion of classified reads (Figure 4.25). The presence of DNA in these samples is likely due to reagent contamination, commonly referred to as the "kitome", originating from DNA extraction kits and lysis solutions [319].

The results presented here suggest that the lysis solution may contribute to background contamination, as detectable DNA was present in samples where it was used (e.g. "Beat-Lys"). However, the absence of detectable DNA in other samples containing the same lysis solution (e.g. "Pur-Lys") indicates that it is unlikely to be a major source of contamination. A limitation of this experiment is the lack of a control containing lysis solution and a filter but excluding PowerSoil beads, which prevents a definitive assessment of potential contamination from other components, such as the beads. Given the negligible number of pass-filter reads in the control samples, the AirSeq protocol will continue to use the current lysis solution.

Despite the low number of classified reads, many negative control samples were dominated by human (*H. sapiens*) sequences, indicating that human contamination is the primary source of extraneous DNA (Figure 4.24). Contamination was highest in the "Off" samples and markedly greater than in the corresponding "On" samples collected at the same location while the sampler was actively collecting air. Furthermore, the "On" samples exhibited a lower proportion of taxonomically classified reads (40%) compared to the "Off" samples (68%), possibly due to their higher total read count. It is plausible that the additional reads in the "On" samples include a broader range of eDNA, which is more difficult to classify, whereas the "Off" samples may contain a higher proportion of easily identified human DNA. As the proportion of human reads was negligible in field-collected samples ("On"), this form of contamination is unlikely to substantially affect experimental airborne samples, which are dominated by environmental taxa.

Given that eliminating all sources of contamination in a laboratory setting is impractical, the inclusion of a negative control in every experiment remains essential. This negative control will be processed from a Cub filter that has never been placed in a sampler or removed from the laboratory, to ensure it is taken through every stage of the DNA extraction and library preparation pipeline. Alongside sterile working procedures, this allows for the detection of taxa likely introduced during sample processing, thereby improving accuracy of conclusions drawn from the data.

Positive control As described previously with the maize controls, barcode cross-talk can occur during ONT library preparation and sequencing. This may result in negative controls displaying a taxonomic composition similar to that of high-DNA samples, despite the absence of target material. Barcode cross-talk can also arise from sequencing errors; for example, one study reported that 0.056% of multiplexed reads were incorrectly assigned to barcodes during ONT sequencing [398].

To assess this risk, a positive control experiment was conducted in which lambda DNA, a negative filter control and a water blank underwent library preparation and sequencing. The results confirmed the occurrence of barcode cross-talk, as both the negative filter and water blank contained reads that aligned to lambda DNA (Table 4.14). Although the total number of aligned reads was low (182 and 88, respectively), they accounted for the majority of pass-filter reads in these samples (73% and 89%), indicating that cross-talk is detectable at low input concentrations.

Based on these findings, all future AirSeq experiments will include a lambda DNA positive control alongside the negative. These controls will be barcoded and sequenced separately from experimental samples to minimise the impact of barcode cross-talk on downstream analyses.

In summary, the protocol refinement experiments confirmed that the current DNA extraction and filter storage procedures used in the AirSeq pipeline are broadly effective for detecting airborne microbial communities, including fungal spores. Short-term filter storage at -80°C is feasible, particularly when starting DNA concentrations are high, though further time-course studies are needed to define safe storage durations. A 20-second bead beating step proved sufficient for recovering bacterial and yeast DNA without introducing excessive fragmentation, though its performance on filter-bound or fungal-rich samples remains to be fully validated. These refinements and future considerations will be critical for ensuring the robustness and reproducibility of the AirSeq workflow in routine pathogen surveillance.

4.7 Conclusion and Future Work

The goal of this chapter was to refine the AirSeq protocol across all stages of the experimental workflow, with the aim of enabling unbiased detection and identification of airborne fungal plant pathogens.

Initial sampler comparisons identified the InnovaPrep Cub sampler as the most suitable for these experiments due to its consistent taxonomic detection, portability and cost-effectiveness. In the maize dispersal study, pollen abundance was highest at 10 m and declined sharply towards 100 m away from the plot, with detection strongly influenced by wind speed and direction. Across all field trials, environmental variability emerged as a key factor affecting sample comparability.

Laboratory-based refinements showed that while filter storage at -80°C is feasible, it can reduce DNA yield which may influence downstream analyses. A 20-second bead beating step was also shown to be sufficient for recovering taxa in mock samples and presents a practical compromise between lysis efficiency and DNA integrity. Control experiments demonstrated the importance of including both positive and negative controls, ideally processed on separate sequencing runs, to account for barcode cross-talk.

Looking ahead, future work could explore additional air samplers as new technologies

become available, compare sequencing platforms and library preparation methods, and assess the performance of targeted amplicon sequencing versus WGS approaches. The protocol refinements presented in this chapter form a robust foundation for airborne fungal surveillance and are applied in the following chapter to monitor plant pathogens across two growing seasons in an agricultural setting.

Chapter 5

Season Monitoring

5.1 Abstract

The AirSeq method was applied to monitor airborne plant pathogens in an agricultural setting across two growing seasons. Regular sampling was conducted in an untreated wheat field during 2023 and 2024, with disease scoring from both years and disease observations from 2024 used to validate the method's effectiveness in detecting known pathogens. Environmental data were also integrated to assess whether airborne spore abundance aligned with conditions known to favour spore release and dispersal.

The analysis includes community-level comparisons at the phylum level from weekly samples, as well as detailed profiling of nine fungal and oomycete genera containing important wheat pathogens. AirSeq successfully detected key pathogens during periods when their presence would be expected based on pathogen biology and environmental conditions. In some cases, early-season detections corresponded with initial field symptoms, demonstrating the potential of this method for early warning and disease monitoring. These findings support the utility of airborne eDNA as a tool for pathogen surveillance and improving disease forecasting in crop systems.

5.2 My Contributions

Dr Darren Heavens and I both collected the samples from Church Farm in 2023 and I extracted the DNA from these samples and sequenced them. Throughout 2024 Dr Heavens collected the samples and processed them to sequencing on my behalf. Darryl Playford conducted disease scoring in the plots where sampling occurred. I carried out all of the bioinformatic work independently, with the assistance of ChatGPT (GPT-5) [273] to refine my code. I also used ChatGPT (GPT-5) to improve the grammar and flow of my writing.

5.3 Introduction

Building on the previous chapter (Chapter 4), the optimised AirSeq protocol was applied to monitor fungal pathogens in a wheat field over two consecutive growing seasons (2023 and 2024). This dataset provides a valuable baseline for understanding the airborne fungal microbiome and offers insight into the temporal dynamics of key fungal or oomycete genera. Additionally, it allows assessment of airborne eDNA abundance in the context of observed disease presence and severity. The same dataset was also mined for potential emergent pathogens using a custom bioinformatics pipeline, the results of which are presented in Chapter 7. Finally, such an extensive dataset also permits the evaluation of the potential of AirSeq as a tool for pathogen monitoring and disease management.

Wheat is a globally important crop, providing approximately 20% of the calories and protein consumed by the human population [69, 103]. Its production is threatened by insect pests, fungal pathogens and environmental stressors, all of which are likely to be exacerbated by climate change, through expanding host ranges and more frequent extreme weather events [341]. Although disease pressures vary by region and season, yield losses due to pathogens are estimated to range between 10–50% globally [69].

As illustrated in Figure 5.1, a large proportion of arable land in the UK has suitable climatic conditions for the development of five major wheat diseases: stem rust, stripe rust, leaf rust, Fusarium head blight, and Septoria tritici blotch [69]. Long-term monitoring data from England and Wales (1999–2019) demonstrate that wheat crops in the UK are affected by a wide range of fungal diseases, with notable shifts in prevalence over time [369]. Septoria tritici blotch was consistently the most prevalent and damaging disease throughout the monitoring period. In contrast, the incidence of powdery mildew declined significantly, while Fusarium head blight increased in both frequency and severity. Since 2009, tan spot has emerged as the third most common foliar disease affecting UK wheat crops.

These patterns highlight the importance of continued monitoring and improved understanding of wheat fungal pathogens in UK agroecosystems. Because pathogens often co-infect hosts, complicating diagnosis and disease management, simultaneous monitoring of multiple taxa is essential [1].

Given the importance of monitoring multiple pathogens simultaneously, this study specifically monitored nine agriculturally significant fungal or oomycete genera: *Blumeria*, *Claviceps*, *Fusarium*, *Magnaporthe*, *Parastagonospora*, *Puccinia*, *Pyrenophora*, *Ustilago*, and *Zymoseptoria*. While not all were detected in the dataset, each was included in the analysis due to their known relevance to wheat production. Further details on their dispersal mechanisms, causal diseases, yield impacts, regional distribution, seasonality, and optimal infection conditions are summarised in Table 5.1.

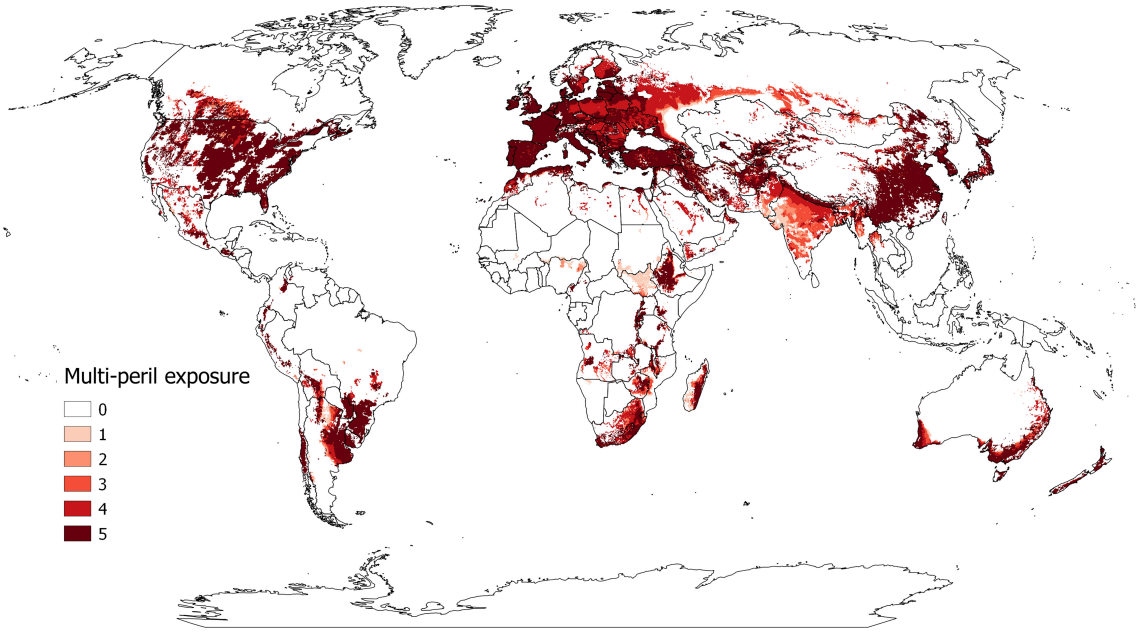


Figure 5.1: Map depicting the number of wheat fungal diseases with suitable climatic conditions (CLIMEX annual growth index >5) across global wheat production areas. The modelled diseases include stem rust, stripe rust, leaf rust, Fusarium head blight, and Septoria tritici blotch. Data are based on the Spatial Production Allocation Model by Yu et al. [407]. Reproduced from Chai et al. [69].

Table 5.1: Summary of fungal and oomycete genera with associated transmission, disease, and environmental conditions

Genera	Species	Dispersal	Disease Caused	Impact	UK Distribution	Optimal Conditions	Ref
Blumeria	*B. graminis* f.sp. *tritici*	Wind	Powdery mildew	Yield losses up to 20% in susceptible varieties; typically under 10%	Across the UK; highest in S & E England	15°C optimal; range 5–30°C; >80% humidity needed	[235]
Claviceps	*C. purpurea*	Wind, insect or splash	Ergot	Minimal yield impact; grain contamination poses health risks and may lead to rejection	Low UK presence; reports increasing in recent years	Cool, wet conditions	[99] [362]

Genera	Species	Dispersal	Disease Caused	Impact	UK Distribution	Optimal Conditions	Ref
Fusarium	*F. graminearum (and other spp.)*	Wind, insect or splash	Fusarium head blight / scab	Mycotoxin production and yield reduction; losses usually minor in the UK	Highest risk in SE England	25–30°C with wet conditions	[39, 117, 307]
Magnaporthe	*M. oryzae*	Wind or splash	Wheat blast	Severe yield losses (>50%) in affected regions	Not present in the UK; problematic in parts of the Global South	15–27°C with wet conditions or high humidity	[46, 88, 267]
Parastagonospora	*P. nodorum*	Wind, splash or seed-borne	Septoria nodorum / seedling blight	UK losses generally under 3% but potential for up to 50% loss	Across the UK; higher risk in SW England	20–27°C, high humidity (75–95%); rain triggers spore release	[186, 198, 328, 329]
Puccinia	*P. triticina, P. graminis, P. striiformis* f.sp. *tritici*	Wind	Leaf, stem and stripe rust	Can cause significant yield loss; severity varies by species and conditions	All present in the UK; highest risk in S & E England	15–22°C optimal; 7–25°C supports disease; high humidity required	[49, 57, 199, 388, 403, 410]
Pyrenophora	*P. tritici-repentis*	Wind or seed-borne	Tan spot	Occasional reports in the UK; rarely leads to serious yield loss	Recorded across the UK; typically low incidence	20–28°C with prolonged dew or rain; windy conditions favoured	[236, 358]
Ustilago	*U. tritici*	Wind or seed-borne	Loose smut	Can cause up to 40% loss	Low in the UK due to seed certification schemes	Cool, humid conditions with light rain or dew during flowering	[223, 224, 300]
Zymoseptoria	*Z. tritici*	Wind and splash-dispersed	Septoria leaf blotch	Losses of up to 50% in severely affected crops	Widespread in UK; highest in S & W England	15–20°C optimal	[329]

These genera were selected due to their inclusion of destructive wheat pathogens, four of which were featured in a global review of the top ten fungal pathogens: *Magnaporthe oryzae*, *Puccinia* spp., *Fusarium graminearum*, and *Blumeria graminis* [90], the ranking of which was based on scientific and economic importance as determined by votes from the international molecular plant pathology community. Additionally, all genera are at least partially wind-dispersed (Table 5.1) and are therefore well-suited for detection via the AirSeq method.

The chosen genera represent a broad spectrum of known prevalence within the UK. Including pathogens commonly encountered across the country, such as *Zymoseptoria* spp., *Fusarium* spp. and *Puccinia* spp., as well as those that are rare but present, including *Claviceps purpurea* (ergot) and *Ustilago tritici* (loose smut). Also of note is stem rust (*Puccinia graminis* f.sp. *tritici*), which was effectively eradicated in the UK during the 1960s but has re-emerged sporadically over the past 15 years, with a limited number of recent reports [323]. In addition, *Magnaporthe oryzae* pathotype *Triticum*, the causal agent of wheat blast, has not yet been detected in Europe but has spread from South America to Bangladesh and Zambia via international trade [343]. Its demonstrated capacity for transcontinental dissemination underscores the need for continued surveillance and preparedness in the UK.

Crop losses from these pathogens can be managed through chemical and cultural control strategies. In England and Wales, fungicide applications have risen over the past two decades [369], yet this reliance is unsustainable, driving resistance in many pathogens [82, 283, 360] and coinciding with the withdrawal of active ingredients on environmental grounds. Cultural practices, such as reducing the use of susceptible wheat varieties, are increasingly adopted [342, 369, 376, 383], but the development of resistant cultivars is slow. Taken together, these strategies remain insufficient to fully protect crops.

Accurate and early detection of pathogens would provide growers with actionable information, allowing targeted fungicide applications or cultural interventions at the appropriate time to prevent disease spread. Such an approach could minimise fungicide use, reduce environmental impact, and improve disease management.

Few studies have combined high-volume air sampling with WGS to track agriculturally significant fungal pathogens over an entire growing season. Most DNA-based research has instead relied on amplicon sequencing to survey environmental fungal communities [29], assessed overall fungal diversity without targeting specific taxa [138], or monitored individual pathogens in isolation [113]. AirSeq has previously been used to detect airborne fungal pathogens, but only across a six-week period and without visual confirmation of disease presence [124]. While valuable, these approaches often lack the taxonomic resolution or breadth required for comprehensive pathogen surveillance.

Building on prior work from this thesis, this chapter applies the improved pipeline to a season-long sampling campaign. Here, multiple pathogenic taxa were monitored across two growing seasons, with airborne signals compared to visual disease identification and scoring at the sampling site.

The aim of this chapter is to apply the AirSeq approach to detect airborne fungal pathogens and to evaluate the resulting data in relation to existing knowledge of spore dispersal dynamics. AirSeq successfully identified multiple pathogens across the growing season, with relative abundance patterns often aligning with disease observation reports, favourable environmental conditions, and known pathogen or host biology.

5.4 Methods

5.4.1 Field site

Samples were collected at the John Innes Centre field station, Church Farm (Bawburgh, Norfolk), which covers 110 hectares across 15 fields. These fields host a wide range of experimental trials, including seed multiplication, generation of breeding materials, agronomic and physiological studies, and disease monitoring in Fusarium nurseries. The layout of the trials in the sampling field during 2023 and 2024 is shown in Figure 5.2, together with the location of the sampler.

Sampling was conducted within the CIMMYT (International Maize and Wheat Improvement Center) plot. This plot was chosen because it was not treated with fungicides and contained a mixture of wheat varieties under evaluation for disease susceptibility. Consequently, a wide range of genotypes were present, each carrying different combinations of resistance alleles, providing an opportunity to relate AirSeq detections to visible disease development without the confounding effect of uniform resistance. In 2023, the CIMMYT plot contained 27 unique wheat crosses, increasing to 60 in 2024, of which six were present in both years.

The exact sampling locations differed by approximately 600 m between the two years, as the untreated CIMMYT wheat plot was located in different fields each season. These locations are indicated in Figure 5.3, with GPS coordinates provided in Table 5.2.

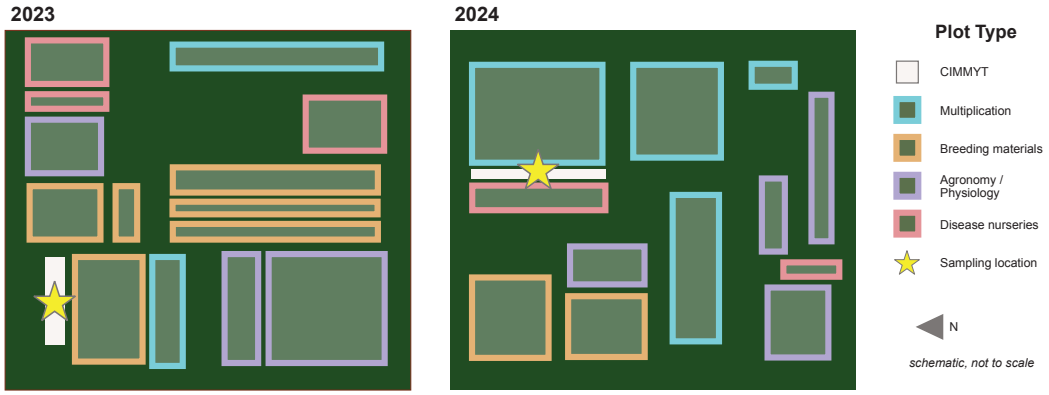


Figure 5.2: Schematic layout of Church Farm Fields where sampling took place in 2023 and 2024, illustrating the diversity of adjacent plots. Outlined blocks represent different trial categories (see legend). The schematic is not to scale.

5.4.2 Sample collection and processing

Air samples were collected using the InnovaPrep Cub sampler at the maximum flow rate (200 L/min) in 2023 and 2024. Collection periods for each year are given in Table 5.2.

Samples were collected in the morning between 09:30 and 12:30. All sampling was carried out in duplicate: in 2023, successive 30-minute samples were collected, while in 2024, simultaneous 120-minute duplicates were taken. On 6 June 2023, a single 60-minute sample was collected. Additionally, on 21 and 28 August 2024, single 1-hour samples were obtained using the InnovaPrep Bobcat sampler. The 30-minute sampling duration was originally chosen to capture 6,000 L of air per sample, aligning with the Coriolis µ collections described in Chapter 3. However, yields were later found to be insufficient,

Figure 5.3: Satellite view of Church Farm showing the sampling locations for 2023 and 2024 and a field where a suspected *Claviceps purpurea* infection occurred in 2024. Satellite image captured in June 2024 and downloaded from Google Earth Pro. Locations marked and labelled by author.

leading to the adoption of longer sampling periods in 2024.

Table 5.2: Summary of air sample collection and processing for 2023 and 2024

Parameter	**2023**	**2024**
Collection Period	Monthly (Jan 2023–Apr 2023), then weekly (25 May–17 Aug)	Weekly (Feb 2024–Oct 2024)
GPS Coordinates	52.62902° N, 1.17476° E	56.62659° N, 1.17997° E
Sample Duration	Two consecutive 30-minute samples	Two parallel 1-hour samples
No. of Samples	29	56
Whole Genome Amplification	Yes	No
Library Prep Kit	Rapid (SQK-RBK114-96)	Ligation (SQK-NBD114-96)
Flow Cell	MinION and PromethION	PromethION
MARTi version	0.9.23	0.9.20 and 0.9.23

Samples were processed as described in the previous chapter (Chapter 4, Section 4.4.1), following the standard protocol involving elution foam, syringe filtration, bead beating, and magnetic bead clean-up. For the 2024 samples, this protocol was modified as follows: after bead beating, the tube contents were transferred to a spin column and centrifuged at maximum speed for 1 minute. In addition, during the magnetic bead clean-up step, the pelleted beads were washed three times with 70% ethanol instead of the standard two washes. Finally, DNA was eluted into 21 μl of elution solution (CD6 from the PowerSoil Pro kit) rather than into water.

The 2023 samples underwent WGA prior to sequencing, whereas the 2024 samples did not.

Following DNA extraction and amplification of the 2023 samples, library preparation and sequencing were carried out using the standard ONT protocol. The specific library preparation kits used are listed in Table 5.2.

To complete the 2024 weekly dataset, additional sequence data were incorporated from experiments conducted at the same location (Church Farm) using the InnovaPrep Cub sampler. These included the 10:00–12:00 samples from the 24-hour experiment on 14 May and 18 June (described in detail in Chapter 6), as well as 1-hour samples from the NorfolkSeq study collected on 25 July and 17 October (described in Chapter 7).

A negative filter control and a positive lambda control were also prepared and sequenced alongside the experimental samples.

5.4.3 Environmental data

Meteorological data were collected using an on-site weather station, which recorded minimum, maximum and average daily air temperature, as well as hourly measurements of electrical conductivity, rainfall, barometric pressure and RH. Wind speed was recorded at ten-minute intervals. All data were exported as CSV files and used for subsequent analysis described in section 5.4.4.

Environmental data were quality-checked, after the identification of an anomalously high wind speed the wind speed data were filtered to remove any values exceeding 100 m/s (∼200 mph) as likely sensor errors. This threshold was selected based on realistic conditions, and caused the removal of 7 observations from the 2023 data. The remaining wind speed data were validated by comparison with records from the nearest publicly available station at Tibenham Airfield, UK (52.457° N, 1.162° E), located 9.7 km from the Church Farm weather station. Agreement between the two stations was quantified using Pearson correlation coefficients, calculated after aggregating the 10-minute site measurements to hourly means and aligning them with the hourly observations from the public station.

Within the CIMMYT lines where sampling was conducted, informal disease assessments were carried out throughout the 2024 season by Darryl Playford, Field Experimentation Manager. In addition, for both 2023 and 2024, formal disease severity was assessed at a single time point at the end of the season. Severity was reported as a percentage and included assessments for mildew, yellow rust, brown rust, and septoria.

5.4.4 Bioinformatics

Raw read data were transferred from the P2 Solo or GridION to the EI HPC for downstream analysis. The reads were rebasecalled using Dorado (v0.7.2) with the super accuracy model (`dna_r10.4.1_e8.2_400bps_sup@v4.3.0`). If multiple sequencing runs had been conducted for the same samples, the data were combined based on shared barcodes. Pass reads were then randomly subsampled to 200,000 reads per sample using the script `subsample_reads.sh`, available in the GitHub repository `https://github.com/Mia-FGB/hpc-scripts`. If fewer than 200,000 reads were available for a given sample, all reads were retained. Subsampling was performed to reduce the computational time and memory required for taxonomic assignment. As demonstrated in Chapter 4, section 4.5.1.2, subsampling minimally alters the relative taxonomic composition of the samples.

Taxonomic assignment of the 200,000-read subsets with BLAST nt was performed using MARTi [281] (version listed in Table 5.2), with the following parameters: `LCAMaxHits:100, LCAScorePercent:90, LCAMinIdentity:70, LCAMinLength:150`, and `LCAMinReadLength:200`. The read count table was then exported from the front-end with a LCA cut-off of 0.1.

Custom Python and R scripts were developed to analyse and visualise the MARTi output data. All scripts described in this section are available on GitHub at `https://github.com/Mia-FGB/church_farm_scripts`. Taxonomic composition at the phylum and genus levels was visualised using stacked bar charts generated in R with the phyloseq package (`Stacked_Bar_Charts.R`). Analyses were filtered to include only fungal or oomycete genera within the phyla Basidiomycota, Ascomycota and Oomycota, and limited to weekly samples collected over the same seasonal period in each year. `Abundant_Phyla_Top_Species.R` was used to identify the top 10 genera for the three most abundant phyla in R.

Line graphs showing the temporal dynamics of selected plant pathogens were generated using `Pathogen_Graphs.R`, with data filtered to include the samples from 2023 and 2024. Additional analyses were performed in Python using Jupyter notebooks: pathogen disease scoring data were processed in `analyse_disease_scores.ipynb`, DNA yield trends were explored in `metadata_exploration.ipynb`, and environmental data were summarised and plotted using `Weather_plots.ipynb`.

5.5 Results

5.5.1 Sample information

The number of pass reads per sample ranged from 55,927-1,358,036 for the 2023 samples, and from 244,268-962,001 for the 2024 samples. Fifteen samples from 2023 had fewer than 200,000 pass reads and were retained in full without subsampling (Table 5.3). There was no clear seasonal effect on low read count, as these samples were distributed across all months. These difference are likely related to how well the sequencing libraries were balanced at the pooling stage.

The 2024 samples were collected with a sampling time four times as long as those in 2023 and subsequently had higher DNA concentrations per sample (Figure 5.4). DNA concentrations in 2024 ranged from 0.13 to 13.30 ng/μl, compared to 0.27 to 2.01 ng/μl in 2023, with eight of the 2023 samples initially yielding undetectable amounts. This difference, attributed to the shorter collection time in 2023, was mitigated with WGA, which increased DNA concentrations in the 2023 samples to a range of 0.60 to 144 ng/μl. While WGA successfully boosted yield, it also introduced greater variability in DNA concentration across the treated samples.

5.5.2 Phylum diversity

The abundance of phyla differs between 2023 and 2024 (Figure 5.5). In both years, Ascomycota are prevalent. In 2023, many samples are dominated by Pseudomonadota, whereas in 2024, there is a greater number of plant-derived reads observed from the phyla Streptophyta. The detection of a high abundance of Pseudomonadota is unexpected, and possible explanations for this observation are considered in the discussion section of this chapter.

Table 5.3: Samples with fewer than 200,000 pass reads retained without subsampling

Date Collected	Pass Reads
03/10/22	65,158
01/11/22	74,190
01/12/22	194,955
01/01/23	136,552
01/01/23	98,473
01/02/23	55,927
25/05/23	127,392
06/06/23	126,734
14/06/23	175,840
14/06/23	123,528
21/06/23	139,611
12/07/23	75,051
03/08/23	134,309
09/08/23	69,741
09/08/23	128,929

5.5.2.1 Phyla abundance 2023

The earlier samples from 2023 (May–June) are largely dominated by Pseudomonadota, while those collected from July to August show a shift towards Ascomycota dominance. Actinomycetota is consistently detected across most samples, with higher relative abundance earlier in the season. Streptophyta, Chordata and Basidiomycota appear in high abundance in select samples but not consistently across timepoints. Bacteroidota is present at low abundance, mainly during June and July. Arthropoda is detected in only one collection each year, but these occur at a similar time, on 20 July 2023 and 1 August 2024.

In 2023, several phyla are detected only at low abundance and in single samples, including Acidobacteriota (25 May, replicate 1), Cyanobacteria (6 June, replicate 1), Mollusca (14 June, replicate 2), Arthropoda (20 July, replicate 2), and Oomycota (9 August, replicate 1). Notably, Cressdnaviricota was observed at high abundance on 17 August (replicate 2) and at low abundance in other August samples. Chordata is detected at low abundance in several samples, with a notably high proportion on 25 May (replicate 2). There does not appear to be a seasonal pattern in Chordata abundance. While the 2023 dataset shows general consistency in phylum composition between replicates, several timepoints (25 May, 21 June, 7 July, and 20 July) do show differences in detected phyla between replicates, suggesting stochastic variation in airborne eDNA as these samples were collected consecutively.

5.5.2.2 Phyla abundance 2024

The 2024 samples show greater consistency over time and between replicates compared to 2023. The most distinct samples are from 14 May, which are dominated by Ascomycota, followed by Basidiomycota, with smaller proportions of Actinomycetota, Pseudomonadota, and Streptophyta. The 25 May samples are predominantly Streptophyta, the relative abundance of this phyla gradually decreases in subsequent weeks through to August.

All samples contain a substantial proportion of Ascomycota reads. From 23 May onwards, Ascomycota abundance generally increases over the monitoring period, before beginning to decline again in August. Actinomycetota, Basidiomycota and Pseudomonadota

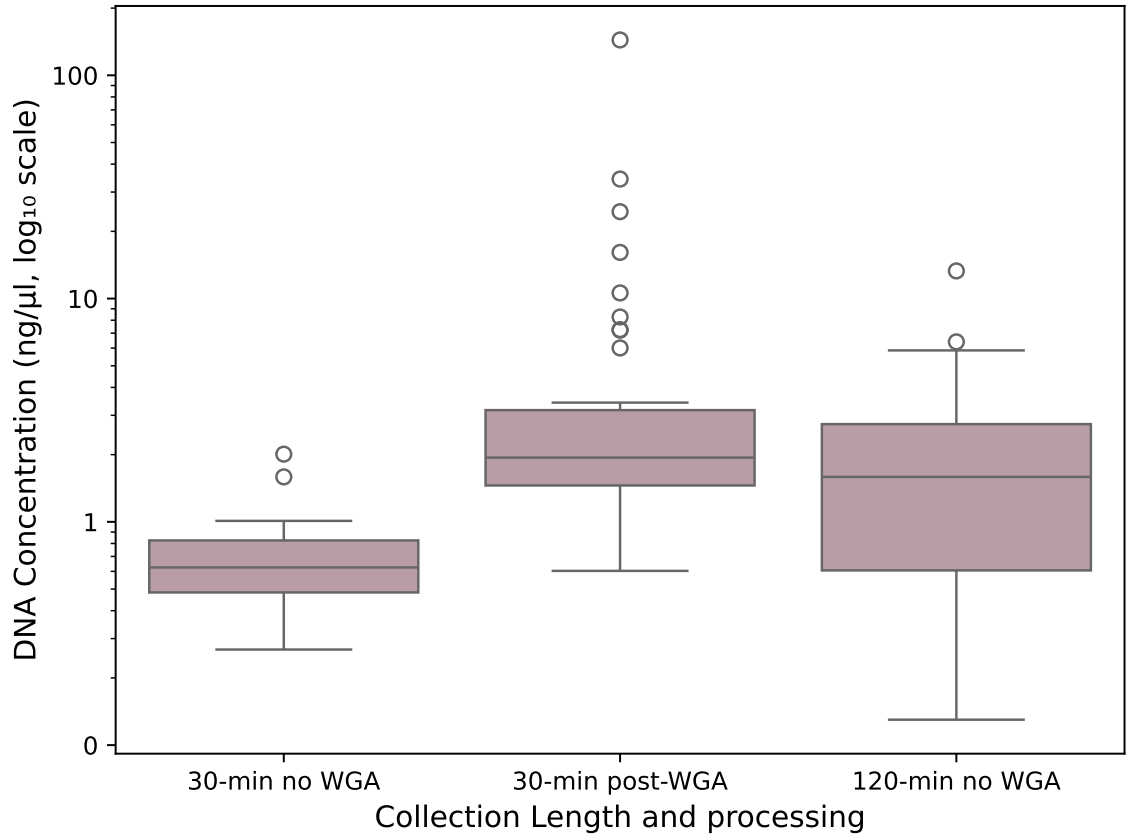


Figure 5.4: DNA concentration (ng/μl, $\log_{10}$ scale) for air samples collected over 30 minutes (2023) and 120 minutes (2024). WGA was applied only to the 2023 30-minute samples.

are consistently present at low abundance across the dataset. Oomycota are also present at low levels in many samples and are particularly abundant in those collected on 11 June and 7 August.

Bacillota and Bacteroidota are detected at low abundance in several samples collected between 18 June and August. Cyanobacteria is detected at low abundance in three samples: 26 June (replicate 2) and both 14 August samples. Arthropoda is detected only in the 1 August samples. Chordata is observed sporadically, generally at low abundance, but is notably higher in 11 June 2023 (replicate 1).

Overall, the relative abundance of phyla differs between the two monitored years (May–August), with a higher proportion of Pseudomonadota in 2023 and more Streptophyta in 2024. Ascomycota, Actinomycetota, and Basidiomycota are consistently detected in majority of the samples in both years. Seasonal shifts are observed in the dominant phyla, with some taxa increasing or decreasing in abundance over time, while others appear only sporadically in a few samples.

5.5.2.3 Top 10 genera from reads aligned to Ascomycota, Streptophyta and Pseudomonadota

The three most abundant phyla across the dataset were Ascomycota, Streptophyta and Pseudomonadota, the top ten genera from each were collated into Table 5.4.

Ascomycota top genera were dominated by well known plant and crop pathogens, including *Alternaria*, *Botrytis*, *Ramularia*, *Zymoseptoria*, and *Pyrenophora*, all of which cause foliar diseases in cereals and other crops. Opportunistic genera such as *Aspergillus* and *Penicillium* were also present but at lower relative abundances.

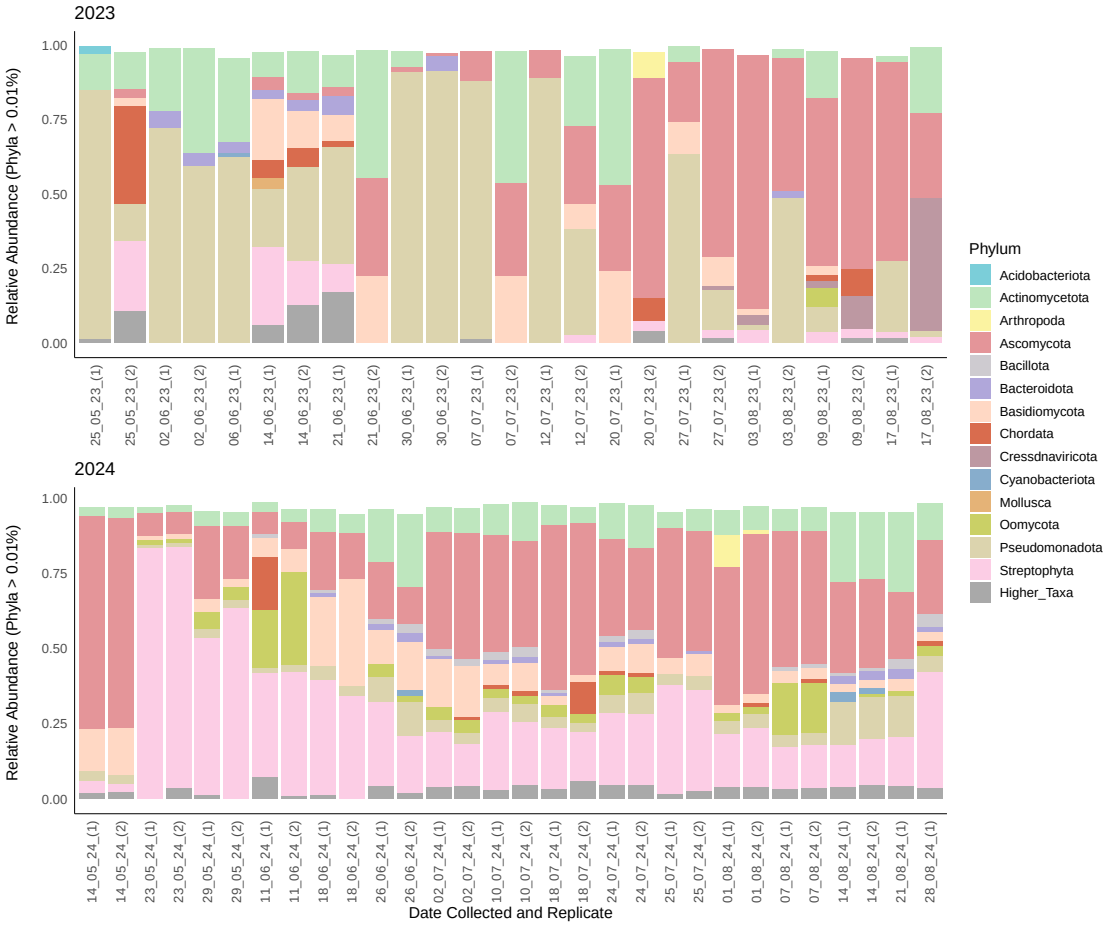


Figure 5.5: Relative abundance of phyla across timepoints in air microbiome samples collected during the seasonal monitoring experiments. Stacked bar plots show phyla representing more than 0.01% of total reads. The top plot presents data from 2023, and the bottom plot from 2024. Data are filtered to include only weekly replicate samples collected between May and August.

Table 5.4: Top 10 Genera from reads aligned to Ascomycota, Streptophyta and Pseudomonadota and their ecological relevance, ordered from most to least abundant.

Phylum	**Genus**	**Ecological Relevance**
Ascomycota	*Alternaria*	Major plant pathogen
	Parastagonospora	Plant pathogen (*P. nodorum* on wheat)
	Botrytis	Plant pathogen (*B. cinerea*, grey mould)
	Ramularia	Plant pathogen (leaf spots on cereals)
	Zymoseptoria	Plant pathogen (*Z. tritici* on wheat)
	Cladosporium	Mostly saprophytes
	Pyrenophora	Plant pathogen (net blotch, leaf spots on cereals)
	Aspergillus	Opportunistic human/animal pathogen
	Diaporthe	Plant pathogens (stem cankers, seed decay)
	Penicillium	Saprophytes; some opportunists
Streptophyta	*Pinus*	Tree (conifer)
	Quercus	Tree (oak)
	Urtica	Herbaceous weed (nettle)
	Alnus	Tree (alder)
	Triticum	Crop (wheat)
	Cryptomeria	Tree (conifer)
	Avena	Crop (oat)
	Brassica	Crop (cabbage, oilseed rape, mustard, etc.)
	Lolium	Grass
	Chamaecyparis	Tree (conifer)
Pseudomonadota	*Roseateles*	Environmental (soil/water bacteria)
	Sphingomonas	Environmental (leaf surface)
	Psychrobacter	Environmental (soil/water)
	Pseudomonas	Many species plant pathogens but also beneficial/rhizosphere species
	Paracoccus	Environmental
	Bradyrhizobium	Symbiotic (legume root nodules)
	Variovorax	Rhizosphere-associated, often beneficial
	Janthinobacterium	Environmental (soil/water)
	Ramlibacter	Environmental (soil)
	Aquincola	Environmental (freshwater)

Streptophyta abundant genera included major crops (*Triticum*, *Avena*, *Brassica*), alongside tree taxa such as *Pinus*, *Quercus*, and *Alnus*. These signals are consistent with both agricultural sources (crop pollen or debris) and background pollen from surrounding woody vegetation.

Pseudomonadota was represented mainly by environmental and soil-associated bacteria (*Roseateles*, *Sphingomonas*, *Psychrobacter*), together with symbiotic or rhizosphere taxa (*Bradyrhizobium*, *Variovorax*). Potential pathogens such as Pseudomonas were also present, though not dominant.

5.5.2.4 Detected Chordata genera

Among the reads aligning to the phylum Chordata, the genera represented by more than 1,000 reads across the dataset were human (*Homo*), pig (*Sus*), and junglefowl (*Gallus*).

5.5.3 Fungal and oomycete genera

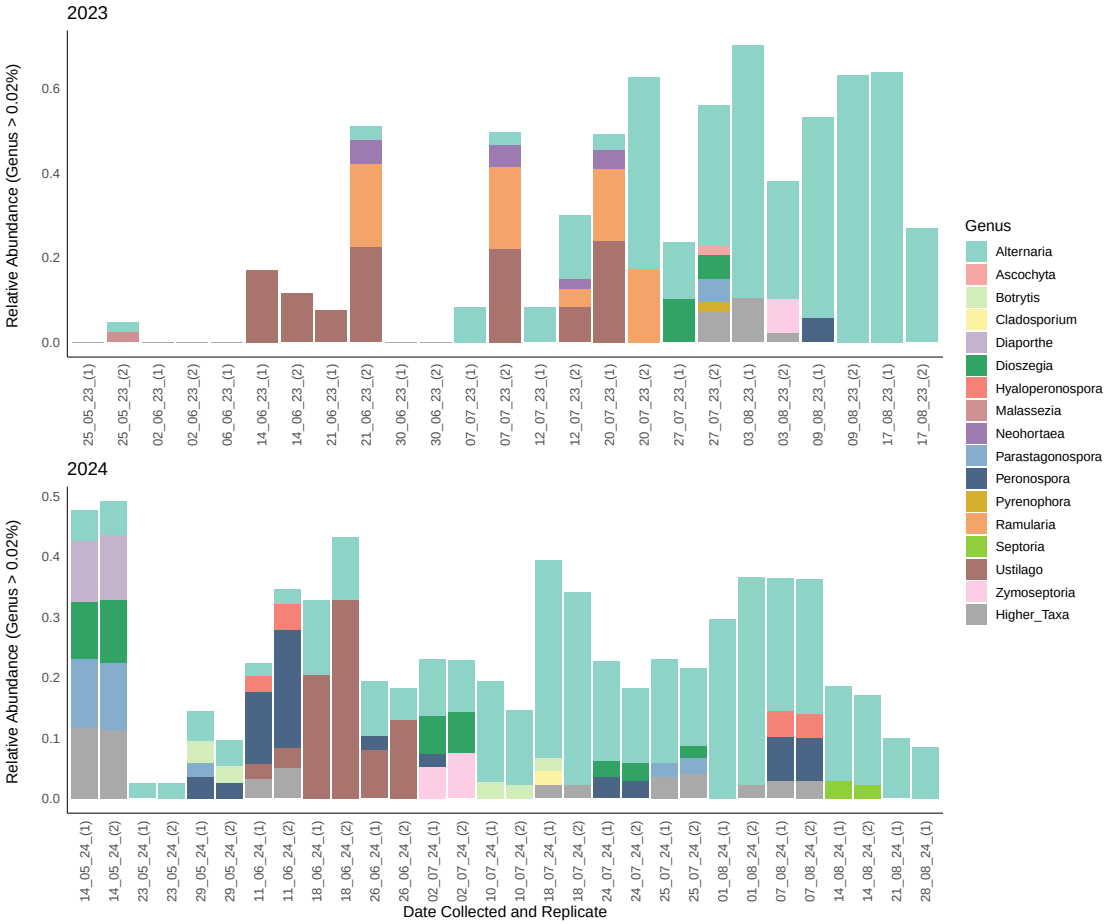


Figure 5.6: Relative abundance of fungal and oomycete genera across timepoints in air microbiome samples collected during the seasonal monitoring experiments. Stacked bar plots show fungal or oomycete genera representing more than 0.02% of total reads. The top plot presents data from 2023, and the bottom plot from 2024. Data are filtered to include only weekly replicate samples collected between May and August.

Across all samples, fungal and oomycete reads represented a relatively small proportion of the total community. At the phylum level, mean relative abundances were highest for *Ascomycota* (10.7% ± 0.8 SE), followed by *Basidiomycota* (2.7% ± 0.5 SE) and *Oomycota* (1.2% ± 0.4 SE) (Figure 5.6).

A closer examination at the genus level (Figure 5.6) showed *Alternaria* to be the most abundant in both years, with its relative abundance increasing toward the end of the season (July-August). *Ustilago* is also present in both years, consistently detected in June, and additionally in July in 2023. *Parastagonospora* is identified in late July in both years (27 July 2023, replicate 2; both 25 July 2024 replicates), and is also present in the initial May samples in 2024. *Dioszegia* is detected in July in both years; in 2023 it is only found on 27 July, whereas in 2024 it is present in multiple July samples.

In 2023, several May and June samples contain no detectable fungal or oomycete genera. The 25 May replicate 2 sample includes a low abundance of *Alternaria* and *Malassezia*. *Ramularia* and *Diaporthe* are detected sporadically in June and July, with *Ramularia* present at higher abundance than *Diaporthe*. The 27 July replicate 2 sample has the highest fungal diversity, containing *Pyrenophora*, *Parastagonospora*, *Dioszegia*, and *Ascochyta*. August 2023 samples are almost exclusively dominated by *Alternaria*, with the exception of *Zymoseptoria* on 3 August (replicate 2) and *Peronospora* on 9 August (replicate 1).

Different fungal and oomycete genera are observed in 2024. The earliest samples, collected on 14 May, have a distinct taxonomic profile and include *Diaporthe*, *Dioszegia*, and *Parastagonospora* alongside *Alternaria*. *Peronospora* is first detected at low abundance on 29 May, increases in abundance in the 11 June 2024, and reappears at low levels in later collections through to August. *Botrytis* is recorded at low abundance on 29 May and in July and August. *Hyaloperonospora* is detected in both replicates of one collection in June and August 2024. *Dioszegia* and *Zymoseptoria* are both present in samples collected on 2 July, the only date where *Zymoseptoria* is observed in 2024. *Dioszegia* is also detected in later July samples. *Cladosporium* appears in only one sample collected on 18 July (replicate 1) at low abundance. *Septoria* is detected exclusively in the 14 August samples.

There are clear yearly and seasonal differences in the detected fungal and oomycete genera. *Alternaria* is consistently present but increases in abundance towards July and August. *Ustilago* shows a consistent detection pattern across both years. Other taxa appear year-specific, such as *Ramularia* in 2023 and *Hyaloperonospora* in 2024. Some genera, including *Pyrenophora* and *Cladosporium*, are observed in only a single sample.

5.5.4 Fungal and oomycete genera of interest

Of the nine genera of interest (Table 5.1), only seven were detected in the regular Church Farm samples; *Magnaporthe* and *Claviceps* were not identified in the dataset (Figure 5.7).

Among the remaining seven genera, a higher number of reads aligning to these taxa (normalised to Hits per Million (HPM)) was detected in 2024 compared to 2023. There were no detections of *Blumeria* or *Puccinia* in 2023, and the other genera peaked in abundance in August, with *Ustilago* detected at lower abundance slightly earlier in the season, from June to July. In 2024, *Blumeria*, *Fusarium*, *Parastagonospora*, *Puccinia*, and *Pyrenophora* all exhibited peaks in abundance between late March and early April, although the exact timing varied among taxa. *Ustilago* and *Zymoseptoria* also showed sharp peaks, in June and July respectively.

Examining each genus in more detail, *Blumeria* was absent in 2023 but abundant throughout 2024 (mean: 463.44 HPM $\pm$ 76.32), with peak abundance observed on 19 March (2760 HPM). *Fusarium* displayed a similar pattern, detected in a single 2023 sample at 350 HPM (27 July). In contrast, in 2024, *Fusarium* was identified across the growing

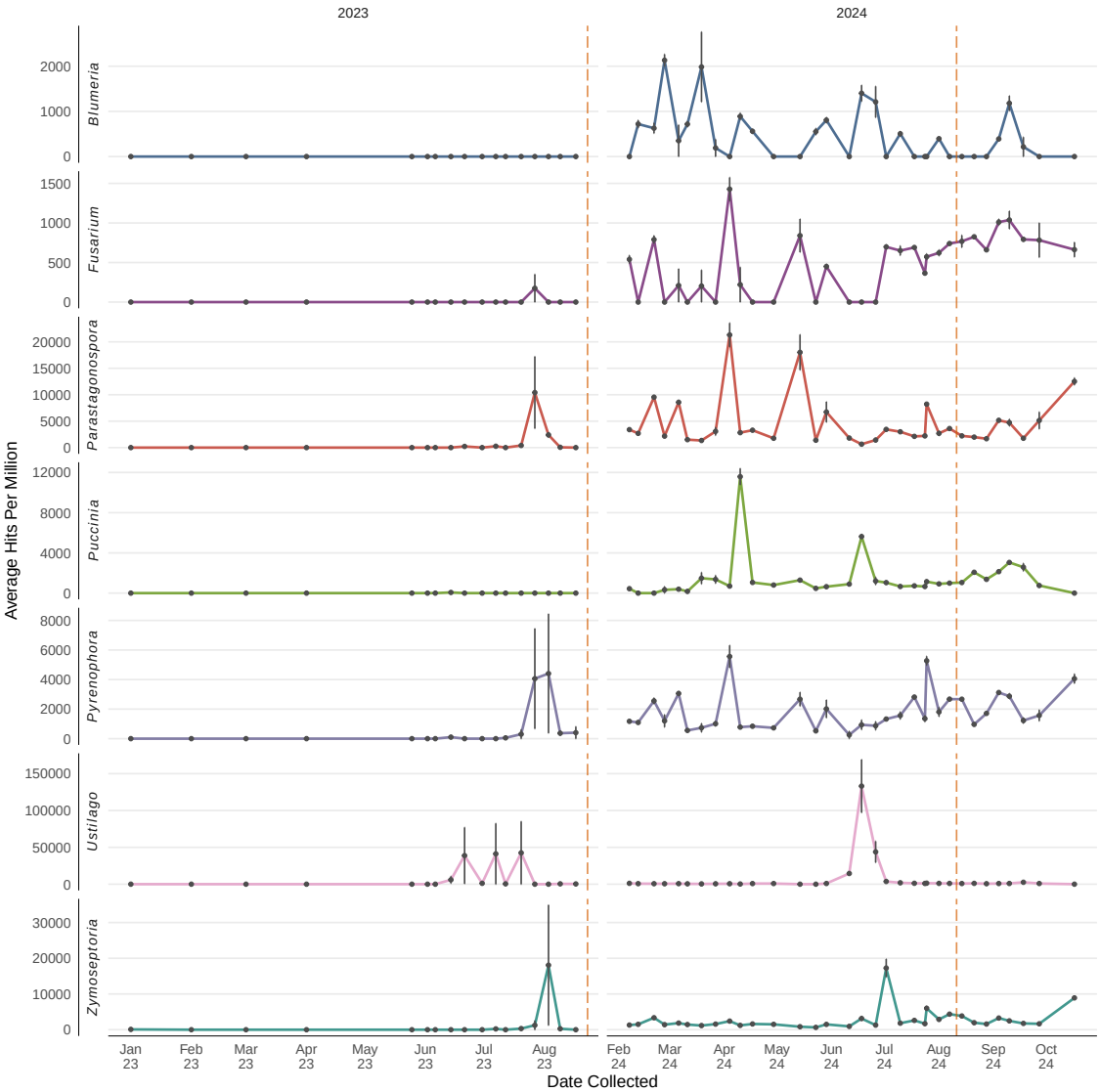


Figure 5.7: Temporal trends in airborne fungal pathogen detections at genus level, faceted by genus (rows) and collection year (columns). Line plots show the average number of hits per million (HPM) for each genus across sampling dates, with vertical error bars indicating standard error. Dashed orange vertical lines indicate the harvest date for each year.

season. Initial detections were variable, but from July to October there was a consistent presence of the taxon (overall mean: 462.89 HPM ± 50.46). The highest abundance of *Fusarium* occurred on 4 April (1575 HPM).

Similar to other genera, *Parastagonospora* was less abundant in 2023 (mean: 834.26 HPM ± 532.25) than in 2024 (mean: 4692.34 HPM ± 604.72). In 2023, a single peak was observed on 27 July (17,210 HPM). By contrast, in 2024, two distinct peaks were recorded on 4 April and 14 May, with abundances of 23,575 and 21,370 HPM, respectively.

Puccinia was detected at significantly higher levels in 2024 compared to 2023. In 2023, the mean abundance was low (4.14 HPM ± 4.14), with a peak of 136.49 HPM recorded on 14 June. In 2024, the mean abundance increased markedly to 1434.53 HPM ± 267.35, with the maximum abundance reaching 12,395 HPM on 10 April. A secondary, smaller peak was observed on 18 June. Following this, the abundance of *Puccinia* remained low for the remainder of the sampling period.

Pyrenophora was detected in both 2023 and 2024, with notably higher and more consistent abundance observed in 2024. In 2023, the mean abundance was relatively low at 587.75 HPM ± 333.10, with very low detection in most samples. However, two later-season samples showed elevated values: 7445 HPM on 27 July and a peak of 8435 HPM on 3 August. Notably, there is a large difference in relative abundance between the two replicates of these collections. In contrast, in 2024, *Pyrenophora* was present more consistently, with a higher mean abundance and lower variance between samples (1882.34 HPM ± 167.22). The maximum abundance occurred earlier in the season, peaking at 6315 HPM on 4 April.

Ustilago was detected at low to moderate abundance in both 2023 and 2024, with notable variation in temporal patterns between the two years. In 2023, several minor peaks were observed from June to July, although overall abundance remained low. The highest value for the year was recorded on 20 July (85,040 HPM), contributing to an overall mean abundance of 7930.34 HPM ± 4127.14. This pattern of detection differed in 2024 with a single pronounced peak occurred on 18 June, reaching 169,165 HPM. This peak was flanked by low-abundance detections in the preceding and following weeks, suggesting a sharp and isolated release event. The mean abundance in 2024 was slightly lower than in 2023 (6830.94 HPM ± 3143.02), differing from all of the other genera considered.

Zymoseptoria was detected in both years, exhibiting different temporal dynamics. In 2023, it was identified only in a single sampling event on 3 August, with a peak abundance of 34,890 HPM and considerable variation between the two replicate samples collected on that day. Meanwhile, in 2024, *Zymoseptoria* was detected intermittently throughout the season at low levels for most of the year, with a distinct peak observed on 2 July (19,780 HPM). The overall mean abundance in 2024 was 2778.28 HPM ± 391.93.

In summary, the results demonstrate clear inter-annual and temporal variation in the presence and abundance of the fungal genera analysed. While some taxa were consistently detected across the sampling period, many exhibited sporadic peaks in abundance, often confined to individual sampling events. Outside of these peaks, most genera were either absent or present at low levels, highlighting the transient and variable nature of airborne fungal DNA in the study area.

5.5.5 Disease presence

5.5.5.1 Timing of disease arrival in 2024

Informal disease assessments were conducted at Church Farm throughout the 2024 season by field staff. Early signs of yellow rust, mildew, and Septoria were noted as early as 23 February, although at low levels. By 6 March, Septoria remained generally distributed across plots, while yellow rust appeared as isolated stripes in a few plots. On 13 March, Septoria had progressed significantly, with some lower leaves completely covered, while yellow rust had not yet increased, which was presumed to be due to low temperatures.

Examples of symptoms observed in the field, including Septoria caused by *Zymoseptoria tritici* and yellow rust caused by *Puccinia striiformis*, are shown in Figure 5.8. The figure presents images of wheat leaves collected at the sampling site in 2024.

Additionally, ergot (*Claviceps purpurea*) was observed on 23 July in a field near the untreated plots, location marked on Figure 5.3.

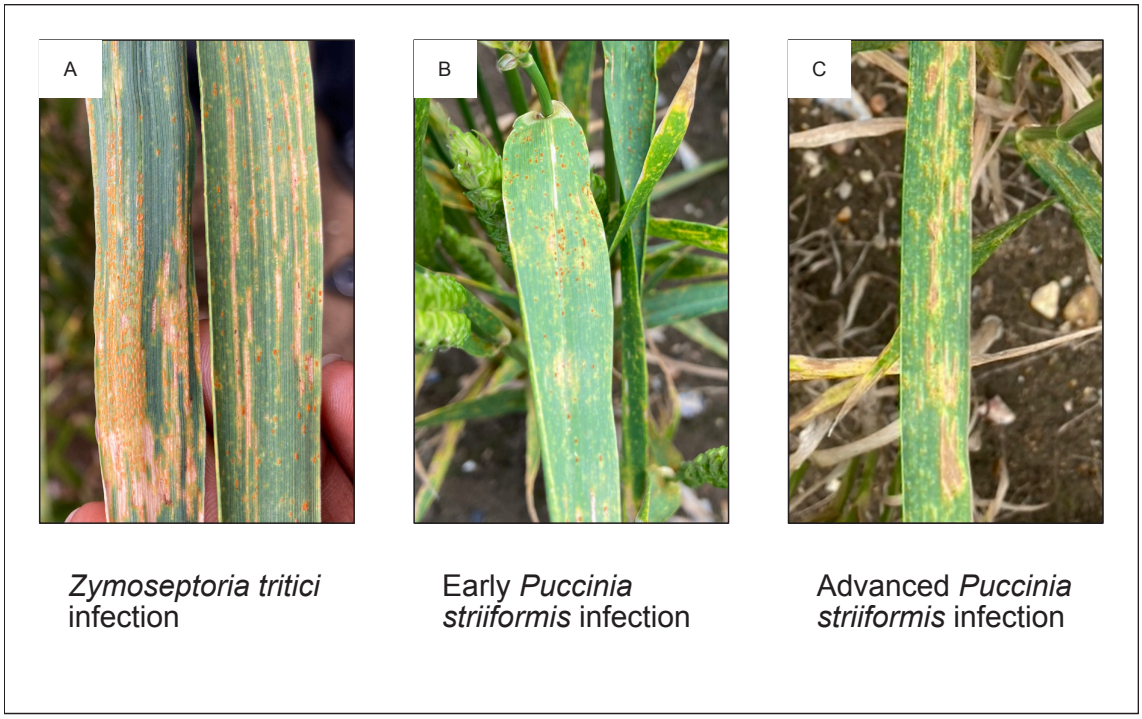


Figure 5.8: Field symptoms of key wheat pathogens identified in airborne samples: (A) typical Septoria tritici blotch lesions with chlorosis and necrosis; (B) early-stage yellow rust pustules aligned with veins; and (C) severe yellow rust infection with dense pustule stripes. Pictures taken on 17 June 2024 in the sampling location.

5.5.5.2 Disease score data

The CIMMYT plots where sampling was conducted were also assessed for disease at the end of the season, with the resulting scores shown in Figure 5.9.

Brown rust, caused by *Puccinia triticina*, and mildew, caused by *Blumeria graminis* f.sp. *tritici*, were almost absent in both years. In 2023, the maximum recorded severity was 0.01% for brown rust and 0.1% for mildew. In 2024, brown rust increased slightly with a maximum score of 10%, while mildew was not detected.

Septoria, caused by *Zymoseptoria tritici*, was present at relatively low levels in 2023 (range: 0–15%; mean: 1.43 ± 3.14). In 2024, the disease was more prevalent, although

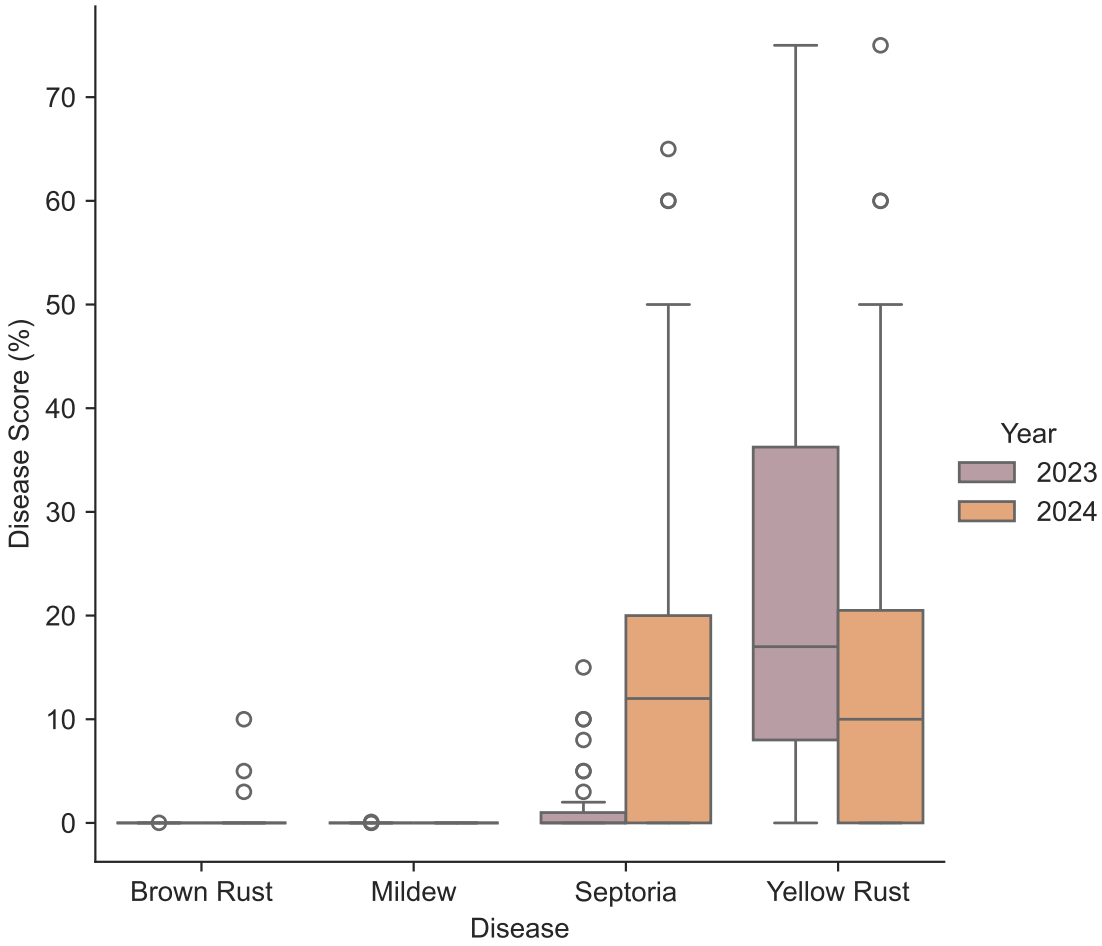


Figure 5.9: Boxplot showing disease severity scores (%) for four wheat diseases (brown rust, mildew, septoria, and yellow rust) assessed on a single end-of-season date in 2023 and 2024 in the CIMMYT plots at Church Farm. Each point represents the disease score of a plot in the field. Data are grouped by disease and year.

severity varied widely across plots (range: 0–65%; mean: 14.35 ± 13.86). This variation may reflect differences in varietal susceptibility or the more conducive environmental conditions in 2024.

Yellow rust, caused by *Puccinia striiformis* f.sp. *tritici*, was prevalent in both years. Average disease severity was higher in 2023 (range: 0–75%; mean: 24.67 ± 22.58) than in 2024 (range: 0–75%; mean: 13.97 ± 16.63), although the maximum recorded score was the same. As with Septoria, the wide variation likely reflects differences in varietal susceptibility and environmental conditions.

5.5.6 Weather data

The environmental conditions during the 2023 and 2024 sampling periods are summarised in Figure 5.10 and Table 5.5. Overall, conditions were broadly similar across both years. Daily temperature data show slightly warmer average values during spring 2024, while the coldest minimum temperatures were recorded in January 2023.

Following quality control, seven wind speed observations exceeding 100 ms^{-1} were removed from the 2023 dataset as likely sensor errors. Removal of these values increased the correlation between the Church Farm weather station and the nearby Tibenham Airfield station from 0.30 to 0.81 (see Appendix, Figure A.1), supporting the conclusion that the remaining wind speed observations were plausible.

Electrical conductivity followed a comparable seasonal trend in both years, decreasing through early summer (June–July) and increasing again from August to October. Average, minimum, and maximum values for this variable showed negligible differences between years.

Average daily rainfall totals were similar across years; however, a notably high single-day rainfall event was recorded in 2023 (11.63 mm on 10 May). In contrast, rainfall was more frequent and intense in early 2024, particularly during March and April.

Mean barometric pressure, relative humidity, and wind speed also remained broadly consistent between the two sampling years.

Across both years, only a limited number of months contained days in which all environmental conditions conducive to infection were met: June to August in 2023, and July to September in 2024 (Figure 5.11). Throughout the year, hourly relative humidity consistently exceeded 75% on at least one occasion per day, and rainfall totals greater than 0 mm were recorded across all months. However, elevated temperatures (>20°C) were largely restricted to the summer months and typically occurred on fewer than 20 days per month. It should be noted that crops were harvested by August, so conditions observed later in the year would not have contributed to infection risk of the plants.

5.6 Discussion

This chapter presents two years of longitudinal data collected from an untreated wheat field, including observations on disease presence and severity, alongside environmental conditions. The following discussion explores key trends, patterns and relationships observed in the dataset. The analysis begins with a comparison of the samples between years then considers taxonomic assignment at the phylum level, before identifying the most abundant fungal and oomycete genera, and finally focusing on the nine target genera.

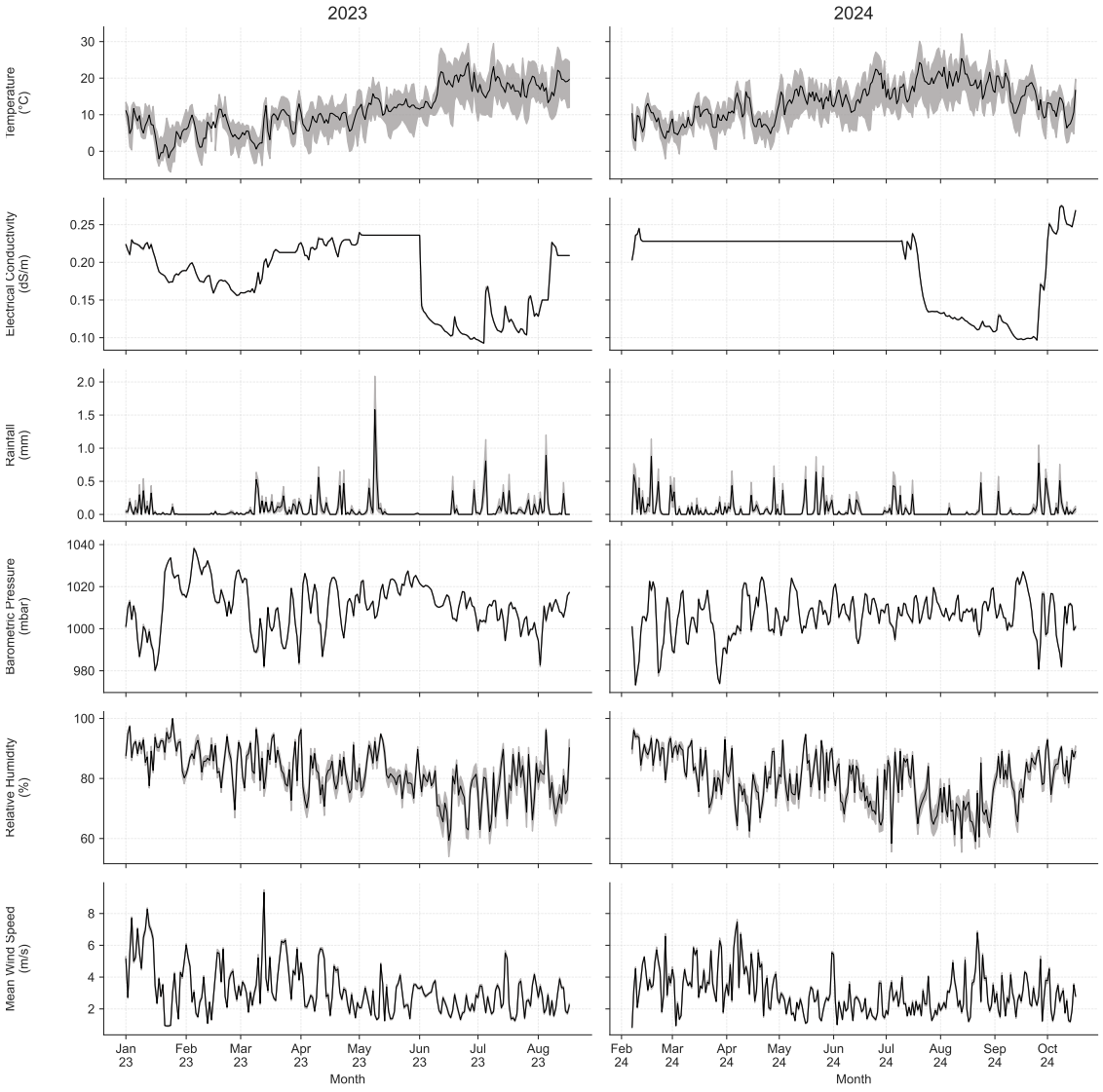


Figure 5.10: Environmental conditions at Church Farm during the 2023 (left) and 2024 (right) sampling periods. Rows show: daily mean temperature with minimum–maximum range (daily data); and daily summaries of electrical conductivity, rainfall, barometric pressure, relative humidity (hourly data), and wind speed (10-minute data). Shaded areas represent the daily range (for temperature) or standard error (for all other variables).

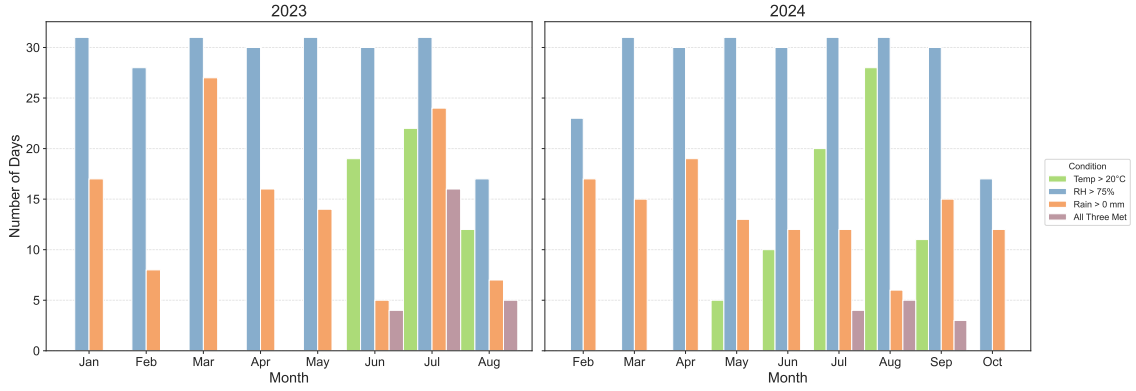


Figure 5.11: Number of days per month in 2023 and 2024 meeting key environmental conditions relevant to airborne fungal pathogen development. Conditions include: days where the average hourly temperature exceeded 20 °C at least once; relative humidity exceeded 75% in any hour; daily rainfall totalled more than 0 mm; and days where all three conditions occurred simultaneously. Bars are grouped by year to allow visual comparison across months.

Table 5.5: Environmental conditions (mean ± SE) over different periods of the year.

Variable	**Frequency**	**Year**	**Early season (Feb–Apr)**	**Mid season (May–Jul)**	**Late season (Aug–Oct)**
Average Air Temperature (°C)	Hourly	2023	7.00 ± 0.08	14.88 ± 0.10	17.02 ± 0.20
		2024	8.57 ± 0.07	14.83 ± 0.09	15.37 ± 0.11
Barometric Pressure (mbar)	Hourly	2023	1011.44 ± 0.29	1012.19 ± 0.17	1006.99 ± 0.43
		2024	1003.53 ± 0.28	1008.25 ± 0.14	1007.40 ± 0.21
Electrical Conductivity (dS/m)	Hourly	2023	0.20 ± 0.00	0.16 ± 0.00	0.19 ± 0.00
		2024	0.23 ± 0.00	0.22 ± 0.00	0.15 ± 0.00
Relative Humidity (%)	Hourly	2023	83.89 ± 0.26	78.17 ± 0.33	78.13 ± 0.80
		2024	82.98 ± 0.24	78.23 ± 0.31	77.66 ± 0.33
Rainfall (mm)	Hourly	2023	0.07 ± 0.01	0.08 ± 0.01	0.10 ± 0.02
		2024	0.10 ± 0.01	0.07 ± 0.01	0.06 ± 0.01
Mean Wind Speed (m/s)	10 minute	2023	3.60 ± 0.02	2.57 ± 0.01	2.53 ± 0.03
		2024	3.88 ± 0.02	2.30 ± 0.01	2.90 ± 0.02

5.6.1 Sample duration

The duration of sample collection differed between years, with samples in 2024 collected over a period four times longer than those in 2023. These longer sampling periods generally resulted in higher DNA yields, allowing for immediate downstream processing (Figure 5.4). In contrast, the 2023 samples required WGA to achieve yields comparable to those obtained in 2024, which may have introduced bias into the resulting taxonomic community [288, 318].

Longer sampling windows are also likely to capture greater diversity, as different fungal species have different environmental or physiological triggers for spore release and therefore release spores at different times of day [206, 352]. These biases in amplification or diversity captured do not directly affect the analysis presented here, which focuses on pathogen presence in relation to environmental conditions. They should, however, be kept in mind when comparing community composition between the two years.

These findings highlight the importance of considering sample duration when planning field experiments, particularly for time-sensitive monitoring. While shorter samples may at first glance appear to lead to quicker results, longer collections often produce higher initial DNA yields, removing the lengthy WGA step enabling faster processing. Additionally, amplified DNA can interfere with pore function and reduce sequencing output [261]. Therefore, longer collections may ultimately result in both faster and more accurate outcomes.

5.6.2 Relative abundance of phyla

The observed differences in phylum composition from the weekly collections between 2023 and 2024 likely reflect a combination of environmental factors and methodological variation (Figure 5.5).

Greater variability between replicates in 2023 may be attributed to their collection in succession rather than simultaneously. Given the highly dynamic nature of airborne eDNA, it is possible that the microbial community shifted between sampling events. Furthermore, the lower DNA yields in 2023 necessitated WGA, which may have introduced additional bias and amplified replicate-level differences. In contrast, the 2024 samples were collected over a longer duration, reducing the likelihood of dominance by stochastically abundant taxa and instead capturing a broader representation of the airborne community present during the collection period, as demonstrated in Chapter 4.

Due to the broad diversity within phyla, there are no clearly defined environmental conditions under which taxa are known to be present in the environment. But there are some seasonal differences observed in this dataset.

The clearest difference between the two years is the high abundance of Pseudomonadota in 2023 compared to 2024, which contained very little Pseudomonadota but a high abundance of Streptophyta. Pseudomonadota is a phylum of both pathogenic and benign Gram-negative bacteria, commonly found in environmental samples including air, soil, and water. It has been frequently detected in airborne microbiome studies [38, 220, 233, 334]. In this dataset the most abundant Pseudomonadota genera were environmental bacteria (Table 5.4). Another study showed that Pseudomonadota increased in airborne abundance in the summer season compared to the rest of the year [52], which may suggest the high detection in 2023 but then does not explain the absence in 2024.

Streptophyta includes land plants and some green algae, in this dataset the most abun-

dant plant genera identified largely reflect those expected in the study region, including tree genera such as *Quercus*, *Alnus*, and *Populus*, and crop species including *Triticum* and *Avena* (Table 5.4). Since these are airborne detections they are likely to be pollen. Pollen release is species-dependent, but is generally correlated with temperature and occurs from spring through to late summer. Other studies have correlated pollen release with RH, though reported relationships vary: one study found a negative correlation [71], while another reported a positive one [313].

Temperature and relative humidity during the 2023 and 2024 sampling periods were broadly similar, with 2024 showing slightly elevated temperatures and lower relative humidity (Figure 5.10). This increase in temperature may have contributed to enhanced pollen release, resulting in higher Streptophyta abundance in 2024. However, as environmental differences were minor, perhaps the longer sampling duration in 2024 resulted in higher pollen detection as the collections more effectively overlapped with peak release. Another potential factor was the change in sampling location which was moved to a different field within the same farm in 2024 (Figure 5.3). Differences in surrounding crops or hedgerows may have influenced the airborne community detected.

The contrasting dominance of Pseudomonadota in 2023 and Streptophyta in 2024 is unlikely to be explained by the slight differences in environmental conditions. Instead, the unexpectedly high abundance of Pseudomonadota in 2023 may reflect differences in how samples were collected and processed, with shorter collections resulting in lower DNA yields and necessitating WGA prior to sequencing. The large differences between datasets across years raise questions about the representativeness of the 2023 data, emphasising the importance of longer sampling durations to reduce the risk of biased detection from transient or sporadic signals, and of avoiding reliance on WGA, which may further distort taxonomic profiles. Together, these findings underscore how both methodological and environmental factors shape community composition in airborne metagenomic surveys.

In both years, Ascomycota were present in many samples, with a general increase in abundance from July onwards. An exception is seen in the 14 May 2024 samples, where both replicates show unusually high Ascomycota abundance. Ascomycota, commonly referred to as sac fungi, comprise a morphologically diverse phylum whose spores are known as ascospores. The top Ascomycota genera identified in this study were predominately plant pathogens (Table 5.4), aligning with the agricultural location where pathogens are likely to be present. Previous studies have shown that airborne ascospore concentrations vary seasonally and geographically. For example, a study in Spain reported peak concentrations in autumn, particularly in September [152], while research in Greece identified highest levels from spring to mid-summer [128]. This suggests that regional environmental conditions play a significant role in determining the timing of peak ascospore abundance. In the UK, Symon et al. documented multiple seasonal peaks in different Ascomycota taxa from June through to October [357], indicating that the Ascomycota detected in this study likely include a range of taxa with distinct sporulation periods across the growing season.

Actinomycetota were more abundant in 2023 than in 2024, with particularly high relative abundance in May 2023 and increasing abundance through August in 2024. This phylum includes high-GC-content Gram-positive bacteria that are commonly found in soil and water, and have been widely reported in airborne microbial studies [138, 229, 331]. The presence of Actinomycetes may be attributable to the large open spaces at the farm, where dry soil is lifted into the air and dispersed by the wind.

One study observed increased abundance of non-spore-forming Actinomycetota in airborne samples following rainfall [172]. In 2023, rainfall was particularly high in early May, with lower but steady levels from July to August, while in 2024, rainfall was lower overall but relatively consistent from May to July and then again from September to October (Figure 5.10). These patterns may partly explain the temporal variation in Actinomycetota abundance observed across the two years. However, the relationship is not fully consistent, suggesting additional factors are likely to influence abundance. Moreover, because samples were collected weekly, the dataset may lack sufficient temporal resolution to detect short-term responses to rainfall events.

Additionally, sporadic detections were made from non-microbial phyla such as Arthropoda, Chordata, and Mollusca. The Chordata reads likely reflect human contamination and the presence of local vertebrates (pigs and chickens), both of which have previously been detected with airborne eDNA [80, 228, 389]. The detection of pig DNA is of particular note, as pigs were kept in a nearby field during sampling in 2023. These sporadic detections therefore likely occurred when the wind was carrying material downwind from that field.

Arthropoda was detected in only a few samples (July 2023 and August 2024). As noted in Chapter 4, accidental capture of insects in the air filters has occurred previously, and this likely explains their presence here. Mollusca was detected at very low abundance in a single sample (14 June 2023). This may represent a true positive (i.e. airborne Mollusca DNA), but is more likely a false positive resulting from misclassification, as molluscs are not typically associated with airborne dispersal.

Overall, the observed differences in phyla between years may reflect environmental differences or be artefacts from the different sampling methods applied.

5.6.3 Relative abundance of abundant fungal and oomycete genera

At the genus level, there are both similarities and differences in the detected taxa between the two years (Figure 5.6). As in earlier observations, these differences are likely driven by a combination of environmental conditions and sampling methods.

Alternaria has been detected at high abundance in several airborne microbial studies [18, 345]. In the UK, peak *Alternaria* spore concentrations have been recorded in early August [345], aligning with the elevated levels observed in this dataset. Supporting evidence from Portugal also demonstrates a similar seasonal trend, with low concentrations in early spring and a marked increase through the year with maximum values recorded in August and September [271]. These findings suggest that *Alternaria* exhibits a broadly consistent seasonal pattern across temperate European climates, peaking in late summer. In cereal systems, *Alternaria* species are also among the "sooty moulds" that colonise maturing wheat spikes [292]. This colonisation of wheat may contribute to the elevated airborne levels detected.

Some taxa appeared at notable abundance in only one year, suggesting interannual variation in the airborne fungal and oomycete community. These include *Ramularia* in 2023, and *Hyaloperonospora* and *Peronospora* in 2024. *Ramularia collo-cygni*, is the causal agent of Ramularia leaf spot in barley, and does not ordinarily infect wheat [146, 245]. However, in 2023, the sampling location was adjacent to a barley plot, which may explain the detection of *Ramularia*. This highlights how local vegetation and cropping context can

influence airborne fungal composition. Previous work combining airborne sampling of *R. collo-cygni* with PCR identified peak spore release in July in the UK [145]. This aligns with the 2023 data presented here, in which *Ramularia* was detected at the end of June and in July (Figure 5.6).

Hyaloperonospora and *Peronospora* are closely related oomycetes within the *Peronosporaceae* family. Both are obligate plant pathogens that cause downy mildew diseases. A study of airborne oomycetes in Germany found these genera to be consistently present, with *Peronospora* showing peak abundance in early spring and summer, and a marked decline in autumn and winter. This seasonal pattern was attributed to the availability of suitable host plants and favourable environmental conditions during the growing season [207]. These findings align with the observations in this study, where *Hyaloperonospora* and *Peronospora* were detected intermittently between May and August, though only in 2024 and not consistently throughout the season. The sporadic detection of oomycete taxa may reflect short-lived environmental conditions or microclimatic changes that influence spore release during individual sampling periods.

Several genera, including *Ascochyta*, *Botrytis*, *Cladosporium*, and *Septoria* were detected in only one or two samples and at low abundance. The sporadic detection of these taxa may reflect short-lived sporulation events, microclimatic conditions that favour spore release, or limitations in the weekly sampling resolution. Many fungal pathogens are known to have narrow temporal windows for spore dispersal, often linked to host development or specific environmental triggers [162].

The 14 May 2024 samples appear taxonomically distinct from the other timepoints, showing notably different community composition (*Diaporthe*, *Dioszegia* and *Parastagonospora*). However, this divergence does not clearly correspond to any measured environmental variable; for example, wind speed during the sampling window was close to the seasonal average. The cause of this distinct profile remains unclear but may reflect transient environmental or biological events not captured in the available metadata, such as unrecorded local agricultural activity, plant growth stage, or microclimatic factors. As *Parastagonospora* is one of the target genera, it will be examined in more detail below, with abundance considered across time without the same level of filtering.

One anomalous detection is that of *Malassezia* in three samples. This genus is typically associated with the skin of humans and other animals [119], and is not commonly found in environmental air samples. Its presence in this dataset could be from airborne animal skin or fur particles but could also be the result of human contamination during sampling or sample processing.

5.6.4 Genera of interest

The nine fungal and oomycete genera containing wheat pathogens (Table 5.1) varied in relative abundance over the monitoring period and between sampling years (Figure 5.7). Overall, pathogen abundance was lower in 2023, with fewer detection events than in 2024. Some genera were not detected in either year, including *Claviceps* and *Magnaporthe*. *Blumeria* was only detected in 2024, with highest abundance early in the season. Other genera, such as *Fusarium*, *Parastagonospora*, *Puccinia*, and *Pyrenophora*, were detected sporadically in 2023 but appeared consistently throughout 2024, with several peaks observed across different months. *Ustilago* and *Zymoseptoria* were detected in both years, each showing a

single peak later in the growing season.

The increased detection of pathogenic genera in 2024 is most likely due to the longer sampling durations. Weather conditions were broadly similar between years (Table 5.5), with 2023 having more days where multiple environmental conditions conducive to fungal spore germination were met (Figure 5.11). Additionally, data were subsampled to 200,000 reads per sample and normalised to minimise sequencing bias. Therefore, the extended sampling window in 2024, from 30-minutes to 2-hours, likely improved the detection of rarer or more sporadically occurring taxa, particularly those released later in the day, resulting in higher relative abundances of the pathogens of interest.

Beyond environmental factors and sampling duration, the observed abundance of airborne pathogenic material can also be compared to crop disease scores. However, there was not consistent agreement between the two (Figure 5.9), with little to no symptoms of brown rust (caused by *Puccinia triticina*) or mildew (caused by *Blumeria graminis* f.sp. *tritici*) in either year, despite airborne detections of both genera in 2024 (Figure 5.7). For brown rust this may be explained by the disease's tendency to develop late on the leaves, meaning symptoms could have been missed during routine scoring. Whereas for mildew this discrepancy may be due to mildew establishing early on seedlings before persisting deep within the crop canopy, meaning symptoms might not have been readily apparent during scoring at the end of the season.

Septoria (caused by *Zymoseptoria triticii*) was more widespread in 2024 than in 2023, yet the airborne abundance of *Zymoseptoria* appears roughly equal in both years. The final disease assessed was yellow rust, caused by *Puccinia striiformis* f.sp. *tritici*, with slightly higher disease scores in 2023 than in 2024, while airborne *Puccinia* was only detected in 2024.

These discrepancies between airborne inoculum and the disease scoring likely reflect that scoring was conducted only once at the end of the season, using material specifically selected for susceptibility testing. The CIMMYT plots themselves contained a wide variety of wheat material, with 27 and 60 different crosses in 2023 and 2024 respectively, only six of which were the same across both years. The resistance characteristics of this material are likely to have had a major effect on disease development, as each line carried different combinations of susceptibility and resistance alleles. Consequently, observed disease differences between years may reflect variation in host susceptibility across the CIMMYT plots, rather than differences in airborne inoculum levels alone.

This highlights that airborne pathogen presence does not always correspond with disease development, as other factors, such as host susceptibility, sporulation timing, and environmental conditions, also play a role [61, 108, 162]. A threshold level of airborne inoculum may also be required to initiate disease or trigger an epidemic, meaning the pathogen can be detected even if symptoms are absent. Additionally, pathogens may originate from nearby fields not included in the disease scoring. Further research is needed to clarify how airborne pathogen data relate to in-field plant health and how this information can support crop management decisions.

5.6.4.1 Genera that were not identified

Neither *Claviceps* or *Magnaporthe* reads were identified in the dataset. The absence of *Claviceps* is intriguing, as although ergot disease is infrequently observed in the UK [99],

sclerotia of ergot (*Claviceps purpurea*) were identified in wheat crops growing in the field adjacent to the sampling site on 23 July 2024, the location is shown in Figure 5.3.

Claviceps survives the winter in the soil as sclerotia, which develop spore-producing structures in spring or early summer. These structures release ascospores that are spread by wind and can infect cereal crops during flowering. Shortly after the initial infection, conidia are produced and transferred to other hosts via insect vectors or rain splash. Sclerotia formation occurs later in the season, as the final stage of the infection cycle [302], at this point spores are no longer released. Therefore, the presence of sclerotia in the nearby field suggests that *Claviceps* spores would no longer be airborne in July.

However, one might still expect to detect airborne *Claviceps* earlier in the season during the initial infection period. The absence of detection may be attributed to the timing of spore release falling outside the sampling window. *C. purpurea* ascospores have been reported to be released predominantly between 00:00 and 06:00 [8], a period not covered by the sampling protocol. Additional contributing factors may include low spore abundance or highly localised dispersal, both of which could reduce the likelihood of detection at the sampling site.

As only the primary ascospores are wind-dispersed, and subsequent conidia rely on insect and rain-splash dispersal, early-season airborne concentrations may be relatively low and therefore more difficult to detect. Research into the dispersal distance of *Claviceps* is limited, although one study suggests ascospores can travel at least 60 m [96]. This is unlikely to be sufficient for detection at the 2024 sampling location, which was situated more than 150 m from the field edge where ergot sclerotia were observed.

Unlike *Claviceps*, the absence of *Magnaporthe* detection aligns with current knowledge of the pathogen's distribution. Wheat blast, caused by *Magnaporthe oryzae* pathotype Triticum, originated in Brazil and has since spread across South Asia and parts of Southern Africa [343]. Due to the rapid transboundary spread of this disease, continued vigilance and early detection systems remain critical [169]. While wheat blast does not currently threaten UK wheat production, the use of AirSeq to monitor for its potential introduction will be an important tool for safeguarding future crops.

5.6.4.2 Early season detection: *Blumeria*

Blumeria graminis, the causal agent of wheat powdery mildew, was only detected in airborne samples in 2024 and was not observed on plants during disease scoring (Figure 5.7 and 5.9). The absence of detection in 2023 may be attributed to the shorter sampling duration, as discussed above, or to lower inoculum pressure at the sampling location that year.

In 2024, airborne *Blumeria* was first detected in February, with peak abundance recorded in March and April, followed by sporadic detections throughout the rest of the season. The early-season presence likely reflects conidia originating from overwintered mycelium or residual inoculum from the previous season, as *B. graminis* is known to survive as dormant mycelium on aerial plant tissues [173]. Additionally, barley seedlings are more susceptible to *Blumeria* infection [366], which may explain the higher spore abundance early in the season. As plants mature, resistance increases, leading to reduced disease and lower airborne inoculum, consistent with the pattern observed here.

These early airborne detections align with disease observations, which reported mildew

(caused by *Blumeria graminis* f.sp. *tritici*) on wheat plants in February at low levels. However, the disease did not appear to progress, and by the end of the season, mildew was no longer observed during disease assessments (Figure 5.9), despite continued low-level detection of *Blumeria* in the air.

Optimal conditions for powdery mildew development include temperatures between 5 and 30 °C and relative humidity above 80 % [235]. These conditions were met during several periods of the 2024 season (Figure 5.10, Figure 5.11), which may explain the continued but sporadic detection of airborne *Blumeria* beyond the spring peak. The apparent discrepancy between airborne abundance and disease severity has been discussed earlier and likely reflects the complex interplay between inoculum presence, environmental conditions, host susceptibility, and timing of infection.

A similar pattern of airborne *Blumeria* abundance was observed in a microscopy-based study conducted across the UK, where *Blumeria* was consistently detected throughout the sampling season. Peak levels were recorded in May, followed by a gradual decline towards the end of sampling in August [287].

5.6.4.3 Mid-season persistent pathogens: *Fusarium, Parastagonospora, Puccinia* and *Pyrenophora*

Fusarium, *Parastagonospora*, *Puccinia*, and *Pyrenophora* genera displayed similar patterns of airborne abundance. In 2023, all except *Puccinia* were detected toward the end of the sampling period (July–August) whereas in 2024, the genera showed a sustained presence throughout the season with peaks in relative abundance in April, along with additional secondary peaks later in the year (Figure 5.7).

The identification of *Parastagonospora* is noteworthy, since *Parastagonospora nodorum* (causing septoria nodorum blotch) was once a prominent pathogen of UK wheat [369], but has since been largely overshadowed by *Z. tritici*. Nevertheless, molecular surveys of leaf samples have continued to detect *P. nodorum*, albeit at low frequencies, often in co-infection with *Z. tritici* [185].

Aside from *Puccinia*, these genera share broadly similar lifecycles, with primary inoculum often originating from crop residues, followed by airborne dispersal of spores and repeated cycles of infection throughout the growing season. In addition, they exhibit similar environmental requirements for germination and sporulation, typically favouring temperatures above 20 °C and the presence of free water or high relative humidity (Table 5.1). Such conditions were most consistently met in July 2023, which aligns with the timing of pathogen detection later in August that year. In 2024, suitable environmental conditions were also present between July and September, potentially explaining the observed secondary peaks in airborne abundance during that period (Figure 5.11).

However, the April 2024 peak in airborne pathogen abundance of multiple taxa is more difficult to explain. Weather data for this period do not indicate any clear short-term triggers, such as sudden increases in rainfall or relative humidity (Figure 5.10). It is possible that the peak reflects a combination of earlier environmental conditions that enabled infection and colonisation, together with subsequent favourable conditions for spore maturation and release. In addition, unmeasured microclimatic factors, such as canopy-level humidity or dew formation, may have contributed to enhanced pathogen development and dispersal, even in the absence of extreme weather events [231].

Puccinia is an obligate biotroph and therefore requires green tissue to establish infection. This likely explains its absence late in 2023, when the other taxa were detected, as no green tissue would have been present. In 2024, early signs of yellow rust, caused by *Puccinia striiformis* f.sp. *tritici*, were observed on wheat plants from 23 February, with slow progression through March. Airborne *Puccinia* was already detectable at very low relative abundance in the initial sample on 7 February. Levels remained low or undetectable in subsequent collections, before rising modestly in late March and peaking sharply in April. These results demonstrate the capacity of AirSeq to detect airborne pathogen DNA several weeks before visible symptoms become widespread.

Importantly, the early low-level detection in February may be more valuable for disease management than the later spike in April, as by that point, substantial infection may have already occurred, reducing the effectiveness of preventative control measures. By the end of the season, a high proportion of plants in the field showed visible symptoms of yellow rust (Figure 5.9), suggesting that earlier action, relying on AirSeq data might have mitigated disease impact. However, a similar level of yellow rust was observed in 2023 despite the absence of airborne *Puccinia* detection. As discussed earlier, this may be due to limitations in the sampling window or biases introduced during laboratory processing

The changing airborne pathogen abundance observed in this study somewhat aligns with the microscopy-based survey reported by Pilo et al. [287]. In that study, *Fusarium* and *Puccinia* spores were consistently detected across all sites using microscopy, with different abundance peaks observed at different locations, supporting the idea that local climatic conditions strongly influence airborne spore dynamics [287]. Metagenomic analysis of pooled samples from the same study also identified *Parastagonospora* and *Pyrenophora*, although the approach lacked temporal resolution due to the need to combine weekly samples to obtain sufficient DNA for sequencing. This highlights a key advantage of the AirSeq approach used in the present study, which enabled detection of all four genera at high temporal resolution, using 120-minute samples.

5.6.4.4 Single late-season Peaks: *Ustilago* and *Zymoseptoria*

The final two genera examined in more detail were *Ustilago* and *Zymoseptoria*, both of which exhibited mid- to late-season peaks in airborne abundance. In 2023, detections occurred between July and August, while in 2024, both genera were detected primarily in July (Figure 5.7). *Zymoseptoria* was additionally identified at very low levels throughout the 2024 season.

Ustilago tritici, the causal agent of loose smut in wheat, is initially seed-borne. When contaminated seed is sown, the pathogen germinates at the same time as the plant and then grows inside it. Spore release then occurs during flowering, and transmission to new hosts occurs via wind-dispersal [31]. Therefore, *Ustilago* spores are most likely to be detected during wheat flowering. In England, wheat typically flowers between mid-May and mid-June, aligning with the mid-June detections of airborne *Ustilago* in this study. This observation is consistent with previous research, including a 13-year microscopy-based study in England that also reported a narrow *Ustilago* detection window during June [357].

These results demonstrate the effectiveness of AirSeq in detecting airborne spores at specific seasonal windows which align with known pathogen biology. However, the relevance of this information for growers is limited in the case of loose smut, as the disease is

managed via seed sterilisation. Nevertheless, airborne detection of *Ustilago* may still have value in broader monitoring schemes. For instance, to track national disease trends, inform predictive models based on environmental conditions, or issue alerts in regions where elevated airborne levels may suggest an outbreak. This alert could guide decisions around seed treatment or discourage seed saving from affected crops.

Airborne *Zymoseptoria* showed a single sharp peak in abundance in both years: in August 2023 and in July 2024. Septoria tritici blotch, caused by *Zymoseptoria tritici*, was observed in both years, though with greater severity in 2024 (Figure 5.9). It is therefore reassuring that AirSeq detected this pathogen in airborne samples, although the relative abundance of *Zymoseptoria* was similar in both years and did not reflect the difference in disease severity. This discrepancy may be attributed to differences in host susceptibility between the wheat varieties grown, or to the shorter sampling duration and use of WGA in 2023.

In 2023, *Zymoseptoria* was only detected in August, whereas in 2024, low levels were present as early as February, followed by a sharp increase in July. This pattern aligns with informal 2024 field observations, which reported Septoria symptoms on plants from February onwards. As with *Puccinia*, this highlights the importance of early-season monitoring for detecting initial infections, which may precede airborne peaks caused by secondary sporulation cycles.

The airborne abundance patterns of *Zymoseptoria* observed in 2024 likely reflect the pathogen's life cycle. Primary infections are initiated by ascospores, produced during the sexual stage on overwintering crop debris, which are typically released from autumn to spring [54]. These spores infect young plants, after which a latent period follows before conidia are produced. Conidial sporulation requires warm and wet conditions, and the resulting spores are splash dispersed to neighbouring tissues [54].

Low-level detections of *Zymoseptoria* in February 2024 may therefore represent ascospores initiating primary infections in the field. A pronounced peak in July 2024, consistent with the spike observed in August 2023, is more likely to reflect conidial production and dispersal. Finally, the high abundance recorded in October 2024 may indicate renewed ascospore release, corresponding with the onset of the 2024–25 growing season.

However, it remains uncertain whether splash-dispersed conidia can be reliably captured by air samplers, and DNA sequencing alone cannot distinguish between spore types. Further validation using microscopy would therefore be required. These observations highlight the importance of sampling beyond the main growing season to better capture the interplay between crop developmental stage, environmental conditions during spore development, and the presence of airborne inoculum.

This deeper analysis of nine key genera highlights the ability of AirSeq to reliably detect airborne eDNA from pathogens known to be infecting crops within the sampling field. In several cases, peaks in airborne abundance aligned closely with environmental conditions favourable to spore release and with observed disease symptoms, demonstrating the method's potential for early detection and in-season disease monitoring.

While some limitations remain, such as the absence of airborne *Claviceps* despite confirmed field presence, these are likely to be taxon-specific. Further research into detection distances and threshold inoculum levels would help improve interpretation and provide clearer guidance for growers.

Overall, AirSeq showed strong agreement between airborne pathogen dynamics and

known environmental and biological triggers. Although some sharp peaks in airborne abundance could not be directly linked to specific environmental factors, this likely reflects the inherent complexity of the airborne microbiome and the influence of unmeasured or microclimatic variables.

5.6.5 Experimental limitations and challenges

There are several limitations to this study, the first being the difference in sample duration between the two years. The increased sampling length in 2024 means the results are not directly comparable to those from 2023, particularly as the 2023 samples were subjected to WGA. The decision to increase sample duration in 2024 was made to improve DNA yield and eliminate the need for WGA, which likely reduced amplification bias and improved sequencing quality. While this limits direct year-to-year comparisons, the consistency of sample length within each year does allow for valid comparisons of taxa across the season.

As with any airborne monitoring study, the length of collection influences which taxa are detected, with longer or more frequent collections generally increasing observed diversity. To balance practicality and cost, weekly collections of 30 minutes (2023) or two hours (2024) were used in this experiment. This frequency and duration likely meant that some species were not detected due to sporadic presence or spore release occurring outside the sampling windows, a risk that was greater in the shorter 2023 collections. Weekly sampling provided much greater resolution than the monthly collections used in the initial chapter (Chapter 3). However, the highly dynamic nature of the airborne microbiome suggests that even finer temporal resolution would reveal additional fluctuations in community composition.

Although the primary focus of this study was pathogens relevant to wheat, the sampling was conducted at an experimental farm where other cereal crops were also grown. As a result, it is likely that the airborne eDNA included pathogens of other hosts, including species of *Ustilago* and *Blumeria* not associated with wheat. These species were not monitored or scored during plant assessments, meaning their presence in the airborne dataset could not be confirmed or linked to visible disease symptoms. Nevertheless, this highlights the potential of AirSeq to capture a broader spectrum of airborne plant pathogens, though species- or strain-level certainty would be required to fully interpret these signals.

A further challenge in interpreting the data arises from the complexity of environmental interactions. With so many taxa detected and multiple dynamic environmental variables, it is difficult to determine which factors are most relevant to the presence or abundance of each pathogen. For example, it remains unclear whether conditions at the time of sampling or those in the days prior, such as sustained humidity or temperature ranges, are more influential, particularly when considering germination, infection, and sporulation cycles.

Moreover, differences in dispersal biology add further complexity [231]. Splash-dispersed pathogens such as *Z. tritici* are likely to reflect local sporulation events [191], whereas rusts (*Puccinia* spp.) produce airborne spores capable of travelling long distances [158], and thus may reflect infections originating further afield. Both spore types have the potential to infect new crops where AirSeq is sampling, but knowledge of their likely source could help refine management strategies. Future studies may benefit from species-specific modelling approaches that consider dispersal biology and lagged environmental variables alongside real-time conditions.

There are also broader challenges highlighted across the experiments in this thesis, which were introduced in the literature review (Chapter 2). These include the measurement of relative rather than absolute abundance, uncertainty over whether the captured DNA is viable, the risk of contamination during sample processing, and difficulty in determining the source of the detected DNA.

5.7 Conclusion and Future Work

The data presented here demonstrate the utility of AirSeq for monitoring airborne taxa and quantifying the abundance of specific fungal and oomycete pathogens throughout the growing season.

The composition and relative abundance of different phyla and dominant genera were shown to fluctuate across the season and between years, reflecting the dynamic and diverse nature of the airborne microbiome.

Focused analysis of nine genera of known wheat pathogens further illustrated the value of AirSeq for disease surveillance, allowing simultaneous monitoring of multiple pathogens. Changes in airborne abundance for these genera generally aligned with known environmental and seasonal conditions that support spore germination and dispersal. However, several high peaks in abundance could not be readily explained by the environmental variables considered in this study. This suggests the influence of additional factors that were not included in this study such as local host material, wind direction, canopy-level microclimates, soil temperature, or dew formation.

Additionally, although many pathogenic genera were detected in the air, only a subset were observed as visible disease symptoms on the plants. Conversely, one disease was identified in a nearby plot but was not detected in the airborne dataset. This highlights the complexity of interpreting airborne eDNA in relation to actual plant infection. The presence of pathogen DNA in the air does not necessarily indicate successful infection or disease development, and the absence of detection does not rule them out, as both processes depend on a range of interacting factors.

Future work should aim to determine critical airborne inoculum thresholds and how these relate to environmental factors that increase disease risk. Identifying such conditions would be highly valuable for growers, enabling more targeted and timely fungicide applications. This, in turn, would support efforts to reduce chemical inputs while maintaining effective crop protection.

One way to evaluate the benefit of AirSeq data would be through a field trial with control and experimental treatments. Control fields would be managed under standard practices, while experimental fields would incorporate AirSeq data to guide disease management strategies. Comparing disease levels, crop yields, and fungicide use between the two groups could provide insights into the effectiveness and practical value of AirSeq.

Additionally further work should consider integrating more detailed spatial and microclimatic data, such as wind trajectories, crop growth stage, and in-canopy humidity, to better understand the drivers of airborne pathogen dynamics. Extending the sampling duration could help capture pathogens that release spores outside the current early morning sampling window, while increasing sampling frequency would generate a richer dataset for correlating environmental conditions with detected taxa. However, these approaches would require significantly more time and may increase overall experimental costs, which

could limit feasibility for large-scale or long-term monitoring efforts.

Chapter 6

Diurnal airborne microbiome composition

6.1 Abstract

The AirSeq method was used to monitor airborne environmental DNA (eDNA) composition over a 24-hour period in August 2023 and again in May and June 2024. During these campaigns, both shorter (2 hours) and longer (up to 24 hours) samples were collected concurrently. The analysis examined species unique to each collection, community-level patterns at the phylum and genus levels (focusing on abundant fungi and oomycetes), and detailed profiles of nine genera containing key wheat pathogens. The results highlight the dynamic nature of the airborne microbiome, with clear fluctuations in detected taxa across the 24-hour cycle, and between months and years. Longer collections captured most taxa observed in shorter intervals, whereas shorter samples typically detected only a subset. However, shorter samples occasionally revealed taxa absent from longer collections, likely reflecting transience or stochastic effects. Importantly, shorter samples also identified peaks in pathogenic spore abundance throughout the day. These findings demonstrate that both the timing and duration of sampling strongly influence the diversity and abundance of taxa detected.

6.2 My Contributions

Dr Heavens collected all 24-hour samples in 2023 and 2024. I performed the library preparation and sequencing for the 2023 samples, while Dr Heavens carried out this work in 2024. I independently conducted all bioinformatic work, including data cleaning, preparation, and analysis. I also used ChatGPT (GPT-5) [273] to refine my code and improve the grammar and flow of my writing.

6.3 Introduction

The 24-hour time-series experiments were designed to investigate how biological and environmental factors shape the airborne microbial community in agricultural settings, and to assess how the timing of sample collection influences the diversity of pathogenic taxa detected. This chapter builds on the knowledge from the regular collections at Church Farm (Chapter 5) with a focus on more intensive sampling over a 24-hour period to understand finer temporal dynamics.

Air is a highly dynamic and fluctuating environment, making it challenging to capture an accurate representation of microbial abundance and diversity. A wide range of environmental conditions, including temperature, pressure, relative humidity, wind speed and UV radiation, can influence which taxa are detected. Previous studies have shown that fungal spore release is often triggered by climatic events such as humidity, morning dew or strong winds, meaning that different species peak in abundance at different times of the day [138, 206].

Although diurnal spore-release patterns are well characterised for a subset of individual pathogens, much less is known about how the broader airborne community fluctuates across a full 24-hour cycle in agricultural environments. Moreover, most diurnal studies to date have relied on microscopy, which either targets specific species [61, 79, 101] or aggregates observations into broad taxonomic groups [138, 287]. There is therefore a need for a more holistic approach capable of simultaneously monitoring a wide range of taxa to determine when community diversity is greatest and when particular species are most detectable. The AirSeq method addresses this need by coupling high-throughput sequencing with high-volume active air sampling.

To address these questions, 24-hour experiments were conducted in an agricultural field in August 2023 and again in May and June 2024. During each 24-hour period, samples were collected consecutively in 4-hour intervals (2023) or 2-hour intervals (2024), with additional 6-hour collections included. In 2024, full 24-hour samples were also taken on the days immediately before, during and after the main experiment. The analysis considered phylum-level patterns, fungal and oomycete genera composition, and detailed profiling of the nine genera known to contain major wheat pathogens which were introduced in Chapter 5.

The experiments demonstrated that community composition shifted over hours, days and months, with patterns more clearly resolved at the genus than the phylum level. Shorter collections were particularly effective for capturing diurnal fluctuations and transient peaks in certain pathogenic taxa, whereas longer collections tended to capture greater overall diversity but sometimes missed rare taxa present only briefly. These findings highlight the trade-off between sampling duration and temporal resolution, and show that finer

temporal resolution reveals more pronounced fluctuations in the airborne community.

6.4 Methods

6.4.1 Sample collection

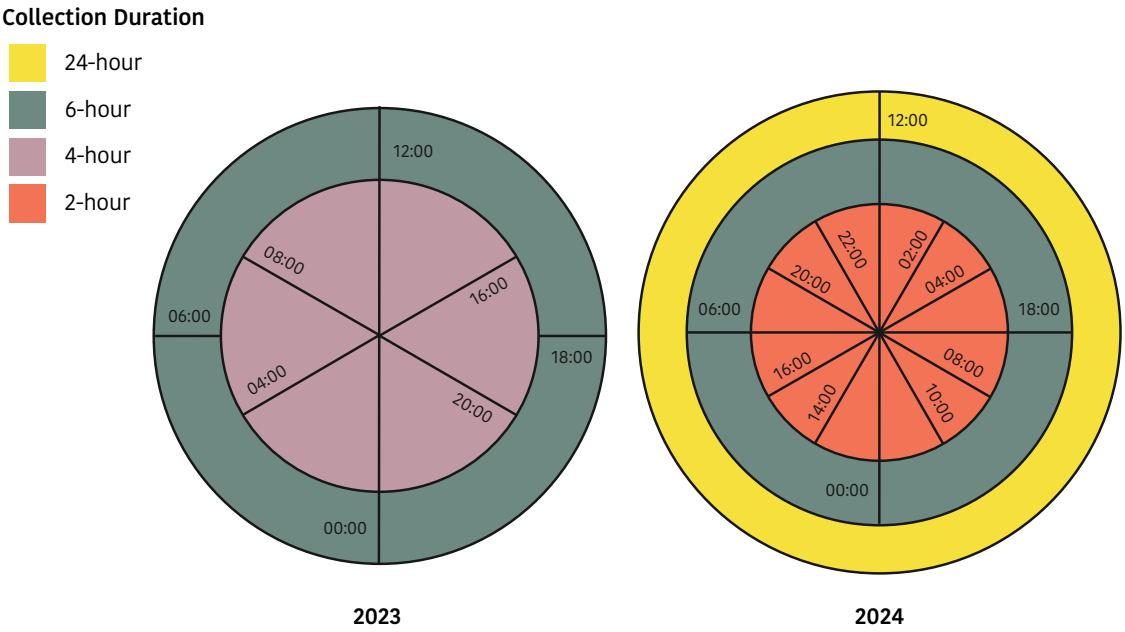


Figure 6.1: Sampling schedule for 2023 and 2024 showing the start and end time for the different collections across a 24-hour period.

Air samples were collected over 24-hour periods using the InnovaPrep Cub sampler, operating at a constant flow rate of 200 L/min. Sampling was conducted at Church Farm, Norfolk, in the same locations as the season-long campaigns described in Chapter 5 (section 5.4.1), within a field that was not being treated with fungicides. Three sampling campaigns were undertaken: 8–9 August 2023, 13–16 May 2024, and 17–20 June 2024. Sampling schedules are illustrated in Figure 6.1.

The 2023 campaign consisted of simultaneous 4-hour and 6-hour collections over a single 24-hour cycle beginning at midday on 8 August. Informed by these results, the 2024 campaigns incorporated increased replication and a broader range of sampling durations. In both May and June 2024, consecutive 24-hour collections were made over three days (13–16 May and 17–20 June). Within these periods, the 2- and 6-hour samples were collected only during the central 24-hour window (14–15 May and 18–19 June), with 24-hour samples collected on the surrounding days. All 2024 samples were collected in duplicate to improve statistical robustness.

6.4.2 Environmental data

Environmental data were recorded at Church Farm using an on-site weather station. Measurements included wind speed (m/s), quantum radiation (μmol m^{-2} s^{-1}), and relative humidity (%) at 10-minute intervals, as well as hourly averages for air temperature (°C) and barometric pressure (mbar). These data were exported as CSV files and integrated with sequencing outputs during downstream analysis to contextualise variation in the airborne microbial community during each sampling campaign.

6.4.3 DNA extraction and sequencing

Samples were processed as described in Chapter 4 (section 4.4.1), using the standard protocol involving elution foam, syringe filtration, bead beating, and magnetic bead clean-up. The 2024 samples were processed according to the modified protocol outlined in Chapter 5. None of the samples underwent amplification prior to sequencing.

Negative and positive controls were included to monitor data quality. For each sampling month, a laboratory negative control was generated by placing a filter in an inactive sampler for 30 minutes before subjecting it to the same processing steps as the experimental samples. The 2024 campaigns also included a lambda DNA positive control, sequenced alongside the field samples.

Library preparation was carried out using the ligation barcoding kit (2023: SQK-NBD114.24; 2024: SQK-NBD114.96) from ONT according to the standard protocol. In each case, 40 ng of DNA in 11 μl was used per sample. The 2023 barcoded library was sequenced on a MinION R10.4.1 flow cell, and the 2024 library on a PromethION R10.4.1 flow cell.

6.4.4 Data analysis

Raw read data from the 24-hour air sampling experiments were rebasecalled using the super accuracy model in Dorado (v0.7.2) to enhance alignment quality. Before taxonomic classification, reads were subsampled to a uniform depth of 200,000 per sample using a randomised selection approach, performed with the script `subsample_reads.sh`, available in the GitHub repository `https://github.com/Mia-FGB/hpc-scripts`.

Taxonomic classification of subsampled reads using BLAST nt was performed with MARTi [281] (2023: v0.9.22, 2024: v0.9.21). The version difference reflects updates applied to the 2023 analysis to ensure consistency with the processed 2024 dataset. Classification used the following parameters: `LCAMaxHits:100, LCAScorePercent:90, LCAMinIdentity :70, LCAMinLength:150, LCAMinReadLength:200`. The read count table was then exported from the front-end with a LCA cut-off of 0.1 % and the resulting taxonomic assignments were extracted for downstream analysis.

Taxonomic and alignment data were analysed and visualised using R and Python. All scripts used in the analysis of the 24-hour dataset are available in the GitHub repository: `https://github.com/Mia-FGB/24hr_analysis`. In R, the following packages supported data wrangling and visualisation analysis: `dplyr`, `ggplot2`, `tidyr`, `reshape2`, `scales`, `lubridate`, `patchwork`, `grid`, and `phyloseq`. Specific analyses included calculation of the proportion of unique and shared taxa (`Total_Uniq_Species.R`), stacked bar plots of phyla and fungal genera (`24hr_stacked_bar.R`), and relative abundance plots of nine key pathogens (`pathogen_plots.R`). Environmental data were visualised in Python using `pandas` and `seaborn`, implemented in the Jupyter notebook `weather_plots.ipynb`.

6.5 Results

To investigate diurnal patterns in airborne taxa, 24-hour sampling was carried out at Church Farm in August 2023 and again in May and June 2024. The 2023 campaign used 4- and 6-hour intervals, while the 2024 design incorporated 2- and 6-hour intervals with replication to improve temporal resolution and robustness.

6.5.1 Species richness and unique species per timepoint

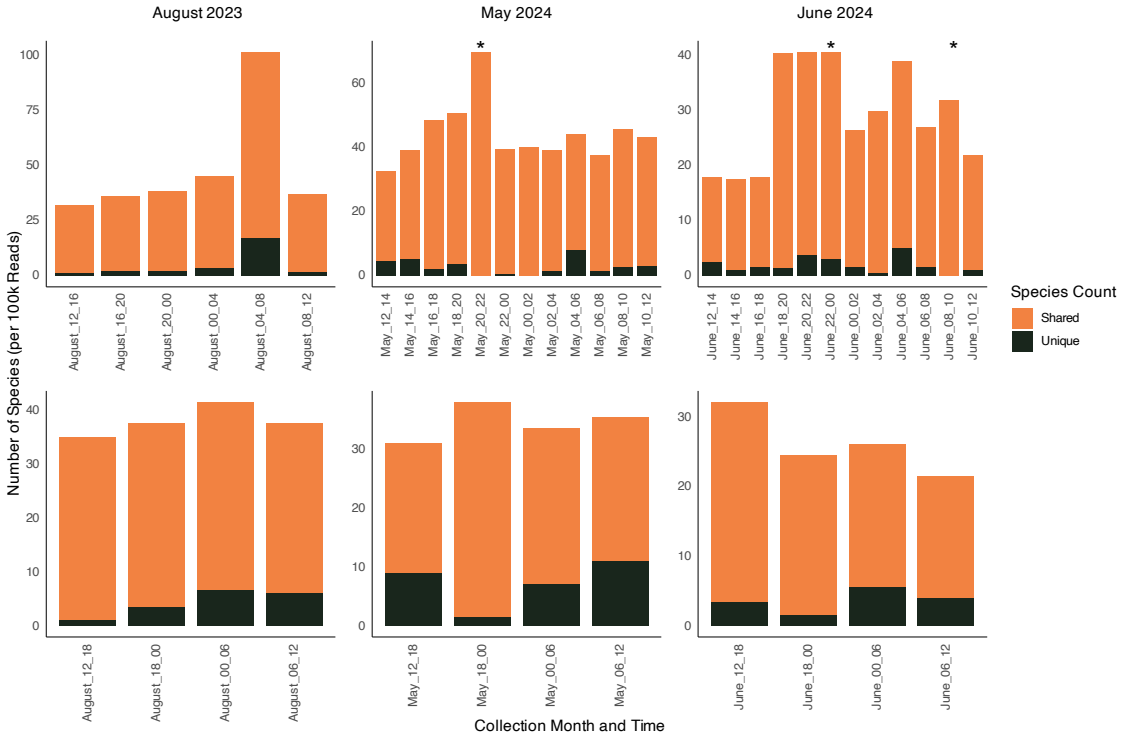


Figure 6.2: Species richness across timepoints in air microbiome samples collected over 24-hours. Bar plots show the number of shared species (orange) and the number unique to each timepoint (black), normalised per 100,000 reads. In 2023, one sample was collected per timepoint; in 2024, each timepoint was sampled in duplicate and bars represent the average across replicates. Samples were collected at 2-, 4-, and 6-hour intervals. Data were filtered to include taxa with HPM ≥ 100 and read count > 5. Asterisks (*) indicate timepoints where one replicate has fewer than 2,000 reads.

The species richness and number of unique species identified per duration and collection date varies across time points (Figure 6.2), although it is important to note that these unique species represent a small normalised read count (Figure 6.3).

In general, 6-hour samples display more consistent species richness across timepoints, while shorter collections show greater variation, highlighting time-dependent peaks in diversity. Across majority of the samples, the number of unique species increases proportionally with total species richness. The exception is early evening in May (18:00 - 00:00) where both the 2 hour and 6 hour 20:00 - 22:00 show high species richness but no or few unique species.

In August 2023, the 04:00–08:00 sample contained nearly twice the number of species per 100,000 reads compared to other timepoints. This pattern was not observed in the corresponding 6-hour samples, which showed similar richness across all intervals (30–40 species per 100,000 reads).

In the 2024 samples, species richness patterns varied between months. In May, the 20:00–22:00 sample had the highest richness, while in June, peaks occurred at 04:00–06:00 and in all the samples from 18:00–00:00. Both months showed a decline in richness during the early afternoon.

Together, these data indicate that airborne microbial species richness fluctuates over a 24-hour cycle and varies between sampling days, likely due to environmental and seasonal factors. However, patterns of increased richness at dawn and dusk, and lower richness in

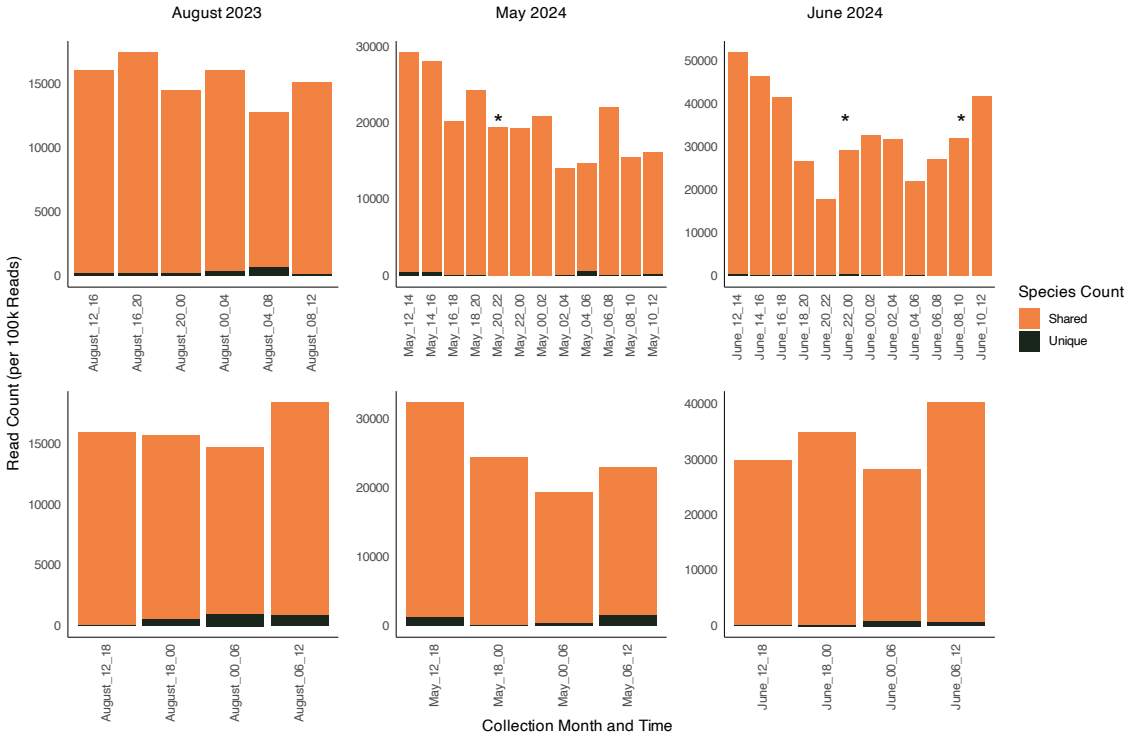


Figure 6.3: Normalised read counts of unique and shared taxa across timepoints in air microbiome samples collected over 24-hours. Bar plots show the number of shared species (orange) and the number unique to each timepoint (black), normalised per 100,000 reads. In 2023, one sample was collected per timepoint; in 2024, each timepoint was sampled in duplicate and bars represent the average across replicates. Samples were collected at 2-, 4-, and 6-hour intervals. Data were filtered to include taxa with HPM $\geq$ 100 and read count $>$ 5. Asterisks (*) indicate timepoints where one replicate has fewer than 2,000 reads.

early afternoon, are consistent. Additionally, longer sampling intervals appear to smooth out temporal variation, suggesting they may provide more stable profiles of the airborne microbiome.

While species richness highlights overall taxonomic diversity, examining community composition at the phylum level provides further insight into shifts in the dominant biological groups present in the air.

6.5.2 Phylum-level composition detected across time and season

Figures 6.4 and 6.5 present the relative abundance of phyla (>0.1%) across samples.

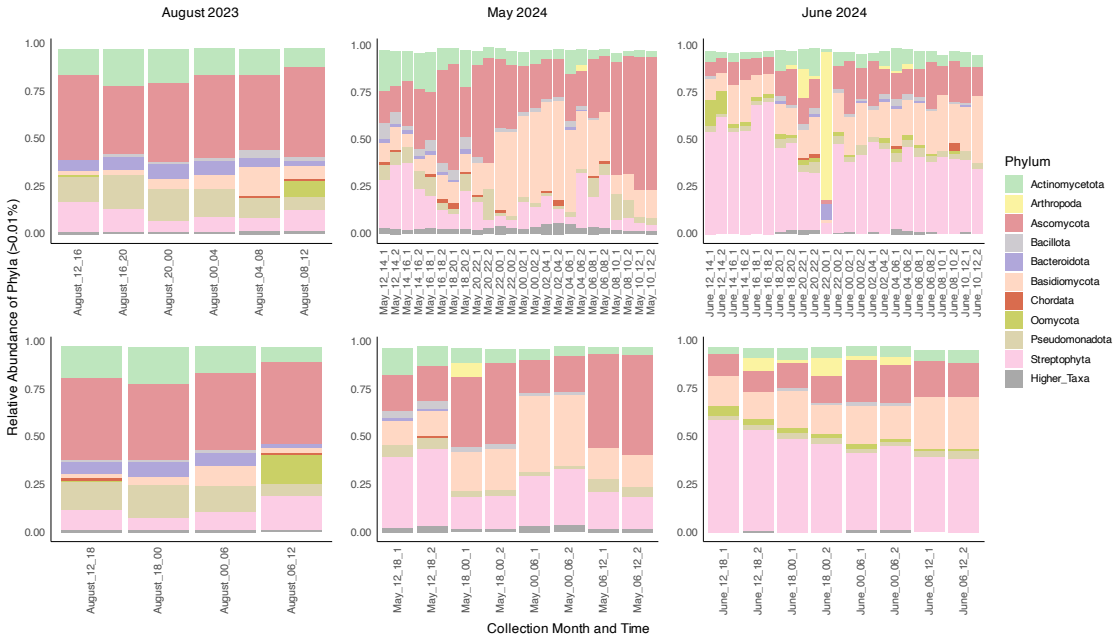


Figure 6.4: Relative abundance of phyla (>0.01% of total reads) in air microbiome samples from the 24-hour experiment. August 2023 (4- and 6-hour intervals) and May/June 2024 (2- and 6-hour intervals with replicates).

Shorter (2- and 4-hour) and longer (6-hour) samples from overlapping timepoints show consistency in phylum-level composition (Figure 6.4) Several phyla appear consistently across all months, including Actinomycetota, Ascomycota, Bacteroidota, Basidiomycota, and Streptophyta. However, their relative abundances vary, for example Basidiomycota and Streptophyta are more prominent in 2024 samples, while Ascomycota dominates in August 2023 and May 2024. June 2024 samples are composed of roughly 50% Streptophyta. A few exceptions occur, such as replicates with elevated Arthropoda content, most notably the first replicate from June 22:00–00:00, which likely reflects accidental insect capture during sampling.

The three-day 24-hour sampling in 2024 (Figure 6.5) reveals similar phylum profiles across days, with most May samples and all June samples dominated by Streptophyta. Notably, Arthropoda reads are detected in all 24-hour samples, particularly in both May 15 collections, where they exceed 50% abundance. This suggests that longer sampling durations may increase the risk of insect contamination, potentially skewing community profiles and reducing sensitivity to rare airborne taxa due to the over representation of non-target DNA.

To explore finer-scale taxonomic changes, the relative abundance of fungal and oomycete

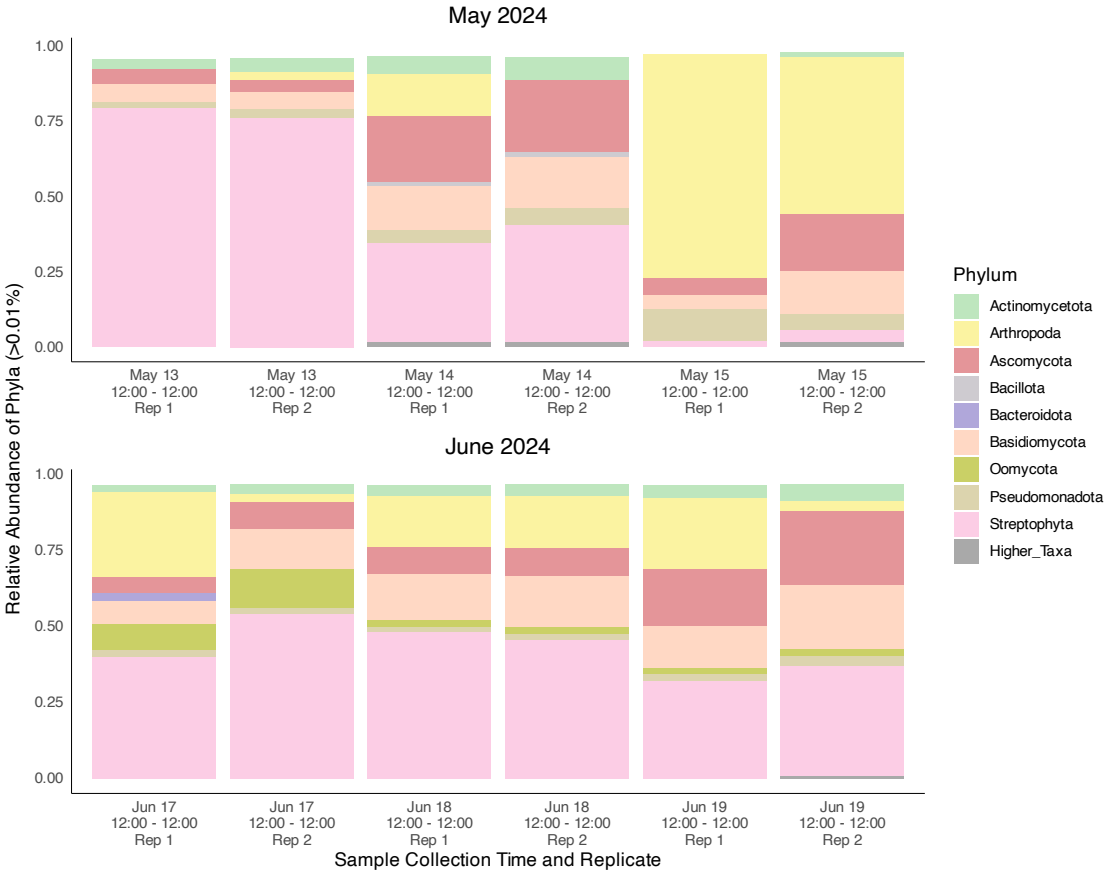


Figure 6.5: Relative abundance of phyla (>0.01% of total reads) in air microbiome samples from the 24-hour experiment. May/June 2024 data were collected over three consecutive days, with two replicate 24-hour samples per day.

genera was assessed across all timepoints and months.

6.5.3 Fungal and oomycete genera detected across time and season

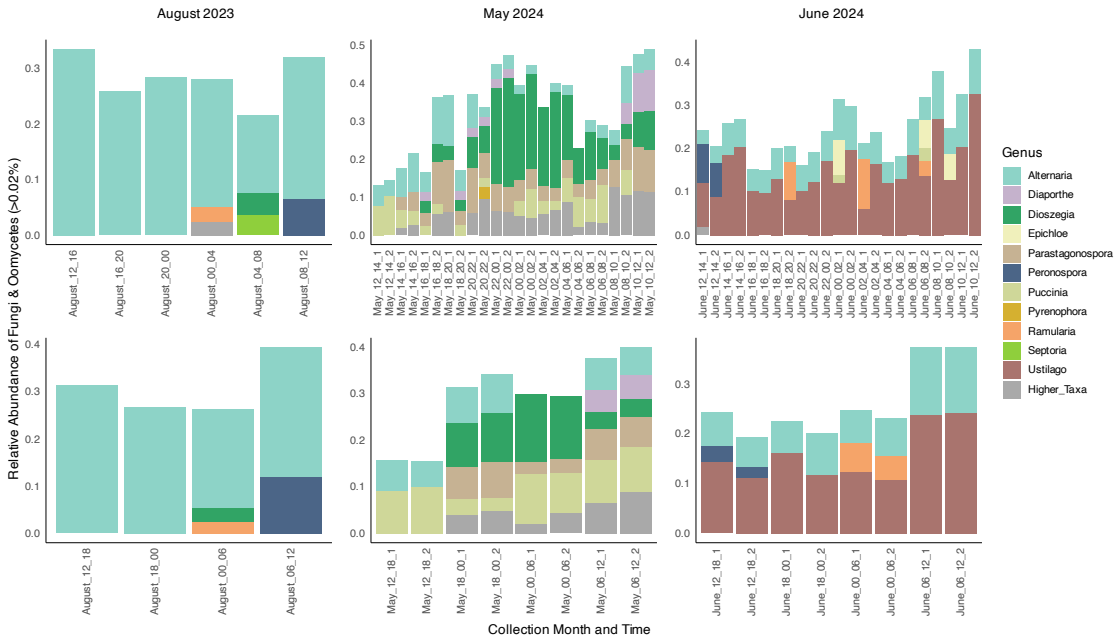


Figure 6.6: Relative abundance of fungal and oomycete genera (>0.02% of total reads) in air microbiome samples. August 2023 (4–6 h) and May/June 2024 (2–6 h) intervals.

Figures 6.6 and 6.7 illustrate the temporal and seasonal variation in fungal and oomycete genera identified in air microbiome samples. Clear differences are observed between months, days and timepoints.

Distinct taxonomic profiles were associated with each month (Figure 6.6). August 2023 was dominated by *Alternaria*, whereas May 2024 showed a more balanced composition, with *Puccinia, Dioszegia, Parastagonospora* and *Alternaria* all present. In contrast, June 2024 was largely dominated by *Ustilago*, with additional genera such as *Peronospora* detected in late afternoon samples, and *Ramularia* appearing around dawn.

The August 2023 4- and 6-hour samples displayed strong internal consistency across overlapping timepoints. For instance, *Peronospora* was only detected in the final collections (08:00–12:00 for 4-hour, and 06:00–12:00 for 6-hour samples). Similar reproducibility was observed in the 2024 datasets, where taxa such as *Dioszegia* in May and *Peronospora* in June were consistently detected or absent across replicates, suggesting time-specific patterns in their presence. Some genera, including *Epichloë* and *Puccinia* in June, were detected only in the 2-hour samples but not in the corresponding 6-hour collections, indicating that shorter sampling intervals may increase sensitivity to transient taxa.

Replicated 6-hour samples collected in 2024 consistently recovered the same genera, indicating a high degree of reproducibility. In contrast, some 2-hour samples exhibited variability between replicates, with certain taxa detected in only one sample. For example, *Pyrenophora* was found exclusively in the second replicate of the 20:00–22:00 May sample.

Figure 6.7 further illustrates daily variation across the three-day sampling periods in May and June. While dominant genera remain consistent within each month (e.g. *Puccinia* in May; *Ustilago* and *Alternaria* in June), day-to-day fluctuations were evident. For instance, *Puccinia* was the sole genus detected on 13 May, whereas *Parastagonospora*,

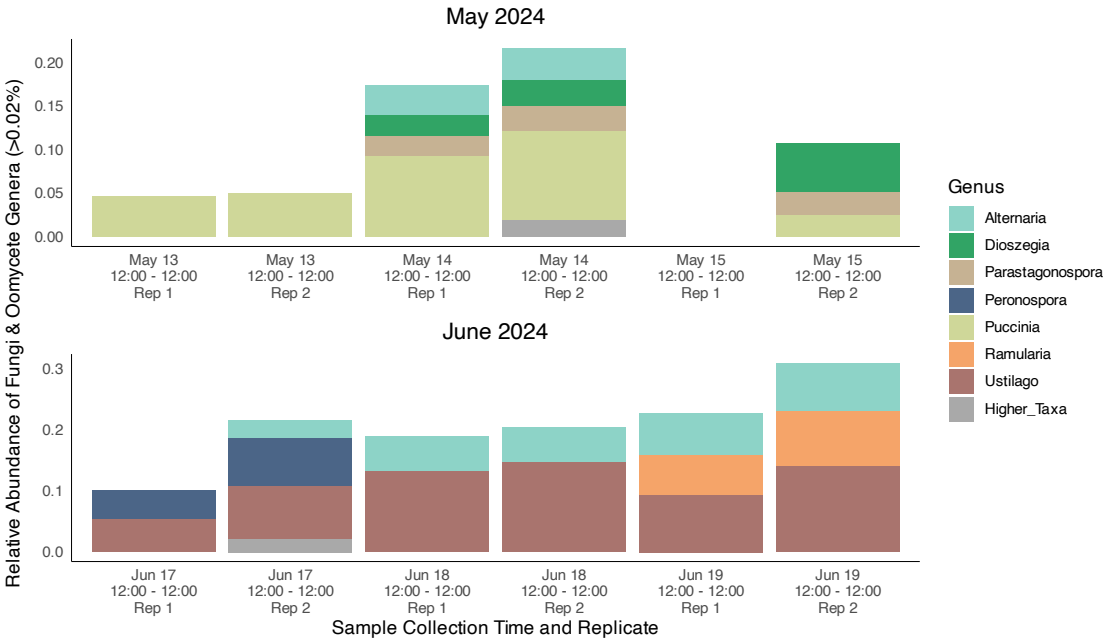


Figure 6.7: Relative abundance of fungal and oomycete genera (>0.02% of total reads) in air microbiome samples. May/June 2024, 24-hour samples collected over three days with two replicates per day.

Dioszegia and *Alternaria* appeared on the following days. Similarly, *Peronospora* was only observed on 17 June and *Ramularia* on 19 June.

A comparison of 2-, 6- and 24-hour samples collected simultaneously on 14 May and 18 June indicates broad consistency in dominant taxa. However, shorter duration samples detected additional genera, such as *Diaporthe* in May and *Peronospora* and *Ramularia* in June. These genera were also detected in 24-hour samples from adjacent days, suggesting that while the air microbiome is relatively stable at the daily scale, shorter timepoints offer finer resolution and may capture ephemeral taxa missed or diluted in longer collections.

In addition to overall fungal composition, specific pathogenic genera of agricultural importance were analysed to investigate their temporal dynamics and potential environmental drivers

6.5.4 Temporal patterns in selected fungal pathogenic genera

To investigate diurnal dynamics in agriculturally relevant fungal pathogens, nine genera were selected for detailed analysis: *Blumeria, Claviceps, Fusarium, Magnaporthe, Parastagonospora, Puccinia, Pyrenophora, Ustilago* and *Zymoseptoria*. The rationale for the selection of these taxa was discussed in detail in the prior Chapter 5. Understanding the temporal dynamics of these genera can guide sampling strategies for known pathogens. Of those selected, seven were detected in at least one collection. No reads were identified for *Claviceps* or *Magnaporthe*.

Figure 6.8 illustrates the variation in airborne abundance of each selected genus across different sampling months and over the 24-hour collection period. The temporal dynamics and monthly patterns for each genus are discussed in detail below.

In August 2023, the dominant genera were *Fusarium, Pyrenophora and Zymoseptoria*. In contrast, the May 2024 samples showed a broader diversity, with all genera detected except *Ustilago*. The June 2024 samples included all target genera, although *Fusarium*

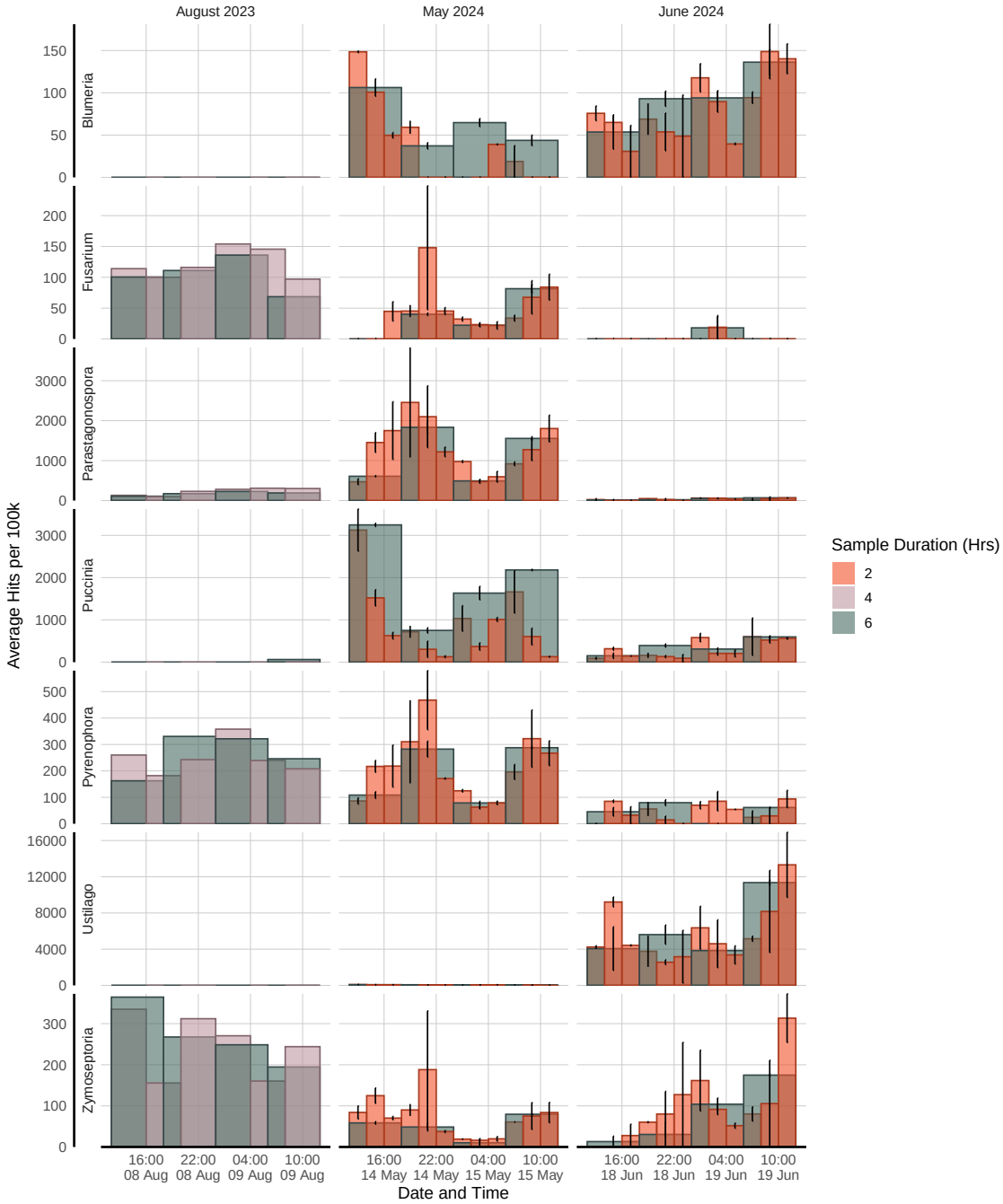


Figure 6.8: Pathogen read abundance (hits per 100,000 reads) over time across sampling months. Each panel shows one pathogen (rows) measured during each month (columns). Bars represent single values for 2023 and the mean of two replicates for 2024, coloured by sampling duration (2, 4, or 6 hours). Column width corresponds to sampling length, with date and time shown on the x-axis.

and *Parastagonospora* were present at considerably lower abundance than in the previous months.

Blumeria was not detected in the August 2023 samples but appeared in both May and June 2024. In May, its relative abundance peaked during the first collections of the day (12:00–14:00 and 18:00), reaching 149 HP100k. In June, *Blumeria* abundance steadily increased across the 24-hour period, culminating in a maximum of 181 HP100k in the final midday collection.

Fusarium was consistently detected across all three sampling months. In August, its abundance remained relatively stable at approximately 114 HP100k throughout the 24-hour period. In May, it was present throughout the day, peaking at 248 HP100k in the 20:00–22:00 sample. However, in June, *Fusarium* was only detected during the early morning, with the highest abundance of 38 HP100k observed in the 02:00–04:00 sample.

Parastagonospora was most abundant in May 2024, reaching a maximum of 3833 HP100k in the late afternoon collection (18:00–20:00). It was also present in the August and June samples but at substantially lower levels, with average abundances of 205 and 35 HP100k, respectively.

Puccinia followed a similar pattern, peaking in May during the 12:00–14:00 and 12:00-18:00 collections at over 3000 HP100k. In August, it was detected only in a single 06:00–12:00 sample (63 HP100k), while in June it was consistently present across all timepoints, with an average of 317 HP100k.

Pyrenophora was detected in all months, showing relatively stable abundance in August (average 255 HP100k) and June (average 45 HP100k), while May exhibited greater fluctuation, with a peak of 579 HP100k in the 20:00–22:00 sample and a decline to 100 HP100k between 00:00 and 06:00.

The levels of *Ustilago* in June were the highest of all genera, reaching a maximum of 16,916 HP100k in one replicate 10:00–12:00 collection. *Ustilago* was not detected in August and was present at low levels in May, with an average of 48 HP100k.

Zymoseptoria was detected across all sampling months. In August, its abundance peaked at 364 HP100k in the initial 12:00–18:00 sample and declined steadily thereafter (average 255 HP100k). In May, *Zymoseptoria* abundance peaked at 331 HP100k during the 20:00–22:00 timepoint (average 66 HP100k), while in June it reached a maximum of 372 HP100k in the final 10:00–12:00 sample (average 89 HP100k), indicating more dynamic fluctuations across the day.

In summary, each pathogenic genus exhibited a distinct temporal pattern, varying in abundance across collection months and over the 24-hour sampling period. For instance, *Pyrenophora* and *Zymoseptoria* were consistently detected in all three months, albeit at differing abundances, while *Parastagonospora*, *Puccinia* and *Ustilago* were identified at high abundance in a single month. Moreover, the timing of peak abundance differed between genera, likely reflecting a combination of biological factors and environmental conditions influencing spore release. The inclusion of 2-hour sampling intervals in 2024 provided greater temporal resolution, revealing short-lived abundance peaks that were not captured in the 4-hour intervals used in 2023.

6.5.5 24-hour environmental data

Weather conditions during the 24-hour sampling periods are shown in Figure 6.9.

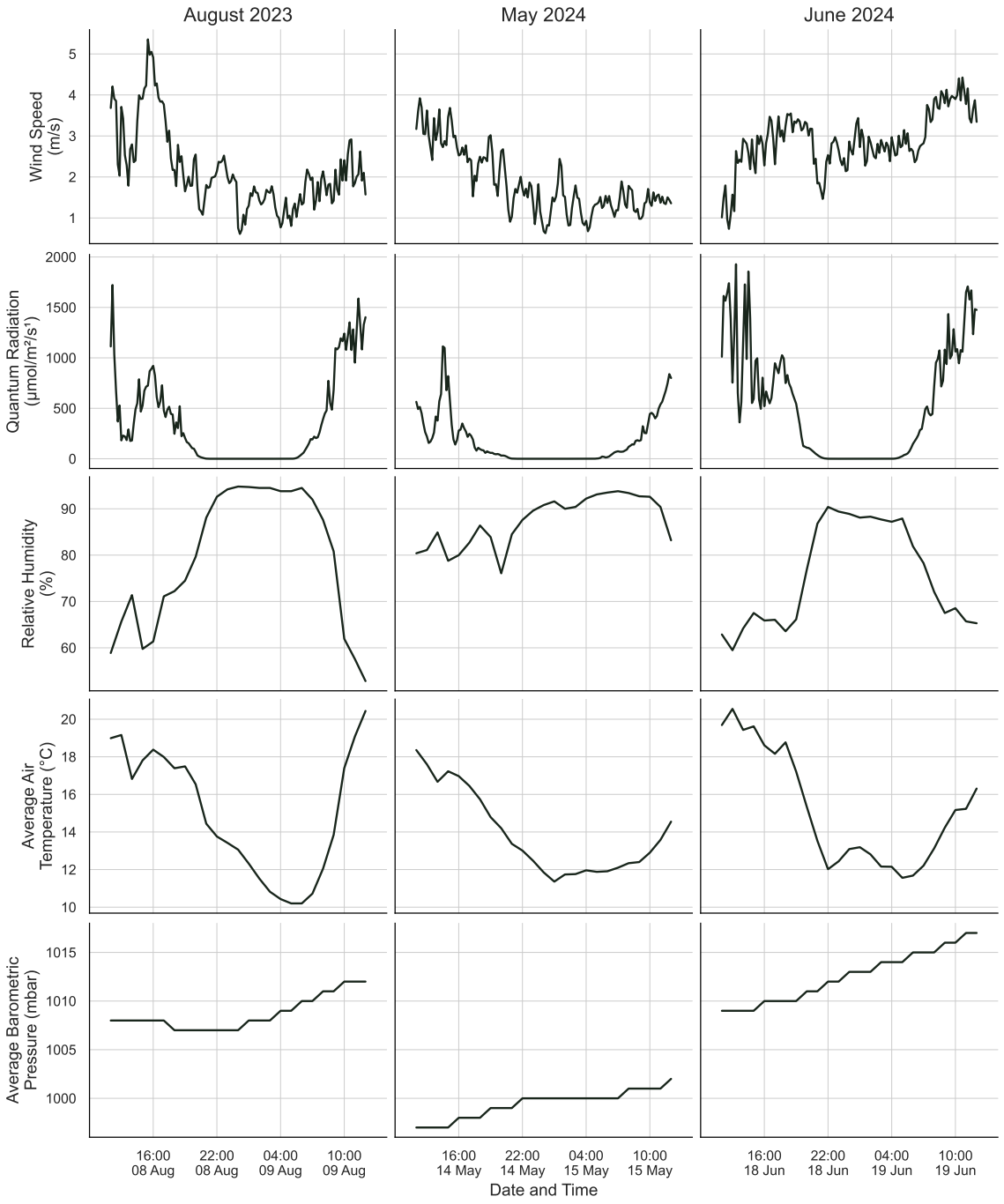


Figure 6.9: Weather conditions during air sampling campaigns in August 2023, May 2024, and June 2024. Each column represents a sampling period; each row shows a different meteorological variable recorded at the sampling site: wind speed (m/s), quantum radiation (μmol m^{-2} s^{-1}), relative humidity (%), average air temperature (°C), and average barometric pressure (mbar). Data were collected at 10-minute or hourly intervals and plotted across each 24-hour sampling window.

Wind speed remained between 1-5 m/s across the three sampling months, with observable variation over each 24-hour period. In August, wind speed peaked sharply at 5 m/s around 16:00 before declining and remaining below 3 m/s for the remainder of the sampling period. In May, the highest wind speed (4 m/s) coincided with the initial sample collection and subsequently dropped below 2 m/s. Conversely, June began with lower wind speeds (1–2 m/s), gradually increasing to 4 m/s by midday on the 19th.

Quantum radiation followed a broadly similar diurnal pattern across months, reflecting seasonal daylight variation. August exhibited the shortest night and highest radiation levels, exceeding 1,500 μmol m^{-2} s^{-1}. during both morning and afternoon periods. June showed a comparable pattern, while May had the longest night and only briefly exceeded 1,000 μmol m^{-2} s^{-1}.

Relative humidity exhibited greater variation between months than other variables. In all months, humidity exceeded 90% overnight. The lowest humidity was recorded in August, dropping to around 50% at midday, with June showing similar midday values (60–70%). In contrast, May maintained consistently high humidity throughout the 24-hour period, never falling below 75%.

Air temperature followed an inverse pattern to humidity across all months. Collections began at similar temperatures (18–21 °C), with the lowest values occurring overnight. In August, temperature dropped to 10 °C at 04:00 before rising above 20 °C by midday. May exhibited a narrower range (11–18 °C), with the minimum at midnight. June showed a similar trend, with temperatures peaking at 21 °C at midday on the 18th, dropping to 12 °C overnight, and increasing to 16 °C by the end of the sampling period.

Barometric pressure exhibited a consistent trend across all months, gradually increasing over the 24-hour sampling period. May recorded the lowest pressure, around 1,000 mbar. Both June and August began at slightly higher levels (<1,010 mbar), with June showing a greater overall increase by the end of the sampling window compared to August.

This environmental dataset highlights the complexity and interdependence of factors influencing spore release. All of the variables change both diurnally and seasonally, underscoring the need to consider multiple interacting conditions when interpreting airborne spore dynamics.

Together, these datasets from the 24-hour sampling campaigns provide a comprehensive view of how airborne microbial communities and key pathogens vary across daily and seasonal timescales. They also highlight how collection duration influences the detection of short-lived increases in airborne taxa.

6.6 Discussion

The 24-hour experiments were designed to track changes in the airborne fungal community over time, aiming to identify the best time of day to sample for maximum pathogen diversity and to better understand daily airborne patterns. The results revealed variations in community composition across hours, months and years, both at the phylum (Figures 6.4, 6.5) and genus levels (Figures 6.6, 6.7). This temporal diversity was further reflected in the abundance patterns of nine focal fungal genera, which were examined in greater detail (Figure 6.8). Longer sampling intervals were shown to detect greater taxonomic diversity but reduced the ability to capture transient or stochastically present taxa.

6.6.1 Species richness

In this study, species richness remained relatively consistent across the 6-hour samples in all three sampling months. However, the shorter 2- and 4-hour collections showed occasional spikes in diversity (Figure 6.2). Peak diversity was observed from 04:00–08:00 in August 2023, 20:00–22:00 in May 2024, and 18:00–00:00 in June 2024. These shorter samples also contained fewer unique species per time point, suggesting that many taxa persist in the air for longer than two hours.

While several microscopy-based studies have examined diurnal patterns for individual species [79, 101], or investigated airborne fungal communities over multiple days or longer periods [18, 98, 188, 324], few have assessed how species diversity varies across a 24-hour cycle. Research into time-resolved airborne microbial metagenomes demonstrates a clear diel cycle in taxa presence, despite many taxa being unassigned to species due to limited reference genomes [137]. Another study investigating diurnal variation in airborne microbes in Siberia reported that bacterial species richness peaked in the evening before declining overnight, while fungal richness was highest at night and decreased during the day [138]. In the dataset presented in this chapter, the timing of peak species richness varied across months when all kingdoms were considered together. However, when fungal and oomycete richness were examined separately, a consistent pattern emerged, with higher richness observed during the evening and at dawn. This trend aligns with the findings from the Siberian study.

6.6.2 Phyla composition

The relative abundance of phyla showed substantial variation between sampling years (2023 and 2024), notable shifts between months (May and June 2024), and smaller fluctuations across 24-hour periods (Figures 6.4, 6.5). A series of 24-hour samples collected on three consecutive days in May and June 2024, also revealed shifts in the dominant phyla from day to day. This was particularly evident in May, when the proportion of Streptophyta reads declined while the relative abundance of Arthropoda increased.

The major phyla identified in this study are consistent with those reported in the Siberian study, which included Actinobacteria, Pseudomonadota, Ascomycota, Basidiomycota and Streptophyta among the most abundant airborne taxa [138]. A global fungal spore sampling study similarly highlighted the widespread presence of Ascomycota and Basidiomycota taxa [4].

6.6.3 Fungal and oomycete genera composition

Building on the phylum-level comparison, genera of fungi and oomycetes with relative abundance $> 0.02\%$ were examined (Figures 6.6, 6.7), revealing substantial variation in composition across sampling months. Several genera showed distinct seasonal trends, suggesting potential biological or environmental drivers of their presence. For example, *Alternaria* was consistently detected across all months, indicating its continuous airborne presence. In contrast, *Dioszegia* and *Parastagonospora* showed seasonal specificity, with *Dioszegia* more abundant in May than August and *Parastagonospora* detected only in May. The dominance of *Ustilago* in June, alongside the presence of *Puccinia* in both May and June, may reflect agricultural activity or the life cycles of host plants during these months.

Alternaria was detected in all sampling months and at most time points, consistent with previous studies that report high airborne abundance of *Alternaria* spores [18, 345]. The data presented in this chapter show no consistent time of day when *Alternaria* was most abundant, and differences in relative abundance between time points were generally minor. This contrasts with microscopy-based studies that have found distinct daily peaks in *Alternaria* abundance, typically from 16:00-20:00 in Portugal (2005–2007) [271], or from noon to early afternoon in Poland (1997–1999) and the United States (1998) [58, 352].

Aside from *Alternaria*, the dominant genera identified in this experiment differed from those reported in a recent global fungal spore sampling study, which found *Cladosporium, Ascochyta*, and *Alternaria* to be the most prevalent genera worldwide [4].

These differences in abundant genera are likely due to variations in sampling years, geographical location, day length and environmental conditions. This highlights the complexity of interpreting temporal and diurnal patterns in the airborne microbiome and underscores the challenge of comparing results across studies that differ in location, timing and methodology.

It was also possible to identify genera present in the air for shorter periods, such as *Ramularia*, which was detected in the early morning samples (00:00–06:00) in both August and June, and *Peronospora*, observed in the late morning in August (08:00–12:00) and early afternoon in June (12:00–14:00). There is limited research on the diurnal presence of these spores, but *Ramularia* species are known to require leaf wetness to sporulate [145], often provided by morning dew, which may explain their occurrence during early morning periods. However, *Ramularia* was also detected at other times in June (06:00–08:00 and 18:00–20:00), suggesting that additional environmental factors may influence spore release.

A few studies have found that *Peronospora* spore release peaks in the morning. For instance, *P. antirrhini* (host: snapdragon) peaked between 05:00 and 12:00 [61], while *P. destructor* (host: onion) peaked around 08:00–09:00, coinciding with decreasing humidity and increasing wind speeds [153]. In August, *Peronospora* was detected after this reported window (08:00–12:00), while in June the peak occurred later than in the other studies, between 12:00 and 14:00. This later detection in June may reflect differing environmental conditions.

6.6.4 Nine genera of pathogenic interest

In addition to the broader fungal and oomycete communities, nine genera of fungal pathogens were specifically analysed to investigate diurnal variation during the 24-hour sampling periods (Figure 6.8). These genera were selected as they represent major wheat pathogens and have already been examined in detail in the seasonal dataset presented in Chapter 5. Here, the focus is on how their abundance fluctuated over shorter timescales, with a few genera of particular interest highlighted and discussed in the context of existing knowledge on spore release timing.

Puccinia was most abundant in the May 2024 samples, with a peak in the early afternoon. It was less abundant in June 2024 and was almost undetected in August 2023. Published research on *Puccinia psidii* in Brazil found that spores were most commonly detected at night, under conditions of high humidity and leaf wetness, along with low temperature, light intensity and wind speed [410]. This does not align with the findings from the current study, where *Puccinia* abundance in May was lowest overnight. Such

discrepancies may be explained by differences in species, geographical location or local environmental conditions.

Pilo et al. used microscopy and metagenomics to track cereal pathogen spore abundance across four sites. *Zymoseptoria* ascospores were consistently detected from May to August at all sites, with abundance peaking in May and early June, declining mid-season, and rising again in August [287]. This late-season increase was attributed to a resurgence in sexual reproduction in preparation for overwintering. Although the dataset presented here includes only a few sampling days per month, *Zymoseptoria* was also detected throughout the sampling period and was most abundant in August 2023, potentially reflecting similar seasonal dynamics.

The Pilo et al. study also reported the highest *Fusarium* spore release in May, with a decline by August. This contrasts with the pattern observed in the dataset presented here, where *Fusarium* was most abundant in August 2023 and least abundant in June 2024. In their findings, *Fusarium* abundance showed a positive correlation with RH, but was only statistically linked to air temperature at one of the four sites, suggesting that the environmental drivers of spore release may vary spatially.

Blumeria graminis was primarily detected in the earlier part of the season (May–June) in the Pilo et al. study, which is consistent with the findings presented: *Blumeria* was present in May and June but absent in August.

One important consideration when comparing the two datasets presented here is that the August samples were collected in the previous year, making it difficult to confidently compare monthly spore abundance. In addition, discrepancies may reflect inter-annual variation or differences in local environmental conditions.

6.6.5 Length of collection

Alongside examining 24-hour patterns in airborne fungal and oomycete spores and comparing them to established literature to assess the accuracy of AirSeq, this study also investigated how sampling duration affects species detection.

To explore this, overlapping 6-, 4-, and 2-hour samples were collected, along with additional 24-hour samples in 2024, including the days immediately before and after the shorter collections. The 24-hour samples captured most of the taxa detected in the shorter intervals, whereas the shorter samples often identified only a subset, likely reflecting those present during the specific collection window. These findings highlight the trade-offs associated with sampling duration, as summarised in Table 6.1.

These advantages and limitations highlight that sampling strategy should be guided by the specific aims of the study. Longer collections may be more suitable for generating a comprehensive list of airborne taxa, which can be useful for comparing different geographical locations. In contrast, shorter collections may be better suited to detecting rarer taxa or examining how the airborne prevalence of specific species varies diurnally in response to environmental conditions.

However, determining the optimal collection length involves balancing several factors. Increasing the duration of a sample generally yields more DNA and a greater number of unique taxa, but it may also dilute the relative abundance of rare taxa and reduce temporal resolution. Airborne microbial studies have employed a wide range of sampling durations, from as short as 10 seconds [6], to 7 days [222], or even over entire seasons [154]. These

Table 6.1: Summary of the advantages and disadvantages of longer versus shorter sample collection durations.

Collection Duration	**Advantages**	**Limitations**
Long (24-hour)	• Detects the majority of prevalent taxa • Higher DNA yields for downstream processing	• May miss rare or stochastically abundant taxa • Unable to determine the timing of spore release peaks • Can demand more time from personnel if the sampler needs to be monitored throughout
Short (2–6 hours)	• Provides fine resolution on when certain taxa were present • Easier to link spore release to the environmental data available	• With a single short collection, taxa abundant in other time points will be missed • Lower DNA yields for downstream processing • Maintaining consistent short-term sampling intervals over a 24-hour period is logistically demanding, as it necessitates round-the-clock human oversight

durations are typically selected based on the target organism(s), study location, type of sampler used and the intended downstream analysis.

6.6.6 Experimental limitations

There are limitations with the 24-hour study. Given the constantly fluctuating nature of the airborne environment, numerous confounding factors may influence the results. With only a single intensive 24-hour sampling period conducted per month, and a total of just three such periods, it is not possible to confidently attribute observed differences between months to either seasonal variation or environmental factors.

6.7 Conclusion and Future Work

In conclusion the data presented here demonstrate that airborne taxa vary across hours, months and years. While longer sampling durations are effective for capturing dominant taxa, shorter intervals often reveal unique or transient taxa that may be diluted in extended collections. Moreover, shorter sampling provides greater temporal resolution, allowing for closer alignment between species abundance and environmental conditions.

Future work could include reducing sample duration to better capture short-term fluctuations in taxa, although this may be constrained by low DNA yields. Extending sampling across a broader range of months or over longer periods than a few days would also provide greater contextualisation of the results.

The variation seen in airborne taxa across samples arises because different spore types respond to distinct biological and environmental triggers, influencing their presence and abundance at particular times of the day. Consequently, the composition of airborne taxa is shaped by environmental changes in the long and short term. These dynamics make it difficult to define a single time of day that is consistently optimal for sampling, as fungal diversity fluctuates throughout the year.

Ideally, sampling strategies should be tailored to the biological characteristics and temporal patterns of the target organisms. However, the sporulation timing of many fungal species remains poorly understood and is further complicated by variable weather conditions. Moreover, because different fungal pathogens may peak in abundance at different times of day, it is unlikely that a single collection window would be suitable for detecting a broad range of taxa.

Although continuous sampling, such as daily 24-hour collections, would improve the likelihood of capturing the full diversity of airborne fungi, this approach is rarely feasible or cost-effective for long-term monitoring. Therefore, shorter, strategically timed sampling intervals represent a practical and efficient alternative for general pathogen surveillance, especially when informed by prior knowledge of local environmental conditions and organism-specific behaviour.

Chapter 7

Metagenomic Alignment and Reporting for Monitoring of Threats - Bioinformatic pipeline

7.1 Abstract

The MARMoT pipeline is presented, including the development of a pathogen-focused reference database, parameter testing and optimisation for taxonomic assignment, and a detailed description of the scripts and analytical stages involved. MARMoT was designed specifically to analyse AirSeq datasets to identify potential emergent airborne pathogens. To validate its performance, a range of datasets described elsewhere in this thesis, along with additional data, were analysed using the pipeline. The results demonstrate that MARMoT can detect a wide variety of pathogens and recover seasonal and temporal patterns consistent with known pathogen dynamics. However, while some potentially emergent high-risk species were identified, these may represent false positives. The detection of Ash dieback, a high-risk pathogen known to be present in the UK, was considered a reliable result. However, often a high level of certainty is required when alerting on pathogenic threats so further development and validation may be necessary before species-level identifications generated by MARMoT can be considered fully reliable. Nevertheless, samples shown to contain high-risk species could be used to target more in-depth analysis of the collection areas.

7.2 My Contributions

I developed the majority of the scripts used in the pipeline, except for those specified in the methods section, with the assistance of ChatGPT (GPT-5) [273]. The datasets analysed were generated from a range of experiments conducted by my lab group during my PhD. Several datasets (Tiptree, Church Farm, Maize Distance, 24-hour) are described earlier in this thesis. Other datasets (Sahara Dust, NorfolkSeq, Breeder Observation Panel) were obtained with my involvement but are introduced here for the first time and were processed by Dr Heavens. Once sequencing was complete, I performed all subsequent analyses presented in this chapter, including rebasecalling and processing through the MARMoT pipeline. I also used ChatGPT (GPT-5) to improve the grammar and flow of my writing.

7.3 Introduction

The MARMoT pipeline represents the culmination of several years of development aimed at identifying the most effective tools and parameters for the rapid and reliable detection of plant pathogens from air samples with ONT sequence data. The parameters and thresholds described in this chapter were selected based on analysis with simulated and real-world data. The pipeline filters raw reads on length, aligns them to a curated reference database of plant pathogen genomes and then filters the alignments with a lowest common ancestor (LCA) algorithm to generate taxonomic assignments. The final results are visualised through heatmaps and stacked bar charts, with a summary output highlighting high-risk pathogen species.

Effective plant disease management relies on the timely and accurate identification of pathogens in the environment. Determining their presence from air samples is essential for protecting crops against incoming diseases, as different pathogens may require distinct chemical or cultural control strategies. This process is further complicated by the potential presence of closely related non-pathogenic taxa in the air, underscoring the need for precise species-level identification to avoid unnecessary fungicide applications. In addition to monitoring common pathogens in the UK, it is also important to identify potential new threats. With a warming climate and an increasingly connected globe, new diseases are emerging that threaten UK food production [85, 265]. MARMoT was therefore developed to support the detection of both established plant pathogens and emerging diseases.

7.3.1 Pathogen databases

In order to build a comprehensive reference genome database of pathogens two different publicly available databases were utilised, PHI-base (Pathogen–Host Interactions database) and the Department for Environment, Food and Rural Affairs (DEFRA) UK Plant Health Risk Register.

PHI-base is a manually curated resource containing experimentally verified pathogenicity, virulence and effector genes from fungal, oomycete and bacterial pathogens. These pathogens infect a wide range of hosts, including animals, plants, fungi and insects. Each entry is supported by peer-reviewed literature and curated by domain experts, ensuring the reliability of the data.

The UK Plant Health Risk Register, maintained by DEFRA, categorises plant pests and pathogens using a structured risk assessment framework. The register includes information on proposed control measures, current regulations and the known distribution of each species both within the UK and globally. It also provides detailed evaluations of each organism, including likelihood of entry, potential for spread, economic and environmental impact, and the value at risk. Each of these parameters is scored and combined to generate an overall risk rating, which is then grouped into five colour-coded risk bands, shown in Figure 7.1.

These bands, ranging from low to high risk, support prioritisation of mitigation efforts and policy decisions. This categorical scheme was used in downstream visualisations to summarise read-level taxonomic assignments by associated risk level. Additional metadata fields in the risk register were also considered such as type of pest and presence in the UK.

Figure 7.1: DEFRA risk rating categories based on combined likelihood and impact scores, with scores and a representative species indicated for each category.

Including species from both PHI-base and the Plant Health Risk Register enabled the creation of a large reference database containing known pathogens with associated metadata.

7.3.2 Bioinformatic tools

MARMoT makes use of a some pre-existing metagenomic analysis tools such as *minimap2* and *LCAParse*.

Minimap2 is designed for mapping and aligning long-read sequencing data generated by technologies such as ONT and PacBio, as well as assembled genomes [217]. The software is designed to handle query sequences ranging from a few kb to around 100 mb, even at error rates of up to 15%. Results are typically generated in either PAF or SAM format, which can then be used for downstream processing [217]. *Minimap2* identifies matches using minimizers, which are short representative subsequences. Minimizers are stored in a hash table for rapid lookup, enabling repetitive regions to be skipped reducing unnecessary computation. This approach allows *minimap2* to achieve high accuracy while being significantly faster than many specialised aligners. Its combination of speed, flexibility, and precision has made it widely used in large-scale genomic and metagenomic analyses.

LCAParse is a command-line tool designed by Dr Richard Leggett for assigning taxonomic classifications to sequencing reads based on alignment results. It uses NCBI taxonomy data and accession-to-taxon mapping files to determine the LCA of the top-scoring hits for each read. Though a separate tool, the LCA assignment code is taken from MARTi. *LCAParse* supports multiple input formats, including BLAST outputs and PAF

files. Users can also apply filters such as minimum identity, coverage, or species-level limits. By parsing alignment results in this way, reads without a clear species-level assignment can still be retained in the analysis at higher taxonomic ranks within metagenomic studies.

7.3.3 Existing tools for metagenomic pathogen detection

A range of tools exist to aid in the alignment of metagenomic reads to reference databases, including *Kraken2*, *Centrifuge*, MARTi and *MEGAN* [166, 196, 281, 395]. There are also specialised tools for the identification of specific plant pathogens, such as MARPLE, which utilises SNPs to differentiate between strains of complex fungal pathogens [301]. However, MARPLE is not purely a software tool, but rather a combination of targeted laboratory methods and custom bioinformatic analysis.

Despite the range of bioinformatics pipelines available, none are specifically tailored to the detection of plant pathogens from metagenomic data. This is the gap that MARMoT addresses. MARMoT is a command-line tool tailored for ONT sequencing data from air-collected samples, aimed at targeted pathogen monitoring. The tool offers flexibility: in addition to scripts for generating a reference database from PHI-base and the Plant Health Risk Register, users can employ any custom reference database - although in such cases, the output will be more limited in terms of the available information on the risk level of identified pathogens.

This chapter presents the parameter optimisation for MARMoT, alongside a detailed description of its scripts and processing stages. Following optimisation, the pipeline was applied to a large number of airborne metagenomic datasets, including those collected during this PhD and others provided by members of the Leggett group. Analysis of these datasets identified three pathogens categorised as red risk. However, there remains uncertainty as to whether these alignments represent genuine detections or artefacts arising from contaminated or low-quality reference genomes.

7.4 Methods

MARMoT is a modular pipeline designed to detect potential plant pathogens in nanopore sequencing data. It processes raw barcoded sequencing reads through a series of steps including quality filtering, taxonomic alignment, and summary reporting, with optional visualisations. The pipeline consists of two main components: the creation of a reference database and the processing of input sequence data.

MARMoT was developed to support the systematic and rapid identification of plant pathogens from airborne sequencing data, with a long-term goal of deployment in bio-surveillance settings. In this project, however, the pipeline was primarily used as a data mining framework to analyse sequencing datasets generated during the study, enabling retrospective identification of airborne emergent plant pathogens.

7.4.1 Construction of the custom pathogen reference database

To enable taxonomic assignment via sequence alignment, a custom reference database of pathogenic species was constructed. This database integrates species from two publicly available sources. PHI-base, (`https://github.com/PHI-base/data`) and the UK Plant Health Risk Register (`https://planthealthportal.defra.gov.uk/pests-and-diseases/`

`uk-plant-health-risk-register`). Both datasets were downloaded as CSV files and processed using a Python-based workflow developed for this study, adapted from an original Bash script by Dr Michael Giolai [124]. Figure 7.2 illustrates the different steps involved.

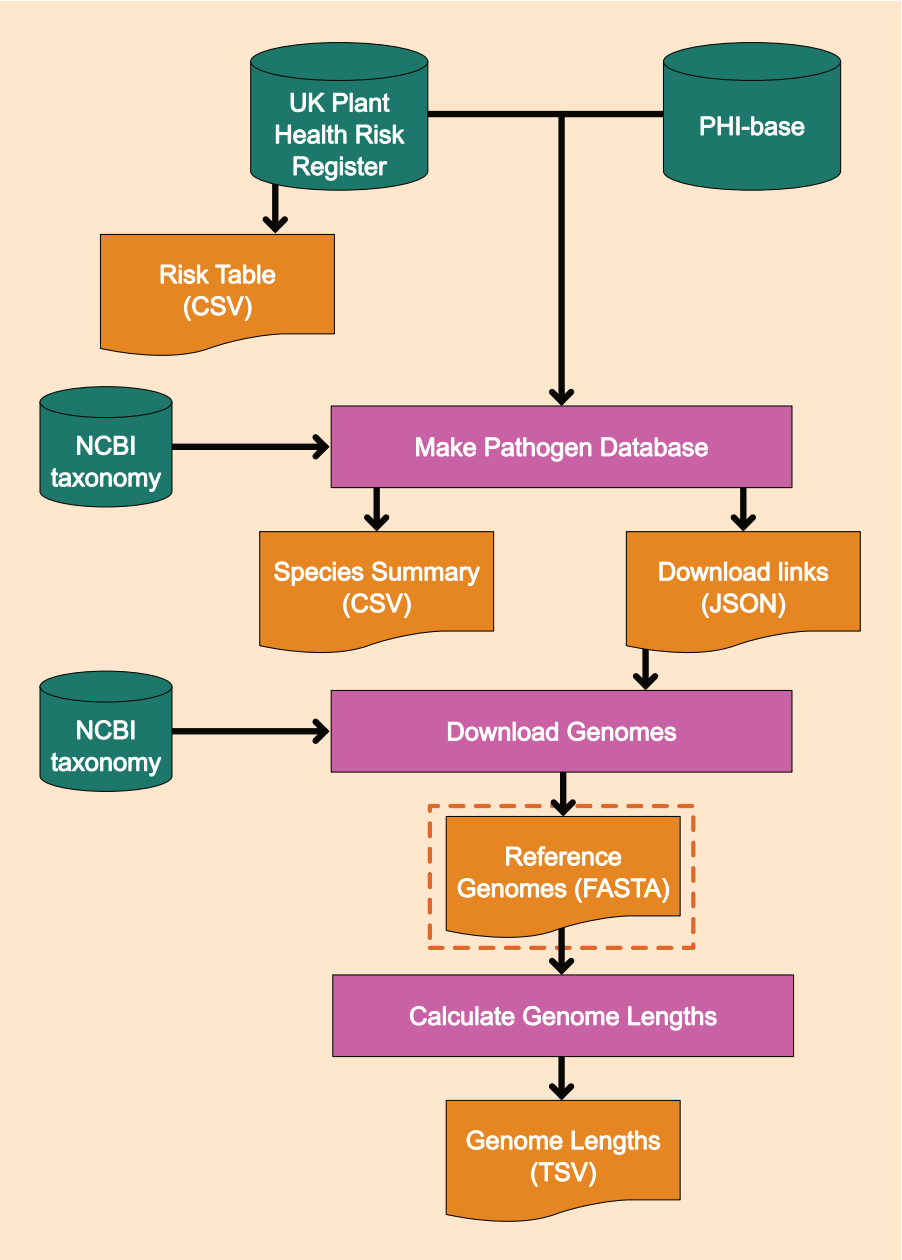


Figure 7.2: Bioinformatic workflow showing the construction of the custom pathogen reference database for MARMoT

The reference database is built using the script `build_reference_database.sh`, available in the GitHub repository (`https://github.com/Mia-FGB/MARMOT_ref_db`). This script coordinates a series of Python programs within a dedicated Conda environment, which can be created from the included `environment.yml` file. It runs locally and requires an internet connection to download genome assemblies and metadata.

The first step is `generate_risk_table.py`, which creates a simplified table of pathogen names and associated risk metrics from the risk register input. This table is later used to generate graphs and summarised outputs.

Additionally, both the PHI-base and risk register CSV files are passed to `Make_Pathogen_Database.py`, which cleans and filters the datasets. Insect, mite, nematode and plant pests are first removed from the risk register, after which the two dataframes are merged and deduplicated to retain only unique species. The `NCBItaxa` module from `ete3` is then used to retrieve the taxonomic ID (taxaID) for each species. The script also downloads the latest assembly summary files from both RefSeq and GenBank and filters them to retain only entries matching the target taxa. For each species, a single representative genome is selected using a defined priority scheme: reference genomes are preferred, followed by complete genomes, chromosomes, scaffolds and contigs. If multiple genomes exist within the same priority level, the largest and most recently released assembly is chosen. The script produces a JSON file with download links for the selected genomes, along with a CSV summary of the included taxa.

Next, `download.py` uses the JSON file to download the selected genomes. Each file is checked against its MD5 checksum to verify data integrity, and the sequence headers are modified to include both the NCBI taxaID and species name, ensuring compatibility with downstream tools. These individual genome files are then concatenated into a single FASTA file. The resulting reference file contains one genome per pathogen and is used for taxonomic alignment with *minimap2* later in the MARMoT pipeline.

Finally, the script `genome_lengths_from_fasta.py` calculates the length of each genome in the generated FASTA file. This information is used in later stages of the pipeline to calculate genome coverage of the taxa. All scripts log their operations to dedicated files, and a user-defined date tag applied during execution provides a versioning mechanism to track changes in database content over time.

7.4.2 Taxonomic alignment and assignment

The MARMoT pipeline is an automated workflow that begins with a configuration file (Table 7.1) and a submission script (`submit_with_config.sh`). Once launched, each barcode is processed independently: sequence data are filtered by length and aligned to a reference database using *minimap2*, generating a PAF output file. The resulting alignments are processed with LCAParse's LCA algorithm to assign taxonomy, and average genome coverage is calculated for each taxon. Finally, the outputs from all barcodes are combined and summarised in R, generating visualisations by genus and plant health risk register status.

Figure 7.3 provides an overview of the key scripts and their functions, whereas Figure 7.4 illustrates the overall bioinformatics workflow, showing the logical flow of data through submission, per-barcode processing, and summarisation. All scripts used in the pipeline are available on GitHub: `https://github.com/Mia-FGB/MARMOT`.

7.4.2.1 Equations used in the pipeline

$$\text{Identity} \;=\; 100 \times \frac{\text{Number of matching bases}}{\text{Alignment block length}} \tag{7.1}$$

$$\text{Query coverage} \;=\; 100 \times \frac{\text{Alignment block length}}{\text{Read length}} \tag{7.2}$$

$$\text{Taxon coverage (\%)} \;=\; 100 \times \frac{\text{Mapped bases}}{\text{Reference genome or alignment length}} \tag{7.3}$$

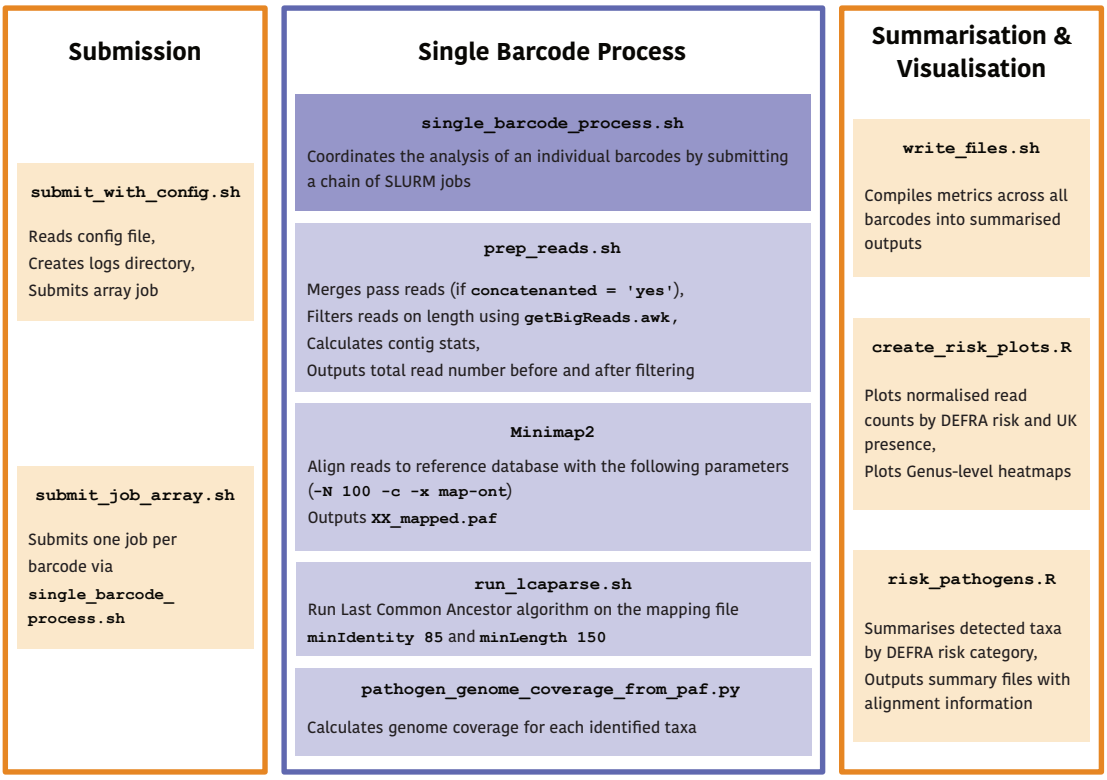


Figure 7.3: Overview of the bioinformatics pipeline for pathogen detection and risk assessment from ONT sequencing data. The workflow comprises three main stages: submission and job control; single barcode processing, which includes read preparation, alignment, taxonomic classification, and genome coverage estimation; and summarisation and visualisation, where combined results are compiled and used to generate diagnostic plots and risk reports. Key tools and scripts used at each stage are indicated, along with relevant parameters.

The pipeline is initiated with `submit_with_config.sh`, which reads a user-defined configuration file and passes its parameters to `submit_job_array.sh` (parameters are listed in Table 7.1). This script sets up and submits a SLURM job array, with each job processing one barcode. In addition, `submit_with_config.sh` creates a dedicated log directory for storing output and error logs. Once all barcode jobs have completed successfully, it triggers the generation of summary files and visualisations via `write_files.sh`, `create_risk_plots.R` and `risk_pathogens.R`

The `submit_job_array.sh` script automatically launches `single_barcode_process.sh` for each barcode specified in the configuration file. For every barcode, a dedicated output directory (e.g. `barcodeXX`) is created to store results and logs. Once initiated, the analysis proceeds without further user input through four sequential stages: (1) read preparation and filtering via `prep_reads.sh`, (2) alignment of filtered reads to the reference database using *minimap2*, (3) taxonomic parsing with `run_lcaparse.sh`, and (4) calculation of genome coverage statistics using `pathogen_genome_coverage_from_paf.py`. Each stage is submitted as a separate SLURM job, with job dependencies managing the correct order of execution, and standard output and error logs recorded in a `logs` subdirectory within the barcode folder.

The first stage of barcode processing is performed by `prep_reads.sh`. This script searches for input reads in either a `fastq` or `fastq_pass` directory within the specified location. If the reads are not already concatenated and the `concatenated` argument is set to no, all files for the barcode are combined into a single FASTQ file, with compressed files

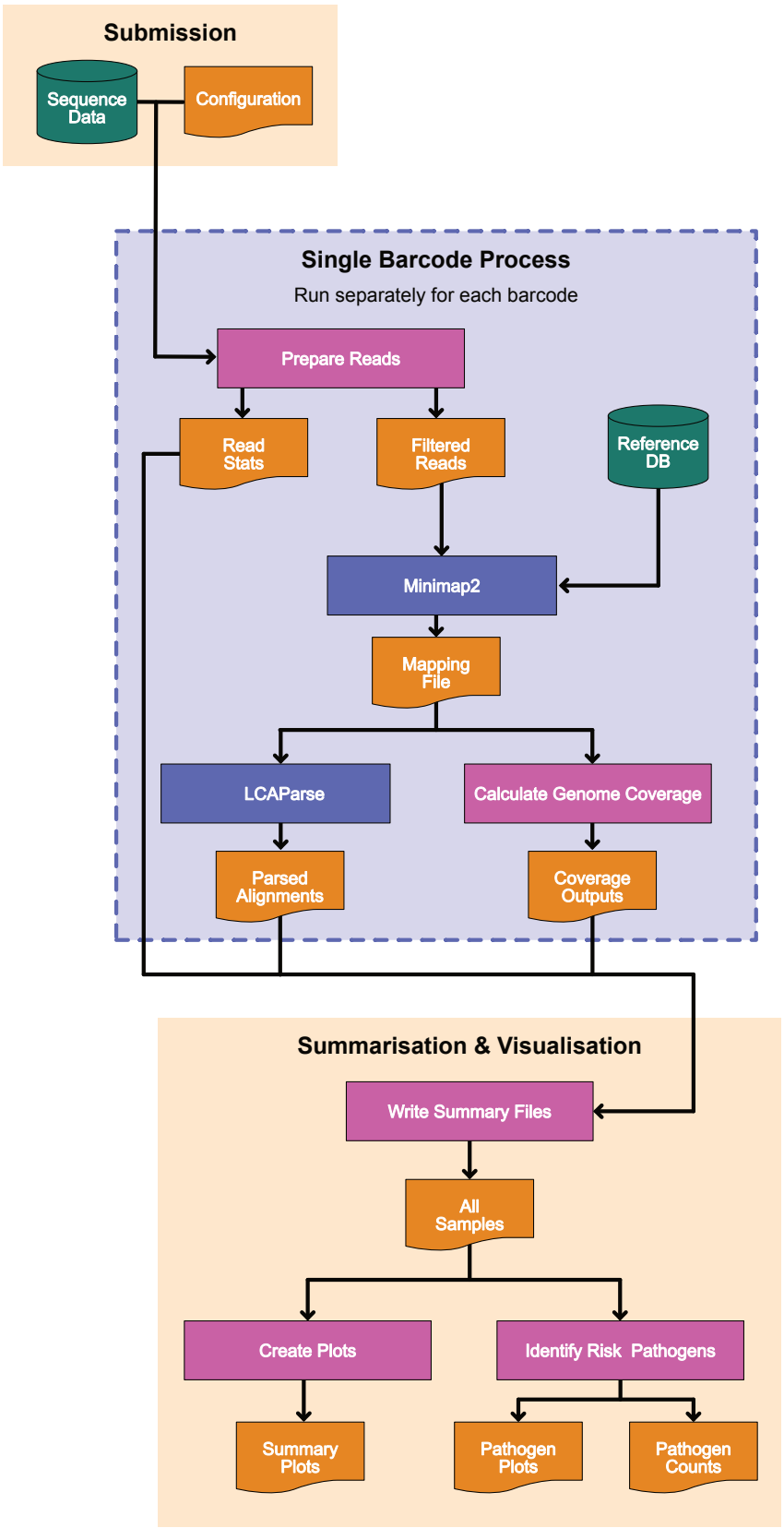


Figure 7.4: Pipeline schematic showing data submission, single barcode processing, and downstream summarisation and visualisation steps. Sequence data and configuration files are submitted, after which each barcode is processed independently through read preparation, mapping, parsing and genome coverage estimation. Outputs are then combined across all samples for summarisation, including pathogen risk assessment and generation of summary plots.

Table 7.1: Explanation of configuration file parameters used to run the MARMoT pipeline.

Parameter	Explanation	Example
`sample`	Name of the sample being processed. Used for labelling and organisation.	`CF_2023`
`location`	Path to raw / rebasecalled reads. Should be the level above `fastq`.	`/path/to/raw/reads`
`filter_length`	Reads shorter than this length will not be included in the analysis.	`300`
`reference_database`	Path to the reference database generated by `build_reference_database.sh`.	`path/to/reference_database.fa`
`scratch_dir`	A directory for temporary large files to be stored. Can be the same as `output_dir`.	`/scratch/pipeline_output/`
`output_dir`	Directory for output files.	`/scratch/pipeline_output/`
`barcode_list`	List of barcodes to process. Can be a sequence or specific list.	`$(seq -w 01 05)` or `("11" "02" "05" "08")`
`concatenated`	Set to `"yes"` if the reads are already concatenated into one file, else `"no"`.	`yes`
`contig_stats`	Set to `"yes"` to calculate contig stats for the barcode, else `"no"`.	`yes`
`genome_lengths_file`	File with TaxaIDs and genome lengths, generated by `build_reference_database.sh`.	`/path/to/genome_lengths.tsv`
`risk_table_file`	File with DEFRA risk information, generated by `generate_risk_table.py` after creating database.	`/path/to/risk_table.csv`
`barcode_labels`	Tab delimited file with barcode number and SampleID to be used as labels in graph creation.	`/path/to/barcode_labels.tsv`

unzipped as required. When a `fastq_fail` directory is present, the number of fail reads is also recorded. The concatenated reads are then length-filtered using `getBigReads.awk`, with the threshold defined in the config file (script provided by Dr Sam Martin). If the `contig_stats` argument is set to yes, `get_contig_stats.pl` is run on the unfiltered data to generate contig-level metrics, and it is always applied to the length-filtered reads. Read counts before and after filtering are written to a summary file (`<barcode>_read_no.tsv`) within the barcode directory.

The length filtered reads are then aligned to the reference database with *minimap2*, producing a PAF output. Alignment is performed using `-N 100 -c -x map-ont`, parameters chosen through experimental evaluation to optimise accuracy and filtering (see section 7.4.5).

Following alignment, the PAF file is provided as input to LCAParse (v0.9.24, `https://github.com/richardmleggett/LCAParse`), which assigns each read to the lowest common ancestor in the taxonomic tree based on the alignment results. This step uses the NCBI taxonomy database and a custom accession map to resolve taxonomic identities. Classification is run with the parameters `-minidentity 85` and `-minlength 150`, using PAF-formatted input. These thresholds were selected based on validation experiments

described in sections 7.4.5 and 7.5.1.

Genome coverage is estimated with a custom Python script (`pathogen_genome_coverage_from_paf.py`). For each barcode, the script parses the PAF file, computing percent identity (Eq.(7.1)) and query coverage (qcov) (Eq.(7.2)) for every alignment. Alignments with identity and query coverage $\geq 80\%$ are retained. For each taxon, the script sums the lengths (`read_len`) of all passing read–taxon pairs to obtain total mapped bases, and coverage is reported as in Eq. (7.3).

The genome coverage script also tracks reads that either pass the filters for multiple taxa or fail to meet the coverage/identity thresholds. For each barcode, the outputs include per-taxon summaries (mapped bases, genome length, coverage percentage, number of contributing reads, and their taxaIDs) together with separate tables listing excluded reads and multi-taxon reads.

Once genome coverage calculations for each barcode have successfully completed, the final summarisation stage begins. This is initiated by the `write_files.sh` script, which is submitted with a dependency on the successful completion of the barcode array jobs. This script collates key metrics across all barcodes, including read retention, fail read counts, *LCAParse* assignments, genome coverage, and total read numbers. It appends each barcode's data to a set of combined summary files, ensuring they are structured consistently for downstream analysis saved in the specified output directory.

7.4.3 Outputs and graphical visualisations

Following taxonomic assignment, the `create_risk_plots.R` and `risk_pathogens.R` scripts are executed. Both scripts perform downstream analysis and visualisation by combining the summarised pipeline outputs with external metadata, including barcode labels and a DEFRA-defined pathogen risk table.

The `create_risk_plots.R` script generates a suite of diagnostic plots, such as stacked bar charts categorised by mitigated risk level, faceted risk plots based on UK presence, per-genus heatmaps of abundance and HP100k and genome coverage bar charts for the nine fungal genera of interest, which were introduced in Chapter 5 (*Blumeria, Claviceps, Fusarium, Magnaporthe, Parastagonospora, Puccinia, Pyrenophora, Ustilago* and *Zymoseptoria*).

`risk_pathogens.R` outputs pathogen-specific read and coverage summaries for each risk level, lists of high-risk read IDs for potential downstream analysis, and an overview of the different pathogens identified, with all results saved in the designated output directory.

The complete set of outputs from all the scripts within the pipeline are detailed in Table 7.2.

7.4.4 Implementation notes

The MARMoT pipeline is designed to run in a Linux environment using SLURM for job scheduling. Most scripts are written in Bash or Python and do not rely on specific Python packages beyond those available in standard distributions. The reference alignment and parsing steps use established tools such as *minimap2* and *LCAParse*, which are executed via SLURM job submissions with resource specifications defined per step.

The visualisation and summary steps require a dedicated R environment. An `environment.yml` file is provided in the GitHub repository to facilitate environment setup via `conda`. This

Table 7.2: Summary of pipeline outputs including visualisations, where XX indicates the barcode number

Generated by & Scope	**Output name**	**Description**
`prep_reads.sh` Barcode	`XX_barcode_percent_retained.txt`	Percentage of reads retained after length filtering
	`XX_contig_stats_filtered_300.txt`	Contig stats of length filtered reads
	`XX_read_no.txt`	Total and filtered read counts
Minimap2 Barcode	`XX_mapped.paf`	Mapping results against reference database
`run_lcaparse.sh` Barcode	`XX_lcaparse_perread.txt`	Taxonomic assignment for each read
	`XX_lcaparse_summary.txt`	Read counts and taxa summary
`pathogen_genome_coverage.py` Barcode	`XX_genome_coverage.txt`	Estimated genome coverage for taxa
	`XX_coverage_excluded_reads.txt`	Reads failing identity or coverage thresholds
	`XX_coverage_multi_taxa_reads.txt`	Reads mapping to multiple taxa
`write_files.sh` Experiment	`genome_coverage_all.txt`	Collated genome coverage across all barcodes
	`read_numbers.tsv`	Total and filtered read counts per barcode
	`lcaparse_summary.txt`	Merged taxon summary across barcodes
	`lcaparse_perread.txt`	Merged taxon assignments per read
	`no_fail_reads.txt`	Notes if fail reads were not present
	`percent_reads_retained_length_filter.txt`	Retained read percentage
`create_risk_plots.R` Experiment	`Graphs/heatmap/*`	Heatmaps of genus abundance (normalised read counts)
	`Graphs/pathogen_graphs/*`	Bar plots and facets of target pathogens (by genus)
	`Graphs/defra_risk/*`	Plots of read distribution coloured by DEFRA risk level
	`genus_summary.tsv`	Normalised genus-level read counts
`risk_pathogens.R` Experiment	`<Level>Risk_ReadIDs_all.tsv`	Read IDs assigned to high-risk (red / orange) species
	`<Level>Risk_ReadIDs_noWidespread.tsv`	Red / Orange risk reads excluding widespread UK species
	`<Level>_Risk_Species_Summary.tsv`	Alignment metrics for the different pathogens, file per risk category
	`Taxon_Presence_Summary.tsv`	Summarised data for each pathogen

file includes the necessary R packages (such as `tidyverse`, `fs`, and `svglite`) for running both R scripts.

7.4.5 Validation and testing of the pipeline

7.4.5.1 Optimisation of *minimap2* parameters for accurate taxonomic assignment

To determine optimal parameters for *minimap2* in the context of pathogen detection, a representative dataset from the Church Farm 2023 sequencing run, also processed with MARTi, was used. Alignments were first performed with the `-x map-ont` preset, tailored for ONT data.

To enhance alignment accuracy and output completeness, two additional flags were tested alongside the standard preset (`-N 100 -c -x map-ont`). The `-c` flag outputs CIGAR strings in PAF format, preventing underestimation of matching bases, while the `-N 100` flag retains up to 100 secondary alignments per read, avoiding the default behaviour of reporting only a single (randomly chosen) alignment when multiple valid matches are present.

Analysis of the PAF output from *minimap2* with both parameter sets included inspection of reported matching bases, mapping length, and the `dv:f` or `de:f` tags. The `dv:f` tag approximates per-base sequence divergence, while `de:f` generated with the `-c` flag represents gap-compressed divergence incorporating mismatches, gaps, and other non-matching bases in the aligned region. Sequence identity was derived from divergence using Eq.(7.1), and percentage of matching bases was calculated from the PAF output using Eq.(7.2).

With both parameter sets, sequence identity and the percentage of matching bases were strongly correlated. Inclusion of the additional flags increased the reported number of matching bases, and the two measures scaled linearly with a stronger relationship, indicative of improved calibration. Detailed results of these comparisons are presented in section 7.5.1.1.

On this basis, *minimap2* was implemented in the MARMoT pipeline with the parameters `-N 100 -c -x map-ont`. All alignments were retained, and taxonomic assignments were subsequently refined with LCAParse, which applies the same LCA algorithm as used in MARTi.

7.4.5.2 *LCAParse* parameter testing

LCAParse parameter values within the pipeline were selected following optimisation tests using simulated ONT reads provided by Dr Ned Peel.

The simulated dataset consisted of 100,000 reads derived from 17 species, of which a subset were present in the emergent pathogen reference database used for alignment (Table 7.3). Reads were aligned to the reference database using *minimap2* with the parameters selected above, and parsed with *LCAParse* using six different combinations of minimum identity and alignment length thresholds (Table 7.4).

Parsed output files were assessed for taxonomic assignment accuracy, with reads classified as either unassigned, correctly assigned, or incorrectly assigned at various taxonomic levels, see results in section 7.5.1.2. From this the false positive, false negative and true negative rates could be calculated. Of the 100,000 simulated reads, only 13,000 corresponded

Table 7.3: Composition of the simulated nanopore dataset used for *LCAParse* parameter testing. Species present in the reference database are indicated in grey.

Species	**Read count**	**Read %**
Akkermansia muciniphila	4000	4
Bacteroides fragilis	8000	8
Bifidobacterium adolescentis	16000	16
Candida albicans	1500	1.5
Clostridioides difficile	4000	4
Clostridium perfringens	500	0.5
Enterococcus faecalis	1000	1
Escherichia coli	2500	2.5
Faecalibacterium prausnitzii	14000	14
Fusobacterium nucleatum	6000	6
Lactobacillus fermentum	6000	6
Methanobrevibacter smithii	3000	3
Prevotella corporis	8000	8
Roseburia hominis	12000	12
Saccharomyces cerevisiae	1500	1.5
Salmonella enterica	2000	2
Veillonella rogosae	10000	10

Table 7.4: Parameter combinations tested for *LCAParse* parsing of simulated read data.

Name	**minIdentity (%)**	**minLength (bp)**
MinID 80, MinLen 100	80	100
MinID 85, MinLen 100	85	100
MinID 90, MinLen 100	90	100
MinID 80, MinLen 150	80	150
MinID 85, MinLen 150	85	150
MinID 90, MinLen 150	90	150

to species present in the reference database.

Based on this analysis *LCAParse* was run with the thresholds set to `-minIdentity 85` and `-minLength 150` within the MARMot pipeline.

7.4.6 Using MARMoT to mine existing airborne datasets

Once complete, MARMoT was run on a number of airborne sequencing datasets collected throughout this PhD and by other members of the lab group. Data from samples collected outside of the scope of this thesis include Sahara dust, Breeder Observation Panel (BOP) and NorfolkSeq collections. The experiment, date, sampler used, location and number of samples from each dataset are shown in Table 7.5.

The Sahara dust samples were collected in 2021 for 1 hour with the Coriolis μ (300 L/min) and Cub samplers (200 L/min), outside of the Earlham institute following the appearance of a dust cloud of Saharan origin [83].

The BOP samples were collected in May, June and July 2024 from wheat fields across East Anglia (Suffolk, Cambridgeshire, Lincolnshire and Oxfordshire) where breeders evaluate common observation panel lines. Collections were coordinated across sites: Limagrain, NPZ, Elsoms and DSV were sampled on day 1, and KWS, Syngenta and RAGT on day 2, with duplicate 2-hour air collections with the Cub samplers (200 L/min) between 09:00

and 11:00.

NorfolkSeq 1-hour Cub samples were collected once a season from April '24 until Jan '25 at different sites across East Anglia (Thetford Forest, Norwich City Centre, Kessingland Beach, Carlton Marshes, Hickling Broad, Church Farm, Brancaster Beach and Foxley Wood).

For all the above samples DNA was extracted following the standard protocol described in Chapter 4, before undergoing ONT sequencing.

An up-to-date reference database was generated using the `build_reference_database.sh` script, as described in section 7.4.1 for the analysis of this data. The input sources were PHI-base (v4.17) and the DEFRA risk register (April 2025). The final database contained 572 genomes: 333 complete genomes, 129 scaffolds, 63 contigs, and 47 chromosomes. Species names, taxonomic identifiers, and accession numbers are provided in Table A.3 in the Appendix.

Once all the data had been processed with MARMoT, the summary files were imported into a new R script to generate visualisations across datasets and to identify commonly detected absent or red- and orange- risk species. Datasets were filtered to remove taxa present at fewer than 10 reads per 100,000. For comparative analyses, the outputs were grouped according to sampling design, as shown in the experiment column of Table 7.5. The Dust and Strawberry datasets were each treated as single groups. The Church Farm samples were divided by year, with two periods in 2023 (Oct–Apr and May–Aug) and two in 2024 (Feb–Jun and Jul–Sep). At Church Farm, additional collections included the 24-hour and maize datasets, which were grouped by collection month and by week, respectively. The BOP and NorfolkSeq datasets were subdivided by collection date, while the sampler comparison experiment was grouped by collection location (Church Farm and NHM)

Genus-level HP100k abundances from the filtered pipeline outputs were used for diversity calculations. Richness was defined as the number of genera detected per experiment, and Shannon diversity was computed from the same abundances. Heatmaps were generated for the nine target genera and for species classified as absent, or as red or orange risk in the DEFRA risk register. In addition, bar charts were produced to compare the proportion of species at different risk levels across experiments, facetted by known presence.

7.5 Results

7.5.1 Validation of the pipeline

7.5.1.1 *minimap2* parameter choice

In order to select the *minimap2* parameters two different alignments were run on the same data and reference database with different flags.

Comparison of the PAF outputs with different parameters demonstrates that the addition of `-N 100 -c` yields a sequence identity measure that scales linearly with the proportion of matching bases (Figure 7.5). This is consistent with the *minimap2* documentation, which notes that the `-c` option performs base-level alignment via CIGAR strings, providing more precise identity estimates albeit at increased computational cost [218].

Table 7.5: Summary of sampling experiments used for MARMoT analysis, showing collection date ranges, sampler used, location and total sample number.

Project	**Chap.**	**Experiment**	**Date**	**Sampler**	**Location**	**Sample No.**
Sahara Dust	N/A	Dust (May '21)	20/05/21 – 21/05/21	Coriolis µ & Cub	Earlham Institute, Norwich	5
Tiptree	3	Strawberry ('22)	10/02/22 – 11/08/22	Coriolis µ	Colchester, Essex	48
Regular Church Farm '23	5	Church Farm (Oct '22–Apr '23)	03/10/22 – 30/04/23	Cub	Church Farm, Norwich	12
		Church Farm (May–Sep '23)	01/05/23 – 17/08/23			22
Regular Church Farm '24	5	Church Farm (Feb–Jun '24)	07/02/24 – 30/06/24	Cub	Church Farm, Norwich	32
		Church Farm (Jul–Sep '24)	01/07/24 – 27/09/24			24
24hr collections	6	24h (Aug '23)	31/07/23 – 09/08/23	Cub	Church Farm, Norwich	13
		24h (May '24)	13–16/05/24			42
		24h (Jun '24)	17–20/06/24			42
Breeder Observation Panel	N/A	BOP (May '24)	08 – 09/05/24	Cub	Suffolk, Cambridgeshire, Lincolnshire & Oxfordshire	14
		BOP (Jun '24)	12 – 13/06/24			14
		BOP (Jul '24)	03 – 04/07/24			14
Sampler Comparison	4	Church Farm Sampler Comparison	09/06/22	All samplers	Church Farm, Norwich	20
		NHM Sampler Comparison	12/06/22		Natural History Museum, London	20
Maize distance	4	Maize (Wk 1 Aug '24)	19/08/24 – 22/08/24	Cub	Church Farm, Norwich	63
		Maize (Wk 2 Aug '24)	27/08/24 – 29/08/24			73
NorfolkSeq	N/A	NorfolkSeq (Apr '24)	14/04/25	Cub	East Anglia	14
		NorfolkSeq (Jul '24)	25/07/24			14
		NorfolkSeq (Oct '24)	17/10/24			14
		NorfolkSeq (Jan '25)	16/01/25			14

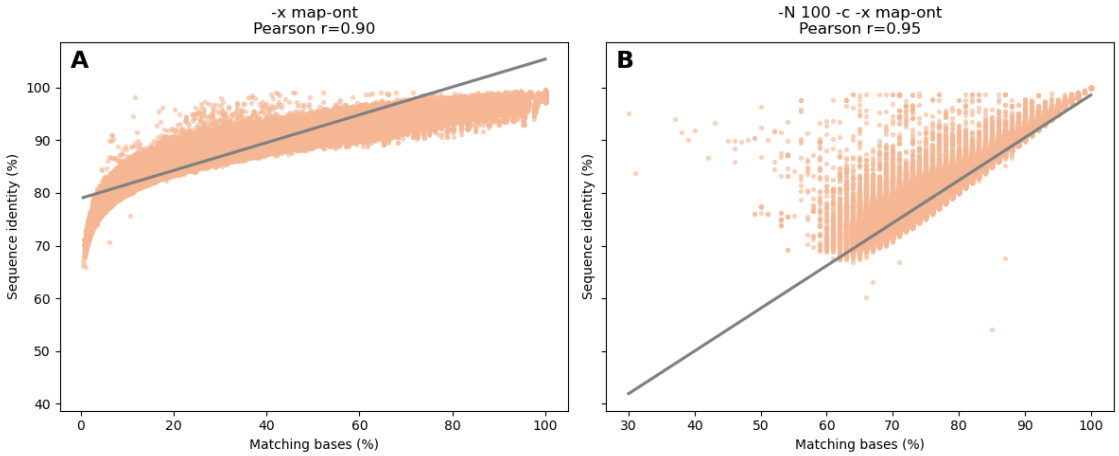


Figure 7.5: Scatter plots showing the relationship between sequence identity (%) and matching bases (%), regression line is shown in grey and Pearson correlation co-efficient is labelled above the graph. A) Data from `-x map-ont`, B) from `-N 100 -c -x map-ont`.

7.5.1.2 *LCAParse* parameter testing

From the 13,000 reads derived from taxa with a reference genome present in the database, the proportions of correct, incorrect, and unassigned classifications are shown in Figure 7.6, with exact values provided in Table 7.6.

At a `minIdentity` threshold of 90%, the majority of reads remain unassigned (61.38%). Reducing this threshold to 85% dramatically decreases the proportion of unassigned reads to 2.66%, while maintaining a low rate of incorrect classifications (0.5%, 6 reads). Further reducing the threshold to 80% increases the number of incorrect assignments substantially (114–123 reads) and shifts many assignments to higher taxonomic ranks (genus or kingdom) rather than species. A `minIdentity` of 85% therefore provides an optimal balance, maximising correct species-level assignments (approximately 200 more than at 80%) while keeping incorrect classifications to a minimum.

Adjusting the minimum length parameter between 100 bp and 150 bp has a more modest effect. Increasing the threshold to 150 bp slightly raises the number of correct species assignments and reduces incorrect classifications, but results in fewer total correct assignments and more unassigned reads. Given the priority of maximising accurate species-level classifications, this trade-off was considered acceptable.

Accordingly, *LCAParse* was run within the pipeline with `-minIdentity 85` and `-minLength 150`.

Table 7.6: Table of the number of reads correctly assigned, correctly assigned to species, unassigned and incorrect

Parameters	**Total correctly assigned**	**Correct to species**	**Unassigned**	**Incorrect**
MinID 80, MinLen 100	12,733	10,557	221	123
MinID 80, MinLen 150	12,715	10,604	221	114
MinID 85, MinLen 100	12,581	10,867	347	63
MinID 85, MinLen 150	12,576	10,899	346	59
MinID 90, MinLen 100	4984	4850	7979	6
MinID 90, MinLen 150	4984	4858	7979	6

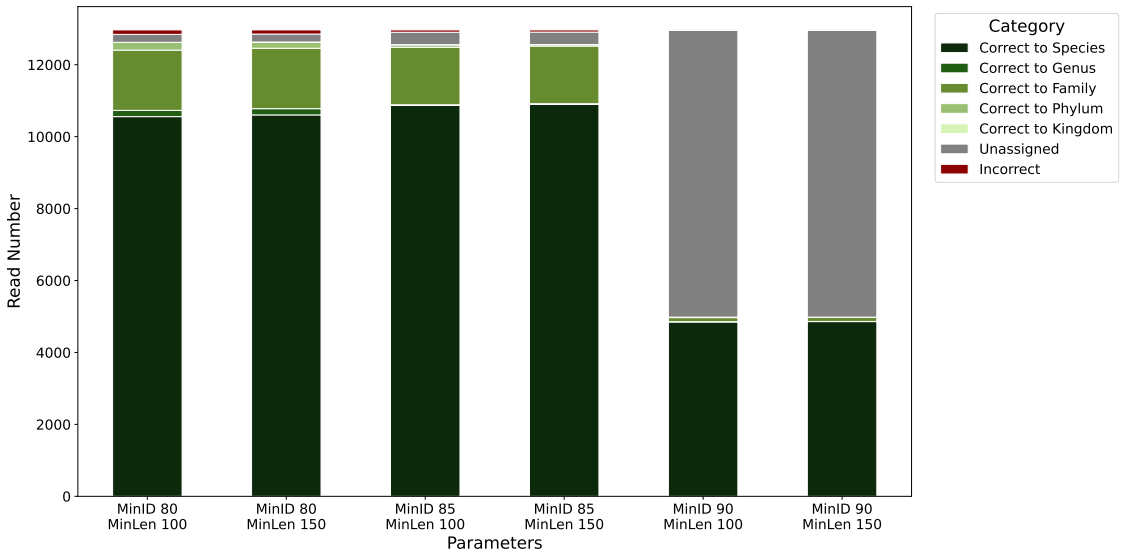


Figure 7.6: Stacked bar chart showing the number of reads assigned by *LCAParse* under different parameter settings. Reads are classified as correct (with the highest resolved taxonomic rank indicated), incorrect, or unassigned. Abbreviations, MinID = Minimum Identity(%), MinLen = Minimum Length (bp)

7.5.2 Example outputs from regular Church Farm 2024

To illustrate the outputs generated by MARMoT, results from one dataset (Regular Church Farm '24) are shown in Figure 7.7. These are the same data presented in Chapter 5, here re-processed with the MARMoT pipeline. Example graphical outputs are shown, with *Fusarium* selected from the nine pathogenic genera to demonstrate HP100k and genome coverage plots. A full list of outputs is provided in Table 7.2.

Plot A is a heatmap showing the genus-level taxonomic composition across the collection period. The five most abundant genera were *Plasmopara* an oomycete, and four fungi: *Alternaria*, *Botrytis*, *Ustilago* and *Phytopthora*. Temporal variation in abundance was evident, with some taxa appearing sporadically and others persisting across multiple timepoints.

Plots B and C focus on *Fusarium*, showing both hits per 100k reads and genome coverage over time. The relative abundance of *Fusarium* generally increased from May to August, with the highest values recorded on 4 Apr (57.38 HP100k $\pm$ 4.36), 14 Aug (49.92 HP100k $\pm$ 11.49) and 10 Sep (54.99 HP100k $\pm$ 19.22). Airborne abundance did not appear to correlate with genome coverage, which on 11 June reached a mean of 0.0030$\times$ ($\pm$ 0.00039). Across the dataset, *Fusarium* genome coverage remained low, with a mean of 0.00112$\times$ ($\pm$ 2.75 $\times$ 10^{-5}).

Plot D shows the distribution of detections according to the DEFRA risk framework across the sampling period. Blue risk taxa dominated at all timepoints, accounting for the majority of hits per 100k reads (mean 126.42 $\pm$ 14.86), followed by orange (53.68 $\pm$ 7.52), green (31.42 $\pm$ 3.00), yellow (30.56 $\pm$ 3.11) and red (11.41 $\pm$ 0.95). Total detections fluctuated over the season but generally increased from Feb to Sep, with notable peaks in several collections between Jul and Sep when overall abundance exceeded 6000 HP100k. The relative contributions of red and orange risk taxa also varied between samples, with red taxa reaching 172.99 HP100k (mean 28.83 $\pm$ 11.92) on 11 Jun and orange taxa peaking at 1009.34 HP100k (mean 252.34 $\pm$ 118.80) on 24 Jul. These results indicate that while

the majority of airborne detections were low-risk taxa, sporadic incursions of higher-risk groups were also recorded.

Together, these outputs show how the MARMoT pipeline summarises community composition, highlights pathogens of interest, and categorises taxa by risk. This example illustrates the information generated per dataset, which is integrated across all samples in the following section.

7.5.3 Diversity of detected genera

Across the experiments, sequencing depth showed little influence on diversity when based on HP100k-normalised data (Figure 7.8). Genus richness, the number of unique genera, was not correlated with total reads (Pearson's r = 0.10, 95% CI -0.36 to 0.52, p = 0.67, n = 20). Similarly, Shannon diversity, which reflects both abundance and evenness, showed no significant association with sequencing depth (Pearson's r = 0.24, 95% CI -0.22 to 0.62, p = 0.30, n = 20). These results indicate that, after filtering spurious low-abundance assignments and normalising by sequencing effort, additional reads did not substantially increase observed richness or alter community evenness.

Cross-project composition from the Bray–Curtis ordination provides a complementary view of between-group differences (Figure 7.8C). The two dimensional solution had low stress, approximately 0.038, indicating a reliable projection. Colouring points by project highlights both similarity and difference among experiments processed in similar ways. Samples from BOP, Sampler Comparison, the Maize collections, and the NorfolkSeq experiments cluster, whereas the Church Farm (Oct '22–Apr '23) sample does not cluster with the other Church Farm experiments, and the 24-hour experiments are dispersed.

7.5.4 Nine target genera per experiment

A heatmap was generated to summarise the nine target genera across all datasets (Figure 7.9). Although a broader range of taxa were detected, focusing on these key genera provides a clearer overview of agriculturally relevant pathogens.

Most genera, including *Fusarium*, *Blumeria*, *Pyrenophora*, *Parastagonospora*, *Zymoseptoria*, and *Ustilago*, were detected across all experiments. *Claviceps* and *Puccinia* were absent only from Church Farm Oct '22–Apr '23, while *Magnaporthe* was not detected in any dataset.

The regular Church Farm samples highlighted seasonal changes in abundance. Oct '22–Apr '23 showed low detections, with *Ustilago* highest (1.20 ± 0.40 HP100k). In May–Aug '23, abundances increased markedly, led by *Ustilago* (1883 ± 952), *Zymoseptoria* (293 ± 263), and *Pyrenophora* (49.36 ± 17.76). In 2024, community composition was broadly similar across both periods: *Ustilago* was most abundant in Feb–Jun (695 ± 336), while *Zymoseptoria* dominated in Jul–Sep (429 ± 128). *Puccinia* and *Pyrenophora* were also moderate, whereas *Fusarium*, *Claviceps*, and *Blumeria* remained low.

Among the 24-hour collections, *Ustilago* in Jun '24 reached the highest abundance overall (4031.05 ± 429.34), followed by *Puccinia* in May '24 (1385.27 ± 178.17) and *Parastagonospora* in Aug '23 (419.02 ± 54.48). *Claviceps* and *Fusarium* were consistently among the lowest.

The BOP samples also showed temporal differences. *Ustilago* dominated in Jun and Jul '24 (1116.85 ± 396.14; 966.02 ± 267.95), but was least abundant in May (7.23 ± 1.22),

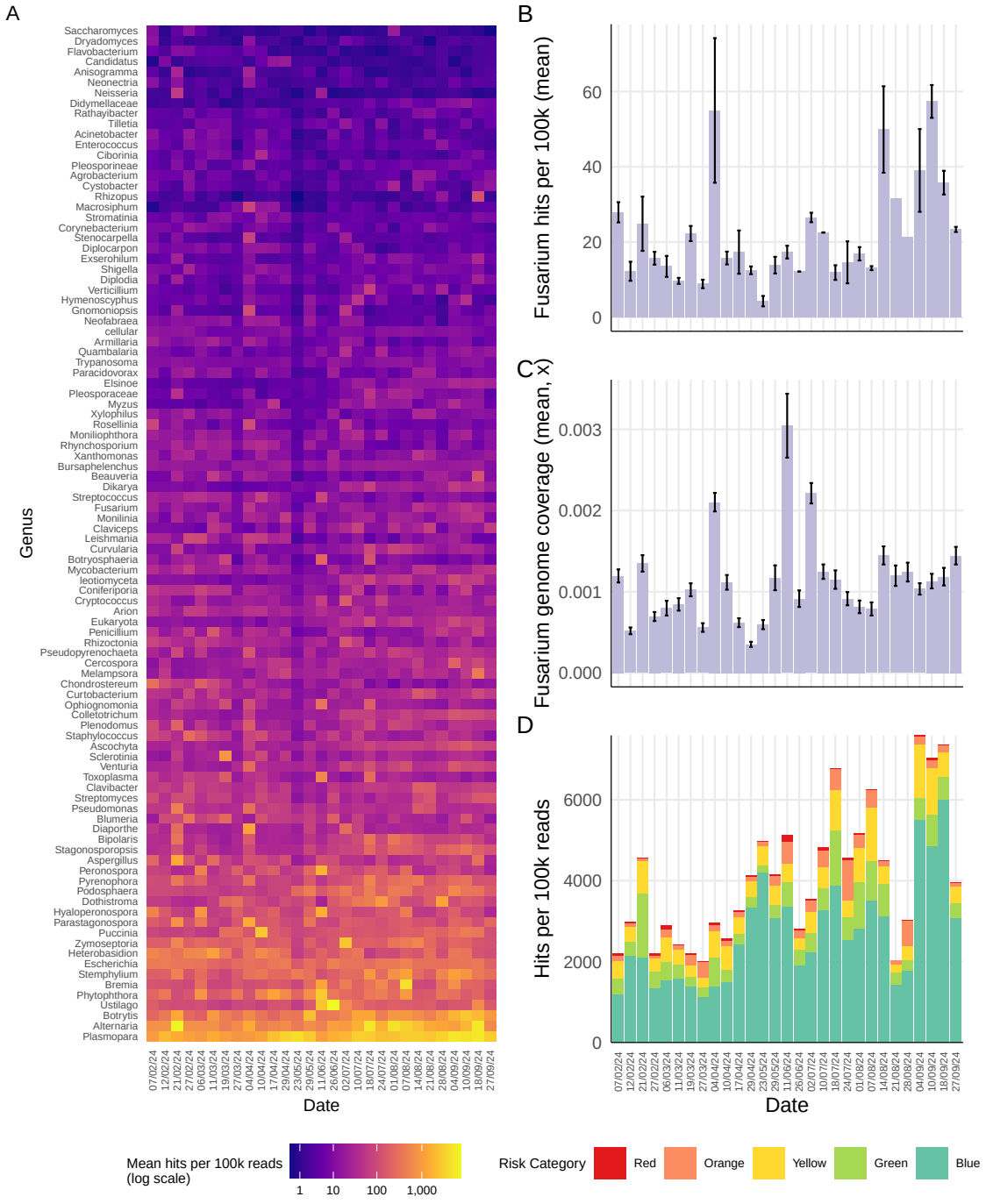


Figure 7.7: MARMoT Outputs for Regular Church Farm 24 collections: Each timepoint represents a weekly collection with two replicates; values are shown as means (A, B, C) or sums (D). (A) Heatmap of genus-level relative abundance across sampling dates, expressed as hits per 100k with a log scale. Genera were included if detected at >10 reads per 100k in at least two samples. (B) Mean *Fusarium* read counts per 100k with standard error. (C) Mean *Fusarium* genome coverage with standard error. (D) Risk classification plot showing hits per 100k reads for DEFRA risk species, grouped by risk category (red = highest risk, blue = lowest risk).

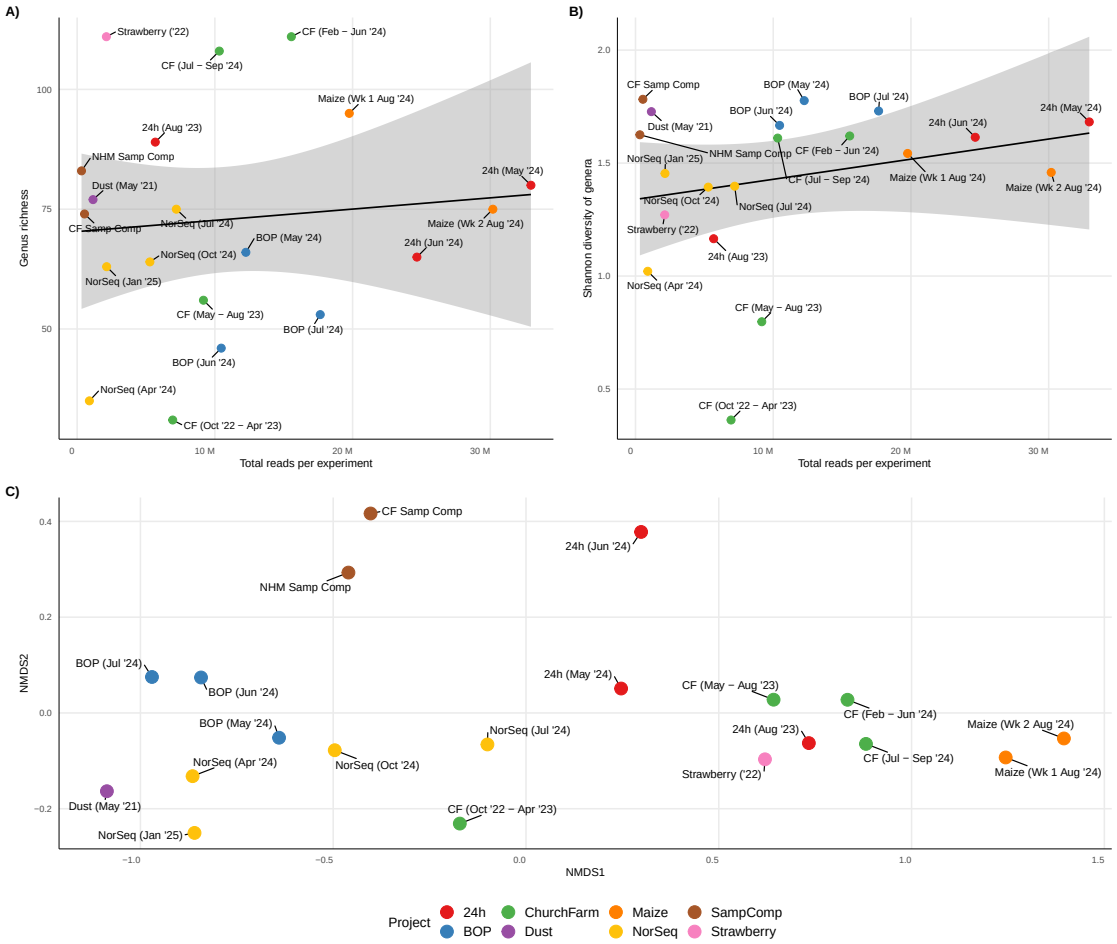


Figure 7.8: Diversity and community composition by experiment, diversity metrics were calculated at genus level using HP100k abundances per group. Richness counts genera with HP100k greater than zero and Shannon diversity is computed from the HP100k abundance vector. Points are coloured by project and labelled by experimental group. **A)** Genus richness versus total reads per group with a linear fit and 95 percent confidence ribbon. **B)** Shannon diversity of genera versus total reads per group, fit and confidence ribbon as in A. **C)** Non metric multidimensional scaling (NMDS) of Bray Curtis dissimilarities on genus abundances, two dimensions (Stress = 0.038). Abbreviations: CF = Church Farm; NorSeq = NorfolkSeq; SampComp = Sampler Comparison.

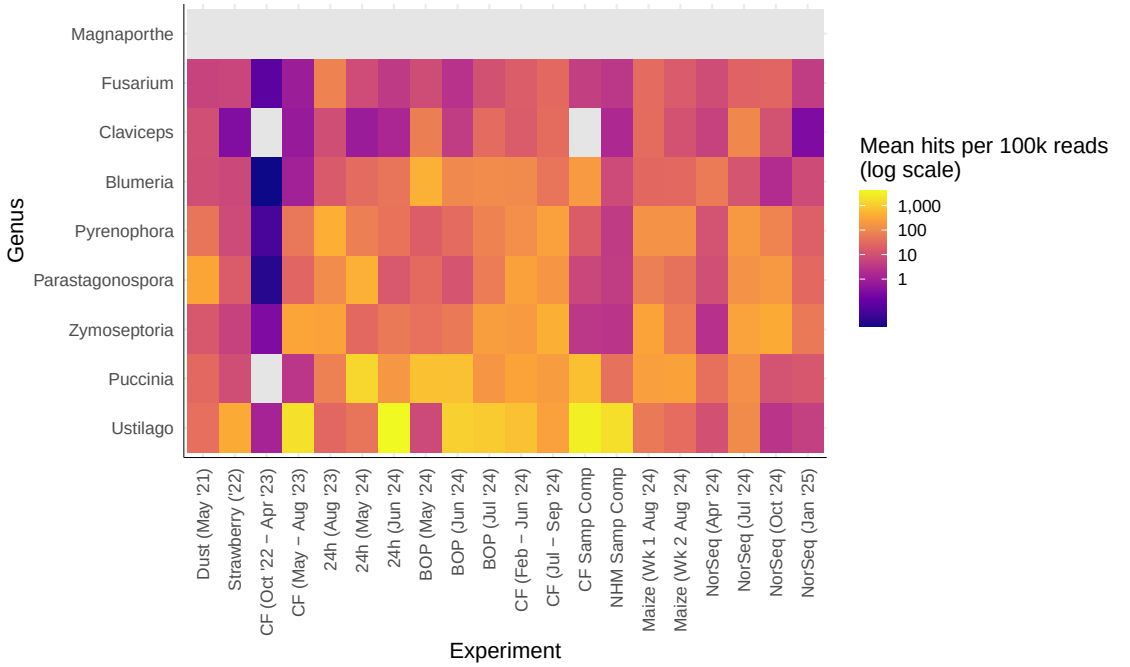


Figure 7.9: Heatmap showing the mean hits per 100k reads for the nine genera of interest across all experiments. Values are averaged across replicates and displayed on a log scale. Abbreviations: CF = Church Farm; NorSeq = NorfolkSeq; SampComp = Sampler Comparison

when *Puccinia* was highest (718.47 ± 284.45). *Blumeria*, *Zymoseptoria*, and *Pyrenophora* were also common, while *Fusarium* abundance remained low throughout.

In the sampler comparison experiment, *Ustilago* again dominated (Church Farm: 2951.92 ± 438.22; NHM: 1767.44 ± 298.81). *Puccinia* and *Blumeria* were elevated at Church Farm (680.56 ± 144.64; 192.85 ± 42.13) compared with NHM (39.67 ± 26.16; 7.63 ± 3.18). *Fusarium* and *Zymoseptoria* were low in both, and *Claviceps* was detected only at NHM.

In the maize experiment, *Zymoseptoria* and *Puccinia* were most abundant in week 1 (277.42 ± 38.08; 236.69 ± 30.41), with *Puccinia* remaining high in week 2 (268.68 ± 17.41) alongside increased *Pyrenophora* (143.87 ± 11.52). *Parastagonospora* and *Ustilago* were moderate, while *Claviceps*, *Fusarium*, and *Blumeria* were low.

The NorfolkSeq samples were dominated by *Zymoseptoria*, especially in Oct '24 (354.65 ± 54.70) and Jul '24 (263.81 ± 42.53). *Pyrenophora* (187.64 ± 14.73) and *Parastagonospora* (179.21 ± 24.14) also reached high levels, while *Puccinia* and *Ustilago* were moderate (100–130 HP100k). *Claviceps*, *Fusarium*, and *Blumeria* remained low.

Overall, these patterns show that while most pathogens of interest were widely detected across experiments, their relative abundances varied substantially by dataset and sampling date.

7.5.5 Risk Register pathogens per experiment and risk category

Abundance and diversity of detected taxa present in the DEFRA risk register were assessed by presence and risk category (Figure 7.10; Table 7.7). Widespread taxa were mainly yellow with some red, while limited taxa were largely orange and red. Absent taxa spanned multiple risk categories, unknown taxa were almost exclusively blue, and the N/A group was primarily green and yellow.

Patterns in species richness mirrored these abundance trends (Table 7.7). Across exper-

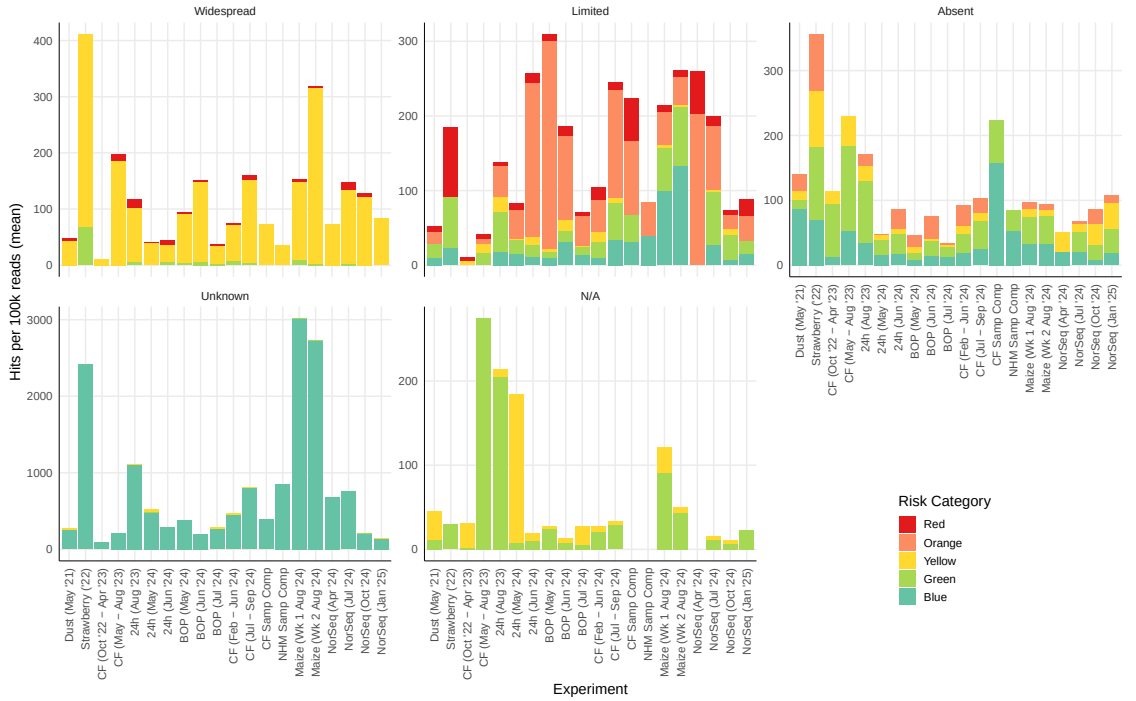


Figure 7.10: Abundance of detected genera grouped by presence status (Widespread, Limited, Absent, Unknown, N/A) across all experiments. Bars represent mean hits per 100k reads, coloured by risk category. Abbreviations: CF = Church Farm; NorSeq = NorfolkSeq; SampComp = Sampler Comparison

iments, relatively few red- and orange-risk species were detected (typically 0–2 per dataset), whereas yellow and green taxa comprised the largest proportion of unique species, and blue taxa were also numerous. Across all datasets, a maximum of 6 orange- and 2 red-risk species were identified in any single experiment, compared with up to 21 yellow, 20 green, and 21 blue taxa.

There was also a clear relationship between sequencing depth and species richness. For example, in the 24 h May '24 collection, which generated a high total read number (33.02 M), 58 species were detected, including 2 red, 6 orange, and 17 yellow taxa. By contrast, NorfolkSeq Apr '24 with only 884 k reads, and the sampler comparison experiments with 516 and 299 k reads, yielded just 8–14 risk category species in total.

Overall, these results show that the distribution of taxa by presence category is closely linked to their assigned risk level and the number of identified risk categorised species is linked to sequencing depth.

7.5.6 Identified species which are absent in the Risk Register

A total of 34 taxa classified as "Absent" in the UK by the DEFRA risk register were detected above the pipeline read count, identity and abundance thresholds. The identified absent taxa spanned a broad range of risk categories (Figure 7.11, Table 7.8). The majority were blue risk (12), followed by yellow and green (10 each) and orange (2), there were no absent red risk species identified. Across the different experiments these species were detected sporadically, with none being present in every sample.

Both Church Farm '24 experiments contained the highest number of unique absent species (21 and 19) alongside the Maize wk1 group (20 species). Meanwhile both of

Table 7.7: Number of distinct species detected per risk category[a]

Experiment	**Total reads**	**Number of unique species**					
		Total	**Red**	**Orange**	**Yellow**	**Green**	**Blue**
Dust (May '21)	1.15 M	39	2	3	10	13	11
Strawberry ('22)	1.94 M	24	1	1	4	9	9
Church Farm (Oct '22 – Apr '23)	6.65 M	21	1	0	5	5	10
Church Farm (May – Aug '23)	9.09 M	26	2	1	5	7	11
24h (Aug '23)	5.67 M	51	3	4	13	15	16
24h (May '24)	33.02 M	58	2	6	17	14	19
24h (Jun '24)	24.67 M	51	3	4	15	14	15
BOP (May '24)	12.25 M	63	3	5	17	20	18
BOP (Jun '24)	10.46 M	39	3	3	11	11	11
BOP (Jul '24)	17.65 M	52	2	6	14	15	15
Church Farm (Feb – Jun '24)	15.55 M	59	3	4	17	19	16
Church Farm (Jul – Sep '24)	10.32 M	58	3	5	15	18	17
Church Farm Sampler Comparison	516 k	14	1	1	2	2	8
NHM Sampler Comparison	299 k	7	0	1	1	1	4
Maize (Wk 1 Aug '24)	19.76 M	56	3	5	13	17	18
Maize (Wk 2 Aug '24)	30.19 M	54	3	5	15	15	16
NorfolkSeq (Apr '24)	884 k	8	1	1	2	0	4
NorfolkSeq (Jul '24)	7.15 M	49	2	4	13	15	15
NorfolkSeq (Oct '24)	5.28 M	45	3	3	13	13	13
NorfolkSeq (Jan '25)	2.07 M	25	1	2	5	9	8

[a] Read counts rounded to the nearest thousand (k) or million (M).

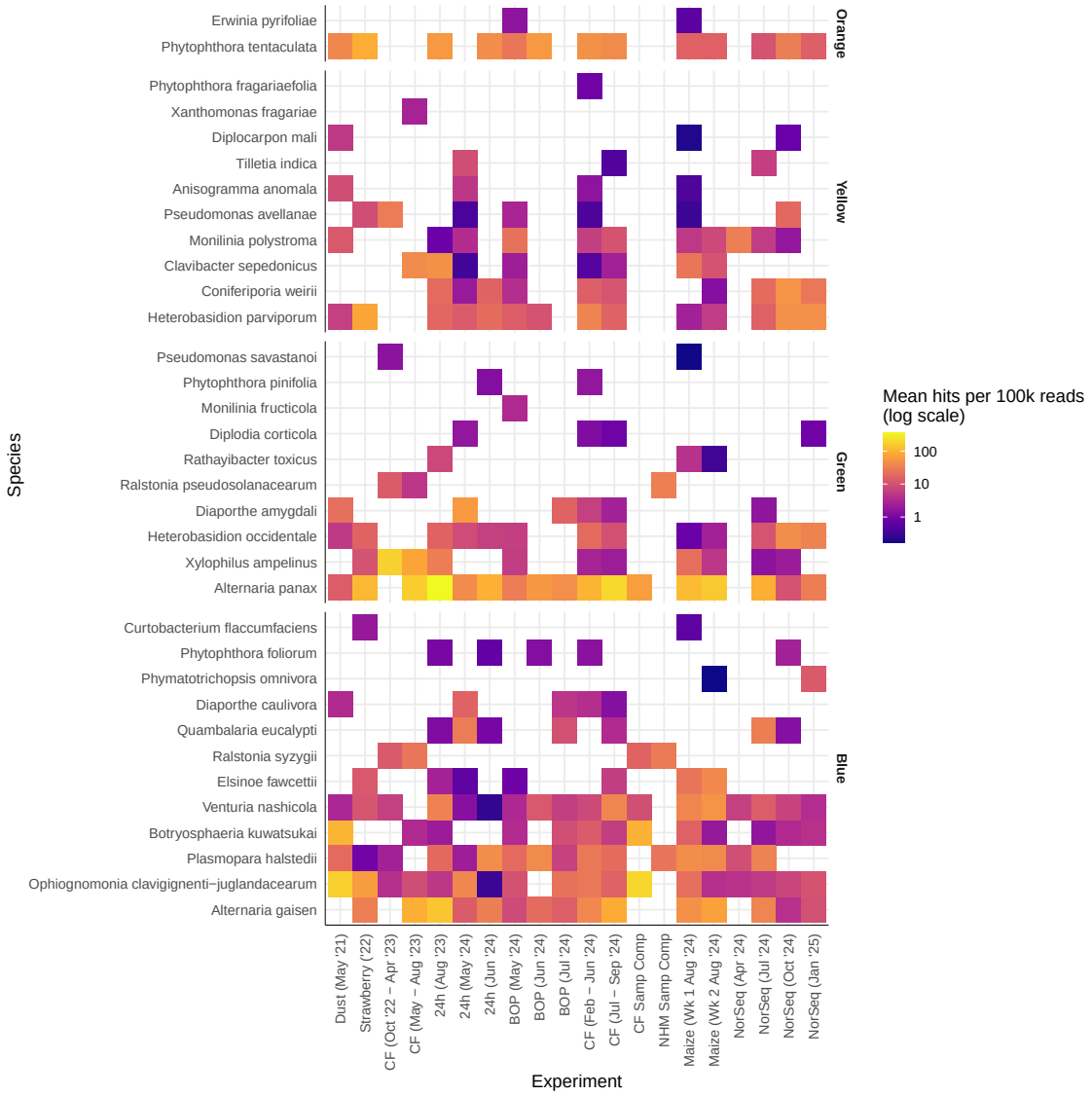


Figure 7.11: Heatmap showing the mean hits per 100k reads for the species which are classified as absent in the UK by the DEFRA risk register. Values are averaged across replicates and displayed on a log scale. Plots are faceted by risk category, data were filtered on HP100k >10. Abbreviations: CF = Church Farm; NorSeq = NorfolkSeq; SampComp = Sampler Comparison

the Sampler Comparison experiments and NorfolkSeq Apr '24 contained the least absent species (3-5). The number of identified absent species was moderately positively correlated with total reads per experiment (Pearson's r = 0.58, 95% CI 0.18 to 0.81, p = 0.008, n = 20), indicating that sequencing depth influenced the likelihood of detecting absent species.

Many of the absent species detected were sporadic: none were found in every experiment, 11 occurred in over half of the experimental groups, and 3 were observed only once. The most prevalent were *Ophiognomonia clavigignenti-juglandacearum*, a fungus of butternut, and *Venturia nashicola*, which infects pear; both were detected in 18 of the 20 experiments. In contrast, *Monilinia fructicola*, *Phytophthora fragariaefolia*, and *Xanthomonas fragariae* were each detected in a single, different experiment. *Alternaria panax* showed the highest abundance, with mean values of 365.2 HP100k in the 24h (Aug '23) experiment and 195.0 HP100k in Church Farm (Jul–Sep '24).

Closer examination of changes between experiments highlights variation in detections across months, seasons, and years. In the 24h experiments, for example, *Ophiognomonia clavigignenti-juglandacearum* peaked in May '24 (37.5 HP100k) but was much lower the year before (5.0 HP100k) and in the following month (0.3 HP100k). By contrast, *Alternaria panax* reached its highest abundance in the 24h Aug '23 samples, with substantially lower levels in the corresponding 2024 collections. Comparable seasonal shifts in species abundance were also observed between the Church Farm, BOP, and NorfolkSeq datasets.

Overall, the detection profile was highly variable, dominated by sporadic low-level hits, with only a subset of pathogens showing recurrent detection across experiments.

7.5.7 Identified species in the red and orange risk categories

Across all the samples 6 species were identified that were labelled as red or orange risk, 3 of each category (Table 7.9). These cover broad pest types and known UK presence.

Several of these species were consistently detected across the majority of experiments, shown in Figure 7.12. For example, *Arion vulgaris* and *Dothistroma septosporum* were identified at comparatively high abundance in most datasets. In contrast, other taxa were detected only sporadically and usually at very low abundance. For instance, *Erwinia pyrifoliae* and *Hymenoscyphus fraxineus* appeared in just a handful of experiments, often near the detection threshold, while *Phytophthora ramorum* and *Phytophthora tentaculata* showed intermediate patterns, being present in several experiments but with variable abundance.

Across experiments, there was also a tendency for a greater number of red and orange species to be recovered when sequencing depth was higher, even after normalisation and filtering (Pearson's r = 0.36, 95% CI: –0.12 to 0.70, p = 0.14). Although this relationship was not statistically significant, it suggests that sequencing effort may still influence the likelihood of detecting rare high-risk taxa.

Together, these results indicate that while a subset of species are persistently detected across sites and sampling strategies, many others occur rarely or only at trace levels, highlighting variability in both prevalence and detectability among high-risk taxa.

Table 7.8: Table of species that are listed as absent in the DEFRA risk register which were identified in the data

Species	**Type**	**Risk**	**Disease**
Alternaria gaisen	Fungus	Blue	Black spot of Japanese pear
Alternaria panax	Fungus	Green	Ginseng Leaf spot
Anisogramma anomala	Fungus	Yellow	Blight of hazel or Eastern filbert
Botryosphaeria kuwatsukai	Fungus	Blue	Blister canker and ring rot of pome fruits
Clavibacter sepedonicus	Bacterium	Yellow	Potato Ring Rot
Coniferiporia weirii	Fungus	Yellow	Root rot of conifers
Curtobacterium flaccumfaciens	Bacterium	Blue	Bean tan spot; Poinsettia Leaf spot
Diaporthe amygdali	Fungus	Green	Canker of almond or peach
Diaporthe caulivora	Fungus	Blue	Stem canker of soybean
Diplocarpon mali	Fungus	Yellow	Sooty blotch of apple
Diplodia corticola	Fungus	Green	Bot canker
Elsinoe fawcettii	Fungus	Blue	Citrus scab
Erwinia pyrifoliae	Bacterium	Orange	Necrotic disease of Asian pear
Heterobasidion occidentale	Fungus	Green	Root and butt rot of conifers
Heterobasidion parviporum	Fungus	Yellow	Root and butt rot of conifers
Monilinia fructicola	Fungus	Green	Brown rot of apple and stone fruits
Monilinia polystroma	Fungus	Yellow	Asiatic Brown Rot
Ophiognomonia clavigignenti-juglandacearum	Fungus	Blue	Butternut canker
Phymatotrichopsis omnivora	Fungus	Blue	Root and soft rots of conifers, soybean, and cotton
Phytophthora foliorum	Oomycete	Blue	Leaf blight of azaleas
Phytophthora fragariaefolia	Oomycete	Yellow	Red stele insStrawberries
Phytophthora pinifolia	Oomycete	Green	Pine foliar damage
Phytophthora tentaculata	Oomycete	Orange	Root and stalk rot
Plasmopara halstedii	Oomycete	Blue	Downy mildew of sunflower
Pseudomonas avellanae	Bacterium	Yellow	Stem dieback of hazel nut
Pseudomonas savastanoi	Bacterium	Green	Infects a variety of plants
Quambalaria eucalypti	Fungus	Blue	Leaf and shoot blight of eucalyptus
Ralstonia pseudosolanacearum	Bacterium	Green	Wilting disease
Ralstonia syzygii	Bacterium	Blue	Sumatra disease of clove; banana blood disease
Rathayibacter toxicus	Bacterium	Green	Annual ryegrass toxicity
Tilletia indica	Fungus	Yellow	Bunt of wheat
Venturia nashicola	Fungus	Blue	Scab of Japanese pear
Xanthomonas fragariae	Bacterium	Yellow	Leaf spot and blight of strawberry
Xylophilus ampelinus	Bacterium	Green	Blight and canker of grapevine

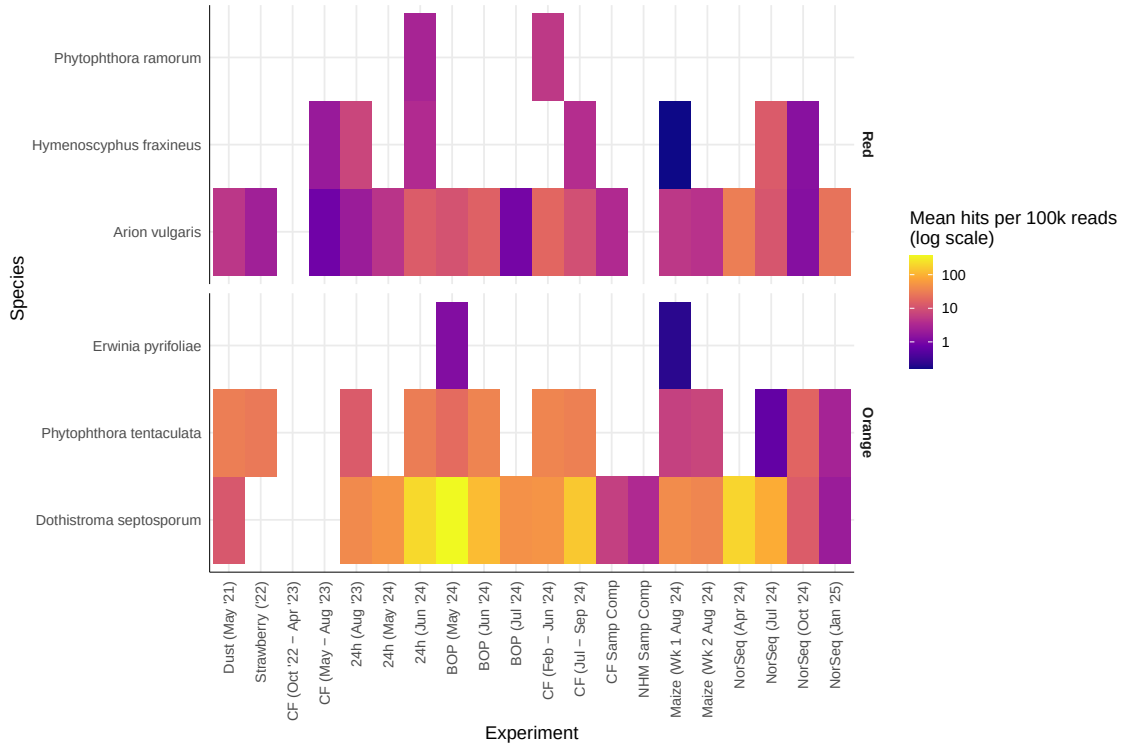


Figure 7.12: Heatmap showing the mean hits per 100k reads for the species which are classified as red or orange risk in the UK by the DEFRA risk register. Values are averaged across replicates and displayed on a log scale. Plots are faceted by risk category, data were filtered on HP100k >10. Abbreviations: CF = Church Farm; NorSeq = NorfolkSeq; SampComp = Sampler Comparison

7.5.8 Genome coverage of detected species

All detected taxa showed low genome coverage (Fig. 7.13), computed as the percentage of mapped bases relative to reference genome length. Compared to the non-red categories, the red risk species exhibited markedly lower coverage (Fig. 7.13A). The highest coverage occurred in the green category (mean $7.73\% \pm 0.41\%$), followed by yellow ($5.67\% \pm 0.34\%$). Within the red risk species, the highest mean coverage was observed for *Phytophthora ramorum* ($2.00\% \pm 0.47\%$), followed by *Hymenoscyphus fraxineus* ($0.16\% \pm 0.02\%$); *Arion vulgaris* showed extremely low coverage (0.01%, ≈ 0).

Across experiments, coverage varied among the three red risk taxa (Fig. 7.13B). *A. vulgaris* was detected in the most experiments but at the lowest coverage (maximum 0.017%). *H. fraxineus* was detected in 7 experiments with 0.01–0.23% coverage. *P. ramorum* was detected in 2 experiments with higher coverage than the other red taxa (1.62% and 2.65%), though still far below whole-genome coverage, both of these experiments contain samples collected at Church Farm in June 2024.

These results show the utility of the MARMoT pipeline to quickly identify the abundance and coverage of different pathogens in a ONT sequenced air sample. There are clear differences in the type and abundance of pathogens detected in the different experiments from across this thesis and beyond.

Table 7.9: Table of species that are listed as red / orange risk in the DEFRA risk register which were identified in the data

Species	**Type**	**UK Presence**	**Risk**	**Disease**
Arion vulgaris	Mollusc	Limited	Red	Spanish slug
Dothistroma septosporum	Fungus	Limited	Orange	Dothistroma needle blight; Red band needle blight
Erwinia pyrifoliae	Bacterium	Absent	Orange	Necrotic disease of apple and pear
Hymenoscyphus fraxineus	Fungus	Widespread	Red	Ash dieback
Phytophthora ramorum	Oomycete	Limited	Red	Ramorum leaf blight and shoot dieback; Rhododendron twig blight; Sudden oak death
Phytophthora tentaculata	Oomycete	Absent	Orange	Root and stalk rot

7.6 Discussion

The parameters used in the MARMoT pipeline were selected through validation experiments that tested alternative alignment settings and filtering thresholds. These assessments showed that the chosen configuration produced well-calibrated results and minimised spurious assignments, providing confidence in the downstream analyses.

Overall, the results demonstrate that the MARMoT pipeline is capable of detecting a wide range of potential pathogens from airborne sequencing data. Eight of the nine genera of known wheat pathogens considered across this thesis were detected in many experiments with varied abundance. Pathogens were then classified into DEFRA risk categories and by their known distribution to determine how abundance changed across experiments. From here the species which were classified as 'absent' in the UK were considered in more depth and shown to have sporadic presence across experiments, as were those classified as red or orange risk.

Genome coverage was evaluated across all risk categories, with particular focus on the three identified red-risk species. Coverage for these species was extremely low, which is consistent with their presumed low abundance in air samples relative to other taxa. The recurrent detection of *Arion vulgaris* at low levels in nearly all samples likely reflects limitations of the reference genome rather than true presence. In contrast, the detection of *Hymenoscyphus fraxineus*, already known to be widespread in the UK, and *Phytophthora ramorum* in two experiments conducted in similar locations and time periods, appears more credible.

7.6.1 Parameter choice

The pipeline parameters were optimised using different approaches depending on the tool. *Minimap2* settings were refined using real sequencing data, reflecting the complexity and variability inherent in environmental samples. In contrast, *LCAParse* parameters were

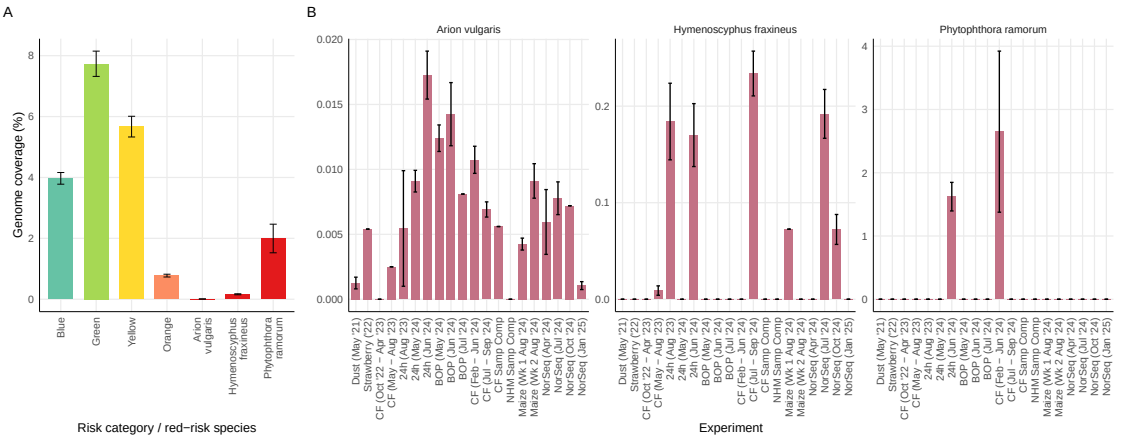


Figure 7.13: Genome coverage by species risk category, across all facets, bar heights represent mean genome coverage (%) and error bars are ±SE. **A)** Mean coverage by risk category; bars show the mean coverage (%) across experiments for each risk category (blue, green, yellow, orange) and for the three red risk species individually. **B–D)** Genome coverage across experiments for the three red-risk species: *Arion vulgaris*, *Phytophthora ramorum* and *Hymenoscyphus fraxineus*. Abbreviations: CF = Church Farm; NorSeq = NorfolkSeq; SampComp = Sampler Comparison

optimised with simulated datasets, which, although less diverse than typical airborne communities (17 species compared with potentially thousands), provide the crucial advantage of a known ground truth. This allows confident assessment of false positive assignments, which is not possible with real-world data.

Whether working with real or simulated data, there is always a trade-off in setting parameter stringency: prioritising the elimination of false positives at the cost of missing some taxa, or maximising detections while accepting the risk of errors. In the context of pathogen detection from air samples, high stringency is preferable to ensure that information passed on to growers is reliable.

In this study, the trade-off was apparent when comparing *LCAParse* parameters for the simulated dataset (Figure 7.6). Increasing the identity threshold to 90% greatly reduced the number of incorrect assignments but also halved the number of reads assigned, leaving many unclassified. An intermediate threshold of 85% was therefore selected as a compromise, reducing the number of false assignments seen at lower thresholds while still retaining a higher proportion of classified reads. A similar, though less pronounced, effect was observed when increasing the minimum alignment length, and a 150 bp cut-off was chosen for the same reason. These parameters reflect the balance between sensitivity and specificity required for species-level pathogen detection, where the consequences of misidentification are more problematic than the underestimation of total community diversity.

This trade-off highlights a central challenge in metagenomic pipeline optimisation: parameter thresholds are rarely universal. Different studies adopt varying levels of stringency depending on their aims, the sequencing depth available, and the ecological or clinical context. Optimal parameter values are also dataset-dependent, with read length, sequencing error profiles, and community complexity all influencing how stringent filters should be set. The difficulty in selecting appropriate filters and thresholds for metagenomic data has been discussed in a number of reviews [50, 150, 195]

Ultimately, there is no clear optimum set of parameters, but rather a balance between sensitivity and specificity. Future development of benchmarking datasets for airborne

metagenomics, ideally representing realistic levels of complexity, would provide a valuable means to assess how parameter choices shape taxonomic outputs. Until then, transparency in parameter selection and acknowledgement of their limitations are essential to enable reproducibility and fair comparison across studies.

7.6.2 Church Farm 2024 example outputs

The example output from the weekly Church Farm collections in 2024 demonstrates the interpretability of the MARMoT pipeline outputs (Figure 7.7). The graphical outputs enable the exploration of broad scale genus level dynamics as well as the ability to highlight specific pathogenic genera, which in this example was *Fusarium*. Additionally, the integration of the `barcode_labels.tsv` file enables the data to be labelled by the date collected and therefore chronologically ordered with repeat samples averaged.

These outputs clearly show the dynamic nature of the airborne community composition, with a few consistently abundant pathogenic genera such as *Plasmopara*, *Alternaria*, *Botrytis* and *Ustilago*. The vast majority of other identified genera were present at low abundance with sporadic increases across the monitoring season, illustrated by the requirement to use a log scale to clearly visualise the different abundances (Figure 7.7A).

The high abundance of *Alternaria* detected in air samples aligns with previous reports from multiple studies [18, 345]. In contrast, the predominance of *Plasmopara* was unexpected. The reference database used here includes only two *Plasmopara* species (Table A.3). *P. halstedii* and *P. obducens*, both of which are listed as absent on the DEFRA risk register. *P. halstedii* is a causal agent of sunflower downy mildew and has been detected in several other experiments within this chapter (Figure 7.11). While it can disperse by airborne spores, its primary infection route is through soil-mediated root infection [136]. *P. obducens*, the causal agent of downy mildew in *Impatiens*, is less well characterised, and its dispersal mechanisms remain unclear. To date, there are no published studies reporting airborne detection of either *P. halstedii* or *P. obducens*, although airborne monitoring of *P. viticola*, the causal agent of grapevine downy mildew, has been demonstrated [67, 258]. Consequently, it remains uncertain whether the high *Plasmopara* signal observed in the 2024 Church Farm samples represents genuine detection of pathogens not currently considered present in the UK, or misassignment from closely related taxa.

Among the nine target pathogens, *Fusarium* was selected for inclusion in the summary output, with both relative abundance and genome coverage tracked over time (Figures 7.7B and C). Abundance fluctuated across the season, peaking in April and remaining relatively high towards September, a pattern consistent with the MARTi analysis presented in Chapter 5. In contrast, genome coverage remained consistently low, indicating that reads did not span large portions of the *Fusarium* reference genomes. It is unclear whether this reflects the low biomass typical of airborne samples, potential misassignment of reads, or sequencing biases. Notably, the timing of peak genome coverage did not coincide with peak abundance, suggesting that genome coverage may provide additional information for assessing the reliability of taxonomic assignments in pathogen surveillance. Further research is required to establish how coverage metrics should be interpreted in the context of airborne monitoring.

Across all species assigned a DEFRA risk register category, the majority were classified as low risk (blue), indicating that most detections were not of immediate concern (Figure

7.7D). Green, yellow, and orange risk taxa were consistently present across all time points, whereas red risk taxa were detected only sporadically. This demonstrates the potential of the pipeline to highlight samples containing higher-risk detections, which can then be prioritised for further scrutiny to assess the reliability of the taxonomic assignments. The detection of red risk taxa is discussed in more detail later in this chapter.

Overall, the summary plot from the 2024 Church Farm dataset illustrates how the MARMoT pipeline can be applied to monitor both the diversity and abundance of airborne pathogens, determine the abundance and genome coverage of the nine target genera and consider the abundance of different risk associated pathogens between airborne samples.

7.6.3 Diversity across the different datasets

A comparison of the genus richness and Shannon diversity of the different experiments showed no significant relationship with sequencing depth (Figures 7.8A and B). This suggests that the filtering and normalisation to HP100k was sufficient to account for the differences in sequencing effort, therefore the observed diversity patterns are more likely biological than technical.

The ordination plot (Figure 7.8C) showed clear clustering of experiments carried out using the same protocol. One example is the sampler comparison experiments, which were divided into two groups based on location (Church Farm and NHM), the clustering of these two experiments suggests that the variety of samplers used may have exerted a stronger influence on diversity than sampling site, consistent with previous findings [155].

The Church Farm (May–Aug '23) sample clustered with the two Church Farm 2024 experiments despite differences in duration and the use of WGA, indicating that location in this case had a greater effect on diversity than processing method. By contrast, Church Farm (Oct '22–Apr '23) did not cluster with the other Church Farm samples, suggesting temporal or seasonal effects. The 24-hour experiments were more dispersed; notably, both 2024 experiments were collected and processed identically, yet did not cluster, again pointing towards seasonal variation.

These results highlight that multiple factors, including protocol, location, and season, can shape observed diversity. Given the heterogeneity in sampling duration, location, and processing across experiments, it remains difficult to disentangle the relative contributions of methodological versus ecological drivers of clustering.

7.6.4 Nine pathogens of interest

Most of the nine target genera were consistently detected across experiments, demonstrating the pipeline's capacity to capture agriculturally relevant taxa (Figure 7.9). *Magnaporthe* was absent throughout, which is expected given that it has not been reported in the UK. *Claviceps* and *Puccinia* were only absent from the Church Farm (Oct '22–Apr '23) dataset, which may reflect seasonal variation in airborne pathogen loads or reduced sequencing performance, as this experiment consistently yielded fewer taxa overall (Figure 7.10).

The heatmap highlights variation in the abundance of these pathogens across experiments. Detections were generally higher at cereal-growing sites (BOP and Church Farm) compared with the NorfolkSeq samples, which included only one wheat farm among the eight locations surveyed. The Church Farm sampler comparison also revealed elevated

Ustilago, *Puccinia*, and *Blumeria* compared with the London (NHM) site, suggesting a stronger influence of local cropping context than geographic location.

Seasonal trends were also apparent. For example, *Ustilago* abundance was elevated in the 24h June sample relative to other 24h experiments, and decreased in the May BOP sample compared with later BOP collections. Similar temporal shifts were evident for several other pathogens.

A notable observation was the detection of *Claviceps* in the Church Farm 2024 samples. This taxon was not identified using MARTi in the previous analysis of the same data (Chapter 5), despite the pathogen being visually observed nearby. This discrepancy may reflect greater sensitivity of the pipeline, but could also be an artefact of using a smaller reference database. Where the earlier analysis employed the BLAST nt database, the MARMoT pipeline used a smaller pathogen specific database. Community composition is known to vary considerably depending on the database employed [346]. However, given the stringent filtering applied in the pipeline, it is also plausible that the detection represents a true identification, underscoring the complexity of interpreting airborne metagenomic data in the absence of a definitive ground truth.

7.6.5 Risk by presence

There are clear differences in the abundance of taxa across the experiments both by risk category and recorded presence in the UK. Figure 7.10 indicates that known presence strongly influences risk category, with taxa of unknown status predominantly classified as blue, widespread taxa largely yellow, and limited taxa containing a greater proportion of red and orange risk species. The DEFRA risk register assigns scores using a framework that combines likelihood, impact, and value at risk, with the likelihood component calculated differently depending on whether a pest is present or absent in the UK. For taxa recorded as present, likelihood reflects the potential for spread to the maximum extent, whereas for absent taxa it reflects the probability of entry and establishment. This relationship highlights how distribution status contributes directly to risk classification and therefore may explain the presence and risk level pattern seen in the data.

Although no statistically significant relationship was observed between overall species diversity and sequencing depth, the number of reads was associated with the total number of species detected across different risk levels (Table 7.7). This likely reflects the fact that only a subset of species in the reference database belong to the DEFRA risk register and are assigned a risk level. For example, the BOP (May '24) dataset with 12.25 M reads contained 63 unique species, whereas the NHM sampler comparison with only 299 k reads yielded just 7 unique risk register taxa. This pattern was not entirely consistent, as the BOP (Jul '24) dataset generated more reads than in May but contained fewer identified risk register taxa. Overall, however, higher read counts tended to correspond with greater risk register species detection, including increased identification of species which were documented as Absent in the UK in experiments with higher sequencing depth. Similar associations between sequencing depth and observed richness have been reported in other metagenomic studies [409].

Differences in sequencing depth are likely influenced by collection length, which affects DNA yield, as well as by the number of samples within each group (Table 7.5) and the period over which the samples were collected. Collection duration alone does not explain

the patterns observed, since both the week-long Maize collections and the 3-day 24h experiments recovered large numbers of unique species. Seasonal variation may also contribute to the observed differences. For example, the NorfolkSeq dataset from Jan '25 contained only 25 unique risk register species, compared with 49 and 45 in the Jul and Oct '24 collections. Yet it remains difficult to disentangle whether this difference reflects genuine seasonal dynamics in airborne communities or the lower sequencing depth of the January sample (2.07 M reads compared with 7.15 M and 5.28 M, respectively).

The increase in the number of detected risk register species also coincided with an increase in the number of high-risk taxa identified, although no dataset contained more than three red risk species. These were the same three species detected consistently across experiments and are discussed in greater detail below.

Together, these results demonstrate that the pipeline is capable of resolving both abundance and diversity across DEFRA risk levels, while also identifying instances where higher-risk taxa emerge in samples, even at low abundance. This highlights the potential value of the approach for integration into surveillance frameworks, where early warning of such detections is critical.

7.6.6 Species considered absent in the UK detected with the pipeline

Across all the experimental data the MARMoT pipeline detected 34 species which are considered absent in the UK according to the DEFRA risk register (Table 7.8). There are two main explanations for these findings: either the species are present in the UK and the risk register is incorrect, or the taxonomic assignments are erroneous.

The first explanation is plausible, as the register relies on published sources that require accurate field identification and subsequent reporting. Moreover, many entries are several years old, and low-risk species are not notifiable, making the accumulation and maintenance of reliable presence data more difficult. Among the species identified as absent, only two were categorised at the higher orange risk level: *Erwinia pyrifoliae*, with its record last updated in 2020, and *Phytophthora tentaculata*, updated in 2022.

Alternatively, these detections may reflect incorrect assignments of sequence reads. Contributing factors could include the limited size of the reference database, insufficiently stringent filtering parameters, or the absence of genome coverage thresholds in the classification process. The potential for and consequences of such erroneous assignments are considered further in the limitations section.

These two possibilities can be illustrated with the case of *P. tentaculata*, which is reported as absent in the risk register. This species was recorded in the UK in 2013 [32], demonstrating that the risk register is not always a fully accurate record. However, there have been no confirmed reports of its presence since, raising uncertainty over whether it remains in the UK. Therefore, *P. tentaculata*'s apparent detection in more than half of the experiments presented here (Figure 7.11) is perhaps more plausibly explained by misassignment to a more abundant *Phytophthora* species, as a pathogen occurring at such prevalence would almost certainly have been detected by conventional surveys.

Overall, these results highlight the difficulty of distinguishing between genuine detections of under-reported taxa and erroneous assignments, underscoring the need for careful interpretation of presence data obtained from the MARMoT pipeline.

7.6.7 High risk species detections

The occurrence of red and orange risk species identified by the pipeline varied across experiments (Figure 7.12). To assess the reliability of these detections, the three red risk species are now considered in greater detail.

Across all red risk detections, genome coverage was markedly lower than in the other risk categories, ranging from 0.01 to 2% (Figure 7.13A). The aim of examining genome coverage alongside read abundance was to gain additional insight into how reads mapped across reference genomes. Within the MARMoT pipeline, coverage was calculated by summing the lengths of all reads that aligned to a given taxon with at least 80% identity and 80% alignment coverage, and dividing this total by the length of that taxon's reference genome. Reads meeting these thresholds for more than one taxon were counted towards each. This measure therefore reflects cumulative mapped bases only; it does not capture the evenness or breadth of read distribution across the genome, nor does it distinguish between primary and secondary alignments.

From the results it is difficult to conclusively determine the presence of red risk species. The lower genome coverage seen may be since these taxa are genuinely rarer and therefore yield lower levels of airborne eDNA. At present, there are no benchmarks for expected genome coverage from airborne samples, making it difficult to determine whether low values reflect incorrect assignment or the detection of rare taxa. Future analyses could address this uncertainty by incorporating genome evenness. Reads distributed across the genome would provide stronger evidence of presence, whereas clustering in conserved regions or single loci may indicate misassignment or contamination of the reference genome.

Arion vulgaris is a mollusc with only limited reported presence in the UK, yet it was detected in almost every experiment and often at a reasonable relative abundance. This pattern is unlikely to represent a true biological signal, as mollusc eDNA would not be expected to occur in airborne samples, and the associated genome coverage was extremely low (Figure 7.13B). The combination of apparently high abundance but minimal genome coverage suggests that the alignments were predominantly short in comparison to the reference genome. The consistent detection across all experiments further indicates that the reference genome itself may be of poor quality, potentially containing contaminant sequences that spuriously attract read mappings. Notably, the reference genome in the database is reported as a chromosome-level assembly, which makes contamination or assembly artefacts a plausible explanation for the repeated detection.

Hymenoscyphus fraxineus is considered a widespread pathogen and was detected in a subset of experiments, including several conducted at Church Farm and in the NorfolkSeq dataset, which spans eight sites, some of which are forested. Within NorfolkSeq, detections occurred in the July and October experiments, but not in April or January. The presence of this pathogen at sites surrounded by trees is consistent with its primary host (ash). Moreover, detections between May and October align with the seasonal pattern of spore release, which peaks in July according to [59]. However, the average genome coverage for *H. fraxineus* is low at 0.2 % for the experiments where it was detected (Figure 7.13B) and since there are not other studies on the anticipated genome coverage from airborne samples it is difficult to determine if this level is sufficient to prove presence. Overall, the detection of *H. fraxineus* in UK collected samples appears credible, but further research could be carried out to validate its presence.

Lastly, *Phytophthora ramorum*, a species with limited presence in the UK, was detected in only two experiments, both conducted at the Church Farm site in June 2024. Detection in two independent experiments from the same location and time period lends support to the possibility that this represents a true signal. The timing also aligns with previous studies, which reported that isolates of *P. ramorum* were only collectable between March and July [100]. Similarly, an airborne PCR study identified peak *Phytophthora* abundance in May–June [253]. However, this study did not detect *P. ramorum* despite using specific validated probes, suggesting an absence of airborne spores [253]. One explanation is that *P. ramorum* spores are primarily splash rather than wind-dispersed, as demonstrated by [255]. Consequently, the detections observed here may instead represent alignments to closely related airborne *Phytophthora* species.

Genome coverage for *P. ramorum* in the two experiments (1.5 and 4%; Figure 7.13C) was higher than for any of the other red risk species, but still low overall. This again raises the possibility that the mapped reads originated from a related species sharing large portions of the genome. The challenge of achieving confident species-level identification from airborne samples has been highlighted repeatedly throughout this thesis, and remains unresolved in the absence of visual confirmation of disease symptoms.

Ultimately, the evidence suggests it is unlikely that any species of great concern were genuinely detected. The apparent detection of *A. vulgaris* is most plausibly explained by incorrect assignments arising from a potentially contaminated reference genome. *H. fraxineus* may indeed have been present, but this species is already established in the UK and therefore does not represent a novel finding. In contrast, the *P. ramorum* detections are more likely to reflect alignments to closely related *Phytophthora* species, as *P. ramorum* itself is not typically dispersed via airborne spores.

It is also noteworthy that none of the reference genomes used in these assignments were complete assemblies, but rather scaffolds or chromosome-level assemblies, which may have contributed to missasignments of sequence reads.

Nevertheless, the potential identification of high-risk pathogens may still be of practical value in directing further investigation. For example, if samples from a specific field were found to contain reads aligning to high-risk taxa, agronomists could then prioritise visual inspection or implement stricter monitoring in that location. This targeted approach could improve the efficiency of surveillance by focusing resources on areas of suspicion rather than attempting comprehensive coverage across a wider region.

7.6.8 Limitations

Application of the MARMoT pipeline across multiple airborne experimental datasets has revealed limitations, particularly concerning the optimisation of analytical parameters, reference database, variability of input data and challenges of interpreting taxonomic assignments.

Firstly the parameter optimisation was not performed in a uniform manner across all stages of the pipeline. *Minimap2* parameters were optimised earlier in the project using real data, before simulated data were available. By the time simulated datasets had been generated, the chosen *minimap2* settings had already been applied to much of the analysis, and they were therefore retained for consistency. In contrast, the *LCAParse* parameters were optimised using simulated data, which provided a more reliable benchmark than real

datasets alone. Without simulated data, optimisation of this stage would have required comparison with outputs from alternative tools, which themselves may not have provided accurate taxonomic assignments. Perhaps, a future direction could be to test the entire pipeline with simulated data.

The reference database used for metagenomic analysis is known to strongly influence both the diversity and the identity of species detected [346]. In this study, where the primary aim was to detect plant pathogens and particularly those considered emergent, the database was constructed from species names listed in PHI-base and the DEFRA plant health risk register. Reliance on the DEFRA register ensured the inclusion of high-risk and emergent species of concern, but this source is not necessarily the most relevant for growers, as it excludes many widespread pathogens that are routinely managed on a crop-by-crop basis. To address this limitation, PHI-base was also incorporated to broaden coverage. However, PHI-base entries lack the detailed presence and risk information available in DEFRA, and it was not feasible within the scope of this study to verify each species individually. By combining both sources, the pipeline can in principle be adapted to different use cases, whether focused on surveillance for novel or high-risk introductions, or on broader monitoring of well-established pathogens.

Once the species list was created a single reference genome was selected per species according to a defined set of rules: preference was given to RefSeq "reference" or "representative" genomes, followed by the highest available assembly level, and finally the longest and most recent assembly. While this approach ensured consistency and inclusion of as many target species as possible, it also introduces several limitations.

Relying on a single reference genome per species inevitably sacrifices strain-level specificity, which may be important for understanding pathogen diversity. Furthermore, the inclusion of assemblies at contig or scaffold level increases the risk of misassignments due to incomplete or contaminated references. On the other hand, excluding such assemblies would prevent the detection of many species of interest altogether. A potential improvement to the pipeline may therefore be inclusion of all these reference genomes, but with an included genome evenness metric in the output per species for downstream filtering.

A further limitation concerns the calculation of normalised abundance (HP100k) after taxonomic assignment. This requires defining the total read number against which species counts are scaled. In this study, normalisation was based on the total number of reads passing sequencing quality filters, but prior to removal of reads shorter than 300 bp. The pipeline also provides an alternative normalisation using only reads that pass the length filter. Using the pre-length filter total offers greater comparability between samples but may overestimate the amount of usable data, whereas the post-length filter total better reflects the analysed dataset but can disproportionately penalise samples with shorter read distributions. Pre-filter normalisation was selected here for consistency across datasets, though it is important to acknowledge that the choice of approach can influence comparative abundance estimates.

A final limitation concerns the confidence that can be placed in species-level detections. Several filters were applied in this study to reduce erroneous assignments, including thresholds for identity, alignment length, and a minimum normalised read count per taxon. Nevertheless, incorrect assignments to closely related species remain possible, particularly given the relatively narrow reference database restricted to pathogens. The *LCAParse* algorithm was introduced to mitigate this by assigning reads to higher taxonomic levels

when species-level resolution was uncertain. However, such assignments are of limited value for threat detection and monitoring, where species-level information is essential. As noted earlier, further validation could be achieved by incorporating genome coverage and evenness, since reads mapping evenly across the genome would provide greater confidence than those which cluster within conserved regions. Ultimately, when providing information to growers or issuing alerts for high-risk pathogens, verified species- or even strain-level identifications are required, particularly if management decisions are to be based on these findings.

7.7 Conclusion and Future Work

Pathogen mining of the AirSeq experimental datasets using MARMoT revealed a broad range of taxa, most classified as low risk, some listed as absent from the UK, and only a small number considered high risk. These high-risk detections must be interpreted with caution, as alternative explanations suggest that some may represent false positives. The pipeline was iteratively developed and its parameters optimised to balance sensitivity and specificity, but further refinement and validation are needed to strengthen confidence in species-level assignments. Nonetheless, the detection of *H. fraxineus*, the causal agent of ash dieback, in the expected season and location demonstrates the potential of MARMoT for pathogen surveillance. With continued optimisation and external validation, the pipeline could provide a reliable framework for early warning of emerging plant diseases.

Future development should focus on improving interoperability to facilitate wider use, generating user-friendly report style outputs for non-specialists, and further validating filtering thresholds. Validation could be achieved through controlled studies with ground-truth data, for example by sampling enclosed environments with known pathogens of interest such as greenhouses. Complementary in silico analyses using simulated datasets, including high-risk pathogens at varying abundances, would also allow systematic assessment of detection thresholds and overall pipeline performance. The addition of a genome evenness metrics would provide an additional layer of validation, helping to distinguish true positive detections from artefacts.

Ultimately, once MARMoT has been shown to operate reliably under both experimental and simulated conditions, it could be integrated into routine surveillance frameworks or applied to monitor broader environments. Such applications would support both the detection of established pathogens and the early identification of novel or emergent threats.

Chapter 8

Discussion

8.1 Overview and Chapter Summaries

The scope of this PhD was to determine the usability of the AirSeq protocol to reliably capture and detect airborne plant pathogens. In order to test, validate and optimise the methodology AirSeq was used in an enclosed greenhouse environment (Chapter 3), then field and lab refinement was carried out to enhance the method (Chapter 4). Following refinement and validation the protocol was used to monitor airborne plant pathogens both over the course of a full season (Chapter 5) and in shorter snapshot samples over 24-hour periods (Chapter 6). Additionally, a custom pipeline was developed for the analysis of airborne eDNA sequence data in order to detect the presence of high risk or potentially emergent pathogens (Chapter 7).

8.1.1 Summary of research chapters

Chapter 3 Preliminary AirSeq methodology was used over the course of a year to monitor three key strawberry pathogens. The sequencing results showed strong correlations with manual disease scores, indicating that AirSeq can track pathogen incidence in line with observed disease dynamics. Crucially, AirSeq could detect high pathogen abundance in the air long before damage was observed on crops. Fungicide applications did not appear to reduce pathogen prevalence in the air or on plants, suggesting potential resistance development. This chapter demonstrated the potential of AirSeq as a reliable surveillance tool for crop pathogens in a commercial horticultural setting.

Chapter 4 Field and laboratory experiments were used to refine the AirSeq protocol. The InnovaPrep Cub was identified as the most suitable sampler, controls were confirmed as essential, while filter storage at –80°C for three weeks was shown to reduce DNA yields, though not catastrophically, and a 20-second bead-beating step proved optimal for extraction. These refinements provided a more robust protocol for future AirSeq experiments. Additionally, a maize pollen dispersal study highlighted the influence of distance and environmental conditions on detection rates.

Chapter 5 The refined AirSeq method was used in an untreated wheat field over two growing seasons to track airborne pathogens. Weekly sampling revealed community-level dynamics and enabled detailed profiling of key fungal and oomycete genera. Pathogen detections often coincided with favourable environmental conditions and, in some cases,

preceded visible symptoms in the field, highlighting the potential of AirSeq for early warning and disease forecasting.

Chapter 6 AirSeq was used to characterise airborne community dynamics over 24-hour periods. Comparisons of short (2-hour) and long (up to 24-hour) collections showed clear daily and seasonal fluctuations, with shorter samples sometimes detecting transient taxa and pathogen peaks. The study showed that both sampling duration and timing influence results, and that the air microbiome fluctuates over short and longer timescales.

Chapter 7 A custom pipeline, MARMoT, was developed to analyse AirSeq datasets using a pathogen-focused reference database. Multiple datasets from this thesis and beyond were run through the pipeline, showing that MARMoT could detect a wide range of pathogens and recover seasonal patterns consistent with known dynamics. While *Hymenoscyphus fraxineus*, the causal agent of ash dieback, was reliably detected. The detection of other high-risk species was associated with some uncertainty, indicating the need for further refinement before species-level alerts can be considered robust. Nonetheless, flagged detections could be used to prioritise more targeted investigations of collection areas.

In summary Across these studies, the AirSeq protocol was shown to reliably detect airborne pathogens, with peaks in abundance often aligning with disease incidence and environmental conditions conducive to infection. The results demonstrate that the method can characterise dynamic pathogen communities across temporal scales and environments. A consistent finding was the complexity of the airborne microbiome, with a vast and highly fluctuating taxonomic composition. Overall, AirSeq shows clear promise for pathogen surveillance, but further work is required before the data can be routinely translated into actionable crop protection measures.

8.2 Evaluation of Findings in Relation to Research Objectives

The objectives of this PhD were to optimise and validate the AirSeq pipeline, from airborne sample capture through to wet-lab and bioinformatic processes, in order to establish its reliability for detecting airborne fungal plant pathogens. A further objective was to assess the extent to which the method could reveal emergent pathogens not currently infecting UK crops but present in the air.

8.2.1 Optimisation and validation of the AirSeq pipeline

The first objective was addressed comprehensively across all stages of the AirSeq pipeline, though further optimisation will always be possible as technologies evolve.

Sample collection was refined through a sampler comparison experiment, which identified the InnovaPrep Cub as optimal for portability, cost-effectiveness, DNA yield, and diversity of captured taxa. Additional experiments demonstrated that sampling duration and timing should be tailored to experimental aims, and that detection of airborne eDNA is strongly influenced by distance from source.

Several wet-lab processes were also optimised. The filter storage experiment showed that delayed DNA extraction reduces yield, reinforcing the need for immediate processing. Bead-beating trials using a mock community indicated that a 20-second duration was sufficient to lyse a broad range of taxa while retaining DNA length. Negative controls confirmed low-level contamination, while positive lambda controls revealed barcode cross-talk during sequencing. Consequently, blank filter controls and positive controls were incorporated into all subsequent experiments.

Bioinformatic optimisation culminated in the development of the MARMoT pipeline, designed to detect pathogens in ONT datasets from airborne samples. Pipeline parameters were refined using real and simulated data, and validation across multiple datasets demonstrated its capacity to detect a broad range of airborne pathogens.

Additionally, the results from AirSeq were validated by comparison with disease observations. For strawberry pathogens, fortnightly disease scores were compared with monthly air samples, showing correspondence between airborne relative abundance and disease incidence. Validation continued at Church Farm, where airborne pathogen abundance was assessed against pathogen biology, conducive environmental conditions, informal field observations, and end-of-season disease scores.

There remain several areas where further optimisation of the AirSeq pipeline could be pursued. Future work could include testing newly developed air samplers as they become available, as well as systematic comparisons of additional sampling durations. In the laboratory workflow, alternative DNA extraction kits or lysis solutions could be evaluated, and bead-beating optimisation could be extended to include variation in speed as well as duration. Finally, the bioinformatic analysis could be benchmarked against additional commonly used software pipelines to assess performance and robustness.

8.2.2 Identification of emergent airborne plant pathogens

The second objective was to evaluate the potential of AirSeq for identifying emergent pathogens that are not yet infecting UK crops but may be present in the air. Achieving this is inherently challenging: the pool of possible emergent pathogens is extremely large, making it unfeasible to generate a comprehensive target list. As a pragmatic alternative, the DEFRA risk register was employed, as it includes pathogens considered a threat to UK food production, including those not yet detected within the country. Nonetheless, this approach inevitably excludes taxa not recognised as risks or not yet characterised. Detection is further constrained by the absence of reference genomes, which precludes the identification of many potentially emergent species. Moreover, emergent pathogens are likely to occur at very low abundance, limiting the ability to distinguish genuine detections from artefacts and increasing the risk that they are excluded during filtering.

Across these experiments, no emergent pathogens were reliably detected. However, this absence does not confirm that such pathogens are not present in the UK. Non-detection may reflect the specific locations sampled or the low airborne abundance of emergent taxa. Given that new pathogens are often introduced via trade, nursery plants, or wind-dispersed spores, targeted sampling in high-risk environments such as nurseries, ports, or coastal sites may increase the likelihood of detection. At present, AirSeq predominantly identifies established UK pathogens, likely reflecting their higher airborne inoculum levels. Although detections at Church Farm demonstrated the sensitivity of the method by capturing some

pathogens at very low abundance prior to symptom development, the detection of rare, wind-borne spores from non-established species would be considerably more challenging. This may necessitate denser sampler networks, and even then, verification of rare detections would remain difficult.

In summary, the objectives of this PhD were met to a substantial degree. The AirSeq pipeline was successfully optimised and validated across sampling, laboratory, and bioinformatic stages. The secondary objective of identifying emergent pathogens proved more challenging, with no reliable detections achieved. Together, these findings demonstrate that while AirSeq has considerable promise as a surveillance tool, further development is required before it can be fully deployed for early detection of novel threats.

8.3 Project Challenges and Limitations

Across the experiments in this thesis, several consistent challenges and limitations were encountered. These centred on the taxonomic assignment and interpretation of sequence data, including dependence on the chosen reference database, the absence of ground-truth validation, taxa counts being relative rather than absolute, the risk of false positives and negatives and uncertainty in species-level resolution.

The choice of database is critical in alignment-based analyses, since species can only be detected if they are represented. When an organism is missing, reads may be misassigned to a closely related taxon, resulting in false positives. Previous studies have shown that community composition can vary markedly depending on the database employed [346]. The experiments in this thesis used either the BLAST nt database or a custom pathogen database. In both cases, a substantial portion of microbial diversity is absent, as many taxa are unculturable and therefore lack reference genomes. Furthermore, errors within reference data, including misannotations, low-quality assemblies, or contamination, may further distort results. Consequently, there are a large proportion of reads that were either unassigned or assigned to higher taxonomic levels than species throughout this thesis.

Another limitation of this study is the absence of ground-truth data against which to validate airborne eDNA detections. Without such validation it is difficult to establish whether observed taxa represent genuine biological presence or artefacts introduced during sampling, sequencing, or classification. This challenge is further compounded by the novelty of airborne sequencing, which limits the availability of comparable datasets and reduces confidence in the detection of rare or unexpected taxa. In this study, disease observations and environmental or seasonal factors known to influence fungal abundance were used as indirect proxies to assess plausibility. However, these indicators are not always reliable: spores may be abundant in the air without visible disease symptoms if host plants are resistant, or conditions may favour germination even when airborne spores are absent. Future work could address this limitation through controlled experiments involving the release of specific spores in enclosed environments, as demonstrated by [124]. However, such approaches do not fully capture the highly diverse and dynamic nature of airborne eDNA in real-world settings.

A further limitation of this work is that the metagenomic datasets are compositional, reflecting relative rather than absolute abundance [126]. As such, apparent shifts in the abundance of particular taxa may not indicate genuine increases or decreases, but instead reflect changes in the wider community. This was especially evident in the seasonal studies,

where the release of pollen coincided with fluctuations in relative taxon abundance, complicating interpretation of pathogen dynamics. The challenge of compositional data could be partially mitigated through the use of synthetic DNA spike-ins at defined concentrations, which have been applied to airborne studies to provide semi-quantitative estimates of fungal DNA abundance [274].

An additional challenge with metagenomic data is the risk of both false positives and false negatives. False positives occur when a taxon is incorrectly identified as present, which in an agricultural context could result in unnecessary fungicide applications or, more concerningly, the application of inappropriate control measures if the organism has been misclassified. False negatives are arguably more problematic, as they create the impression that a pathogen is not present; in practice, this could leave crops unprotected and highly vulnerable to infection. These risks highlight the need for careful interpretation of AirSeq outputs and suggest that, at present, airborne metagenomic monitoring should be viewed as a complementary tool within wider disease surveillance frameworks, rather than as a stand-alone diagnostic. Further research could extend the parameter optimisation in Chapter 7 using more complex simulated datasets to refine detection thresholds and improve confidence in taxonomic identifications.

Alongside the risk of false positives and false negatives, there remains the question of whether species or even strain level identifications can be achieved from airborne eDNA, which often yields very low amounts of DNA. Such resolution is particularly important in pathogen detection, where genera can contain benign and virulent species, or where strain level differences determine host specificity in relation to crops of interest. Achieving this resolution is challenging due to conserved genomes among closely related taxa and incomplete representation in reference databases. As demonstrated in Chapter 7, the proportion of genome coverage for identified species was typically very low, limiting the reliability of species level assignments. In cases where specific pathogens of concern are suspected, one possible solution would be the use of targeted primers capable of distinguishing between strains, as described by [301]. This approach could be combined with WGS to validate initial detections, providing a more targeted strategy once metagenomic analysis has identified potential pathogens.

Bringing these challenges together, it becomes clear that the absence or detection of a pathogen in airborne eDNA does not directly translate to crops being safe or at risk. Reliable detection depends on multiple factors, including the completeness of the reference database, the robustness of the analytical method, the concentration of the pathogen in the air, and the ability to resolve identifications to the appropriate species or strain level. As a whole, the results highlight both the promise of airborne pathogen monitoring and the need for further validation and methodological refinement before it can be fully integrated into routine surveillance frameworks.

8.4 Advancing AirSeq Towards Practical Deployment

Although this thesis has advanced understanding of airborne detection of plant pathogens, several key challenges remain before AirSeq can become a routine tool to assist disease management. Bridging the gap between taxonomic detections and actionable risk assessments will require further experimental work, particularly to address questions of spore viability, turnaround time for on-farm decision-making, and the need for strain level resolution to

capture host specificity.

Viability of detected pathogens is crucial, as crops are only at risk if airborne spores remain capable of germination and infection. One possible approach would be to culture material directly from air samples, though this is limited by the fact that many fungi are not readily culturable [244] and competitive taxa may dominate mixed samples. Alternative strategies include leveraging sequencing-based approaches, such as exploring rRNA-to-DNA ratios as a proxy for metabolic activity [127], or developing nanopore signal ("squiggle") analysis to distinguish between viable and non-viable DNA [373]. While these remain technically challenging and, in the case of nanopore-based methods, at an early stage of development, such experiments could ultimately provide critical insight into which airborne detections represent true epidemiological threats.

Another valuable experiment would be to compare airborne eDNA with DNA from infected plant tissue. This would provide greater certainty that the taxa detected in the air are the same as those responsible for disease symptoms. Preliminary work in Chapter 3 began to address this, but the approach could be scaled up with additional samples and longitudinal comparisons to track whether dominant airborne taxa correspond to those infecting plants over time. Such experiments would strengthen confidence in AirSeq as a tool for anticipating outbreaks by detecting specific pathogens before visible symptoms emerge.

A further area for investigation is determining at what stage in the fungal infection cycle airborne spores can be reliably detected. Detecting primary inoculum would be particularly valuable, as this stage precedes symptom development and secondary spread, offering the best chance for early intervention. However, spore concentrations are likely to be lower at this stage, making detection more difficult, and the window of opportunity for sampling may be short. Controlled environment experiments, where the timing and source of inoculum are known, could provide insights into how well AirSeq captures primary infections and whether targeted sampling strategies could improve detection during this critical phase.

A related knowledge gap concerns the spore abundance thresholds and spatial scales relevant to airborne monitoring. In particular, it remains unclear what concentration of airborne spores is required to trigger an epidemic. This is important as it would be helpful to provide growers both with an idea of what pathogens are present but also which are the most likely to cause infections and therefore where intervention is required. Additionally, although considered with the distance-from-maize source experiment in Chapter 4, there is still much to be learnt about the distances from which a sampler can detect airborne pathogens and how this is affected by environmental factors. Increased knowledge of the detection distance of AirSeq would be helpful in establishing the placement of samplers across agricultural fields to achieve reliable coverage. Controlled experiments introducing a known quantity of inoculum, combined with disease monitoring and spatially distributed sampling, could provide insights into how AirSeq detections relate to disease risk in the field.

Finally, there are technological developments that could further enhance the AirSeq pipeline. A key priority is reducing the time from sample collection to result, as growers often need to respond rapidly to protect crops following pathogen detection. In this thesis, samples were processed in bulk and sequenced together to improve efficiency and reduce costs, but for future applications an automated in-field workflow would be preferable.

Such a system could employ robotics and microfluidics to transfer samples, extract DNA, and prepare libraries for sequencing before transmitting results for analysis. Comparable approaches already exist, for example portable PCR assays for plant pathogens [375] and laser-based pollen detection systems [71]. Recent advances in nanopore sequencing, such as the ElysION platform and the forthcoming TraxION device, further suggest that fully automated, in-field pathogen detection is becoming increasingly feasible. However, the costs of such systems are currently likely to be prohibitively high for most crops, and early adoption may therefore be limited to high-value horticultural or protected cropping systems. Overcoming these barriers would be a crucial step towards making AirSeq a practical component of disease surveillance.

8.5 How AirSeq Can Be Used In The Future

This thesis has demonstrated the potential of AirSeq as a tool for disease surveillance, capable of generating detailed information on airborne plant pathogens. Although a number of limitations remain to be addressed, the findings presented here highlight the considerable promise of AirSeq, both within agriculture and in wider applications for monitoring airborne eDNA communities.

Although AirSeq has so far only been tested at a local scale, in glasshouses and open wheat fields, the approach has considerable potential to be scaled up. Networks of spore samplers could provide a faster and more efficient means of conducting large-scale pathogen monitoring, reducing reliance on labour-intensive surveys. Such networks could operate at regional, national, or even global levels, with the resulting data integrated with meteorological information to advance understanding of spore transport and improve the accuracy of disease forecasting models. Moreover, a wider sampling network would enhance the capacity to detect emergent pathogens and trace their geographical origins.

All of these applications could be further enhanced by the delivery of real-time results. AirSeq already makes use of ONT sequencing technology, which generates output during the sequencing process, and with the addition of automation described earlier it may be possible to obtain results from airborne sampling within only a few hours. Such rapid turnaround would allow growers to respond swiftly to threats, and when implemented within a network of samplers could provide early warning of incoming inoculum before it reaches the crop.

Another exciting possibility lies in the use of WGS, which, unlike amplicon sequencing, captures the full genome of airborne taxa. This provides the opportunity to identify additional information of practical value, such as fungicide resistance genes, which could guide the targeted application of effective chemicals and reduce unnecessary inputs. Furthermore, WGS data could be employed to track the movement of specific strains across regions, offering valuable insights into how pathogenic lineages spread between agricultural sites.

As demonstrated in the data presented here, AirSeq detects a broad range of taxa beyond fungi, and therefore has potential applications outside agriculture, including biodiversity monitoring, human health, and atmospheric science.

In terms of biodiversity monitoring, previous studies have already demonstrated the capacity of air sampling to detect vertebrates [80, 228]. With longer sampling durations and optimised extraction protocols, AirSeq could similarly be applied to identify and charac-

terise biodiversity from airborne material. Such an approach would be particularly valuable in hard-to-access environments, offering a non-invasive means of surveying species presence and enabling large-scale monitoring without the need for extensive visual surveys.

Within human health contexts, AirSeq could be adapted to detect pathogens relevant to humans as well as plants. Air sampling has already been used to establish microbiome baselines [197] and to identify infection hotspots [42, 210, 305]. AirSeq could be applied in a similar way, with the WGS approach enabling simultaneous monitoring of fungi, bacteria, and DNA viruses, and with protocol modifications allowing the inclusion of RNA viruses. In addition, AirSeq data could be leveraged to identify antimicrobial resistance genes in airborne pathogens, supporting infection control strategies such as the maintenance of clean rooms for immunocompromised patients.

Finally, in addition to sampling at ground or canopy level, AirSeq could be taken to new heights by attaching samplers to planes or drones, or deployed on ships to sample over the oceans. This would enable investigation of airborne microbial communities at different heights above sea level and across varying distances from land. While this thesis has shown broad temporal fluctuations in airborne communities, expanding sampling across spatial gradients would generate new insights into how these communities vary and disperse. Such investigations could improve our understanding of long-distance particle dispersion and reveal more about how microorganisms interact with the atmosphere, for example through processes such as ice nucleation [51], thereby influencing broader ecological and climatic dynamics.

In conclusion AirSeq can accurately detect airborne plant pathogens, with changes in abundance often correlating with disease scores and environmental conditions. This thesis has primarily focused on improving and validating the AirSeq pipeline to ensure the reliability of taxonomic assignments, while also generating new insights into pathogen presence in UK agricultural fields. To translate AirSeq into a practical tool for disease monitoring, however, further work is needed to understand how airborne detections relate to infection risk, and to test whether using AirSeq data to guide crop protection improves outcomes in practice. With such refinements and additional experimentation, AirSeq has the potential to become a cornerstone of plant disease surveillance and to find valuable applications beyond agriculture.

Bibliography

[1] Araz S. Abdullah et al. "Host–Multi-Pathogen Warfare: Pathogen Interactions in Co-infected Plants". In: *Frontiers in Plant Science* 8 (Oct. 2017). DOI: 10.3389/fpls.2017.01806.

[2] A. M. Hasanthi Abeykoon et al. "Performance Evaluation and Validation of Air Samplers To Detect Aerosolized Coxiella burnetii". In: *Microbiology Spectrum* 10.5 (Sept. 2022), e00655–22. DOI: 10.1128/spectrum.00655-22.

[3] Nerea Abrego et al. "Give me a sample of air and I will tell which species are found from your region: Molecular identification of fungi from airborne spore samples". In: *Molecular Ecology Resources* 18.3 (2018), pp. 511–524. DOI: 10.1111/1755-0998.12755.

[4] Nerea Abrego et al. "Airborne DNA reveals predictable spatial and seasonal dynamics of fungi". In: *Nature* 631.8022 (July 2024), pp. 835–842. DOI: 10.1038/s41586-024-07658-9.

[5] Eduardo Abreo and Nora Altier. "Pangenome of Serratia marcescens strains from nosocomial and environmental origins reveals different populations and the links between them". In: *Scientific Reports* 9 (Jan. 2019), p. 46. DOI: 10.1038/s41598-018-37118-0.

[6] Igor E. Agranovski and Evgeny V. Usachev. "In-situ rapid bioaerosol detection in the ambient air by miniature multiplex PCR utilizing technique". In: *Atmospheric Environment* 246 (Feb. 2021), p. 118147. DOI: 10.1016/j.atmosenv.2020.118147.

[7] Jaime Aguayo et al. "Combining permanent aerobiological networks and molecular analyses for large-scale surveillance of forest fungal pathogens: A proof-of-concept". In: *Plant Pathology* 70.1 (2021), pp. 181–194. DOI: 10.1111/ppa.13265.

[8] Stephen C. Alderman, Darrin L. Walenta, and Philip B. Hamm. "Timing of Occurrence of *Claviceps purpurea* Ascospores in Northeast Oregon". In: *Plant Health Progress* 11.1 (Jan. 2010), p. 2. DOI: 10.1094/PHP-2010-1123-01-RS.

[9] Anna Aldrighetti and Ilaria Pertot. "Epidemiology and control of strawberry powdery mildew: a review". In: *Phytopathologia Mediterranea* 62.3 (Dec. 2023), pp. 427–453. DOI: 10.36253/phyto-14576.

[10] Rohia Alili et al. "Exploring Semi-Quantitative Metagenomic Studies Using Oxford Nanopore Sequencing: A Computational and Experimental Protocol". In: *Genes* 12.10 (Sept. 2021), p. 1496. DOI: 10.3390/genes12101496.

[11] Johannes Alneberg et al. "Binning metagenomic contigs by coverage and composition". In: *Nature Methods* 11.11 (Nov. 2014), pp. 1144–1146. DOI: 10.1038/nmeth.3103.

[12] Malin Alsved et al. “Sources of Airborne Norovirus in Hospital Outbreaks”. In: *Clinical Infectious Diseases* 70.10 (May 2020), pp. 2023–2028. DOI: 10.1093/cid/ciz584.

[13] Stephen F. Altschul et al. “Basic local alignment search tool”. In: *Journal of Molecular Biology* 215.3 (Oct. 1990), pp. 403–410. DOI: 10.1016/S0022-2836(05)80360-2.

[14] Achour Amiri and Natalia A. Peres. “Diversity in the erg27 Gene of Botrytis cinerea Field Isolates from Strawberry Defines Different Levels of Resistance to the Hydroxyanilide Fenhexamid”. In: *Plant Disease* 98.8 (Aug. 2014), pp. 1131–1137. DOI: 10.1094/PDIS-11-13-1171-RE.

[15] Achour Amiri, Adrian I. Zuniga, and Natalia A. Peres. “Prevalence of Botrytis Cryptic Species in Strawberry Nursery Transplants and Strawberry and Blueberry Commercial Fields in the Eastern United States”. In: *Plant Disease* 102.2 (Feb. 2018), pp. 398–404. DOI: 10.1094/PDIS-07-17-1065-RE.

[16] Liat Amsalem et al. “Effect of Climatic Factors on Powdery Mildew Caused by Sphaerotheca macularis f. sp. Fragariae on Strawberry”. In: *European Journal of Plant Pathology* 114.3 (Mar. 2006), pp. 283–292. DOI: 10.1007/s10658-005-5804-6.

[17] Choa An, Cheolwoon Woo, and Naomichi Yamamoto. “Introducing DNA-based methods to compare fungal microbiota and concentrations in indoor, outdoor, and personal air”. In: *Aerobiologia* 34.1 (Mar. 2018), pp. 1–12. DOI: 10.1007/s10453-017-9490-6.

[18] Godfrey Philliam Apangu et al. “Environmental DNA reveals diversity and abundance of Alternaria species in neighbouring heterogeneous landscapes in Worcester, UK”. In: *Aerobiologia* 38.4 (Dec. 2022), pp. 457–481. DOI: 10.1007/s10453-022-09760-9.

[19] Sara Araújo et al. “From soil to surface water: exploring Klebsiella ’s clonal lineages and antibiotic resistance odyssey in environmental health”. In: *BMC Microbiology* 25 (Feb. 2025), p. 97. DOI: 10.1186/s12866-025-03798-8.

[20] Ravinder Arigela et al. “Effect of relative humidity on passive spore release from substrate surfaces”. In: *Journal of Aerosol Science* 183 (Jan. 2025), p. 106477. DOI: 10.1016/j.jaerosci.2024.106477.

[21] Raymond W. Arritt et al. “Lagrangian numerical simulations of canopy air flow effects on maize pollen dispersal”. In: *Field Crops Research* 102.2 (June 2007), pp. 151–162. DOI: 10.1016/j.fcr.2007.03.008.

[22] George K. Auer and Douglas B. Weibel. “Bacterial Cell Mechanics”. In: *Biochemistry* 56.29 (July 2017), pp. 3710–3724. DOI: 10.1021/acs.biochem.7b00346.

[23] Martin Ayling, Matthew D Clark, and Richard M Leggett. “New approaches for metagenome assembly with short reads”. In: *Briefings in Bioinformatics* 21.2 (Mar. 2020), pp. 584–594. DOI: 10.1093/bib/bbz020.

[24] Donald E. Aylor, Neil P. Schultes, and Elson J. Shields. “An aerobiological framework for assessing cross-pollination in maize”. In: *Agricultural and Forest Meteorology* 119.3 (Nov. 2003), pp. 111–129. DOI: 10.1016/S0168-1923(03)00159-X.

[25] Juliana S. Baggio, Marcus V. Marin, and Natalia A. Peres. "Phytophthora Crown Rot of Florida Strawberry: Inoculum Sources and Thermotherapy of Transplants for Disease Management". In: *Plant Disease* 105.11 (Nov. 2021), pp. 3496–3502. DOI: 10.1094/PDIS-11-20-2476-RE.

[26] Matthew G. Bakker. "A fungal mock community control for amplicon sequencing experiments". In: *Molecular Ecology Resources* 18.3 (2018), pp. 541–556. DOI: 10.1111/1755-0998.12760.

[27] Elisa Banchi, Alberto Pallavicini, and Lucia Muggia. "Relevance of plant and fungal DNA metabarcoding in aerobiology". In: *Aerobiologia* 36.1 (Mar. 2020), pp. 9–23. DOI: 10.1007/s10453-019-09574-2.

[28] Elisa Banchi et al. "DNA metabarcoding uncovers fungal diversity of mixed airborne samples in Italy". In: *PLOS ONE* 13.3 (Mar. 2018), e0194489. DOI: 10.1371/journal.pone.0194489.

[29] Elisa Banchi et al. "Environmental DNA assessment of airborne plant and fungal seasonal diversity". In: *Science of The Total Environment* 738 (Oct. 2020), p. 140249. DOI: 10.1016/j.scitotenv.2020.140249.

[30] Daniel G. Barber et al. "Statistical design approach enables optimised mechanical lysis for enhanced long-read soil metagenomics". In: *Scientific Reports* 14.1 (Nov. 2024), p. 28934. DOI: 10.1038/s41598-024-80584-y.

[31] C. C. V. Batts and Ann Jeater. "The development of loose smut (*Ustilago tritici*) in susceptible varieties of wheat, and some observations on field infection". In: *Transactions of the British Mycological Society* 41.1 (Mar. 1958), 115–IN8. DOI: 10.1016/S0007-1536(58)80015-7.

[32] L. Beal et al. "First report of Phytophthora tentaculata affecting Santolina in the UK". In: *New Disease Reports* 37.1 (2018), pp. 8–8. DOI: 10.5197/j.2044-0588.2018.037.008.

[33] Edward W. Beals. "Bray-Curtis Ordination: An Effective Strategy for Analysis of Multivariate Ecological Data". In: *Advances in Ecological Research*. Ed. by A. MacFadyen and E. D. Ford. Vol. 14. Academic Press, Jan. 1984, pp. 1–55.

[34] Johan Bengtsson-Palme et al. "Metagenomics reveals that detoxification systems are underrepresented in marine bacterial communities". In: *BMC Genomics* 15.1 (Sept. 2014), p. 749. DOI: 10.1186/1471-2164-15-749.

[35] Gaëtan Benoit et al. "High-quality metagenome assembly from long accurate reads with metaMDBG". In: *Nature Biotechnology* 42.9 (Sept. 2024), pp. 1378–1383. DOI: 10.1038/s41587-023-01983-6.

[36] Mia F. G. Berelson et al. "From air to insight: the evolution of airborne DNA sequencing technologies". In: *Microbiology* 171.5 (2025), p. 001564. DOI: 10.1099/mic.0.001564.

[37] Angela Berrie. *Control of grey mould in strawberry crops | AHDB*. Tech. rep. 18/04. East Malling Research: Horticultural Development Council (HDC), Dec. 2004.

[38] Valentina Bertolini et al. "Temporal variability and effect of environmental variables on airborne bacterial communities in an urban area of Northern Italy". In: *Applied Microbiology and Biotechnology* 97.14 (July 2013), pp. 6561–6570. DOI: 10.1007/s00253-012-4450-0.

[39] Dhan Bhandari and Simon Edwards. *Risk assessment for fusarium mycotoxins in wheat | AHDB*. Tech. rep. (Accessed 31/07/25). AHDB Cereals & Oilseeds, 2023.

[40] Richa Bharti and Dominik G Grimm. "Current challenges and best-practice protocols for microbiome analysis". In: *Briefings in Bioinformatics* 22.1 (Jan. 2021), pp. 178–193. DOI: 10.1093/bib/bbz155.

[41] Anil Bhujel et al. "Detection of gray mold disease and its severity on strawberry using deep learning networks". In: *Journal of Plant Diseases and Protection* 129.3 (June 2022), pp. 579–592. DOI: 10.1007/s41348-022-00578-8.

[42] Gabriel Birgand et al. "Assessment of Air Contamination by SARS-CoV-2 in Hospital Settings". In: *JAMA Network Open* 3.12 (Dec. 2020), e2033232. DOI: 10.1001/jamanetworkopen.2020.33232.

[43] C. Blanco et al. "Relationship Among Concentrations of Sphaerotheca macularis Conidia in the Air, Environmental Conditions, and the Incidence of Powdery Mildew in Strawberry". In: *Plant Disease* 88.8 (Aug. 2004), pp. 878–881. DOI: 10.1094/PDIS.2004.88.8.878.

[44] Cesar Blanco, Berta los de Santos, and Fernando Romero. "Relationship between Concentrations of Botrytis Cinerea Conidia in Air, Environmental Conditions, and the Incidence of Grey Mould in Strawberry Flowers and Fruits". In: *European Journal of Plant Pathology* 114.4 (Apr. 2006), pp. 415–425. DOI: 10.1007/s10658-006-0007-3.

[45] Aitor Blanco-Miguez et al. "Extending and improving metagenomic taxonomic profiling with uncharacterized species with MetaPhlAn 4". In: *bioRxiv* (Aug. 2022). DOI: 10.1101/2022.08.22.504593.

[46] Pascale Bodevin. *Wheat Blast: Earth observation and climate forecasts for risk management*. Tech. rep. (Accessed 31/07/25). CABI, Mar. 2025.

[47] Kari Oline Bøifot et al. "Performance evaluation of a new custom, multi-component DNA isolation method optimized for use in shotgun metagenomic sequencing-based aerosol microbiome research". In: *Environmental Microbiome* 15.1 (Jan. 2020), p. 1. DOI: 10.1186/s40793-019-0349-z.

[48] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data". In: *Bioinformatics* 30.15 (Aug. 2014), pp. 2114–2120. DOI: 10.1093/bioinformatics/btu170.

[49] Melvin D. Bolton, James A. Kolmer, and David F. Garvin. "Wheat leaf rust caused by Puccinia triticina". In: *Molecular Plant Pathology* 9.5 (May 2008), pp. 563–575. DOI: 10.1111/j.1364-3703.2008.00487.x.

[50] Hendriek C Boshuizen and Dennis E te Beest. "Pitfalls in the statistical analysis of microbiome amplicon sequencing data". In: *Molecular Ecology Resources* 23.3 (2023), pp. 539–548. DOI: 10.1111/1755-0998.13730.

[51] Robert M. Bowers et al. “Characterization of Airborne Microbial Communities at a High-Elevation Site and Their Potential To Act as Atmospheric Ice Nuclei”. In: *Applied and Environmental Microbiology* 75.15 (Aug. 2009), pp. 5121–5130. DOI: 10.1128/AEM.00447-09.

[52] Robert M. Bowers et al. “Seasonal variability in airborne bacterial communities at a high-elevation site”. In: *Atmospheric Environment* 50 (Apr. 2012), pp. 41–49. DOI: 10.1016/j.atmosenv.2012.01.005.

[53] Vendula Brabcová et al. “Fungal Community Development in Decomposing Fine Deadwood Is Largely Affected by Microclimate”. In: *Frontiers in Microbiology* 13 (Apr. 2022). DOI: 10.3389/fmicb.2022.835274.

[54] C. J. Brennan et al. “A review of the known unknowns in the early stages of septoria tritici blotch disease of wheat”. In: *Plant Pathology* 68.8 (2019), pp. 1427–1438. DOI: 10.1111/ppa.13077.

[55] Eoin L. Brodie et al. “Urban aerosols harbor diverse and dynamic bacterial populations”. In: *Proceedings of the National Academy of Sciences* 104.1 (Jan. 2007), pp. 299–304. DOI: 10.1073/pnas.0608255104.

[56] J. Paul Brooks et al. “The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies”. In: *BMC Microbiology* 15.1 (Mar. 2015), p. 66. DOI: 10.1186/s12866-015-0351-6.

[57] *Brown rust in cereals | AHDB*. (Accessed 31/07/25). URL: https://ahdb.org.uk/brownrust.

[58] M. Burch and E. Levetin. “Effects of meteorological conditions on spore plumes”. In: *International Journal of Biometeorology* 46.3 (Aug. 2002), pp. 107–117. DOI: 10.1007/s00484-002-0127-1.

[59] Paul Burns, Volkmar Timmermann, and Jon M. Yearsley. “Meteorological factors associated with the timing and abundance of Hymenoscyphus fraxineus spore release”. In: *International Journal of Biometeorology* 66.3 (Mar. 2022), pp. 493–506. DOI: 10.1007/s00484-021-02211-z.

[60] Jeroen Buters et al. “Automatic detection of airborne pollen: an overview”. In: *Aerobiologia* (July 2022). DOI: 10.1007/s10453-022-09750-x.

[61] J. M. Byrne, M. K. Hausbeck, and L. E. Sconyers. “Influence of Environment on Atmospheric Concentrations of Peronospora antirrhini Sporangia in Field-Grown Snapdragon”. In: *Plant Disease* 89.10 (Oct. 2005), pp. 1060–1066. DOI: 10.1094/PD-89-1060.

[62] Benjamin J. Callahan et al. “DADA2: High-resolution sample inference from Illumina amplicon data”. In: *Nature Methods* 13.7 (July 2016), pp. 581–583. DOI: 10.1038/nmeth.3869.

[63] Christian Cando-Dumancela et al. “A guide to minimize contamination issues in microbiome restoration studies”. In: *Restoration Ecology* 29.4 (2021), e13358. DOI: 10.1111/rec.13358.

[64] Yue Cao et al. “Airborne bacterial community diversity, source and function along the Antarctic Coast”. In: *Science of The Total Environment* 765 (Apr. 2021), p. 142700. DOI: 10.1016/j.scitotenv.2020.142700.

[65] O. Carisse et al. “Analysis of Incidence–Severity Relationships for Strawberry Powdery Mildew as Influenced by Cultivar, Cultivar Type, and Production Systems”. In: *Plant Disease* 97.3 (Mar. 2013), pp. 354–362. DOI: 10.1094/PDIS-05-12-0508-RE.

[66] Odile Carisse and Julie Bouchard. “Age-related susceptibility of strawberry leaves and berries to infection by Podosphaera aphanis”. In: *Crop Protection* 29.9 (Sept. 2010), pp. 969–978. DOI: 10.1016/j.cropro.2010.03.008.

[67] Lucía Carrera et al. “Airborne Plasmopara viticola Sporangia: A Study of Vineyards in Two Bioclimatic Regions of Northwestern Spain”. In: *Horticulturae* 11.3 (Mar. 2025), p. 228. DOI: 10.3390/horticulturae11030228.

[68] J. J. Carroll and D. R. Viglierchio. “On the Transport of Nematodes by the Wind”. In: *Journal of Nematology* 13.4 (Oct. 1981), pp. 476–483.

[69] Yuan Chai et al. “Multi-peril pathogen risks to global wheat production: A probabilistic loss and investment assessment”. In: *Frontiers in Plant Science* 13 (Oct. 2022). DOI: 10.3389/fpls.2022.1034600.

[70] Induja Chandrakumar et al. “BugSplit enables genome-resolved metagenomics through highly accurate taxonomic binning of metagenomic assemblies”. In: *Communications Biology* 5 (Feb. 2022), p. 151. DOI: 10.1038/s42003-022-03114-4.

[71] Christel Chappuis et al. “Automatic pollen monitoring: first insights from hourly data”. In: *Aerobiologia* 36.2 (June 2020), pp. 159–170. DOI: 10.1007/s10453-019-09619-6.

[72] Themoula Charalampous et al. “Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection”. In: *Nature Biotechnology* 37.7 (July 2019), pp. 783–792. DOI: 10.1038/s41587-019-0156-5.

[73] Syama Chatterton, Andrew C. Wylie, and Zamir K. Punja. “Fruit infection and postharvest decay of greenhouse tomatoes caused by Penicillium species in British Columbia”. In: *Canadian Journal of Plant Pathology* 34.4 (Oct. 2012), pp. 524–535. DOI: 10.1080/07060661.2012.710069.

[74] Jill C. Check et al. “It’s a Trap! Part I: Exploring the Applications of Rotating-Arm Impaction Samplers in Plant Pathology”. In: *Plant Disease* 108.7 (July 2024), pp. 1910–1922. DOI: 10.1094/PDIS-10-23-2096-FE.

[75] Shifu Chen et al. “fastp: an ultra-fast all-in-one FASTQ preprocessor”. In: *Bioinformatics* 34.17 (Sept. 2018), pp. i884–i890. DOI: 10.1093/bioinformatics/bty560.

[76] Wen Chen et al. “Assessing Performance of Spore Samplers in Monitoring Aeromycobiota and Fungal Plant Pathogen Diversity in Canada”. In: *Applied and Environmental Microbiology* 84.9 (May 2018), e02601–17. DOI: https://doi.org/10.1128/AEM.02601-17.

[77] Wen Chen et al. “Optimizing an integrated biovigilance toolbox to study the spatial distribution and dynamic changes of airborne mycobiota, with a focus on cereal rust fungi in western Canada”. In: *Molecular Ecology Resources* 24.6 (Aug. 2024), e13983. DOI: 10.1111/1755-0998.13983.

[78] Samuel D. Chorlton. “Ten common issues with reference sequence databases and how to mitigate them”. In: *Frontiers in Bioinformatics* 4 (Mar. 2024). DOI: 10.3389/fbinf.2024.1278228.

[79] R. A. Choudhury et al. "Spatiotemporal Patterns in the Airborne Dispersal of Spinach Downy Mildew". In: *Phytopathology®* 107.1 (Jan. 2017), pp. 50–58. DOI: 10.1094/PHYTO-04-16-0162-R.

[80] Elizabeth L. Clare et al. "eDNAir: proof of concept that animal DNA can be collected from air sampling". In: *PeerJ* 9 (Mar. 2021), e11030. DOI: 10.7717/peerj.11030.

[81] Marcus Clauss. "Particle size distribution of airborne micro-organisms in the environment-A review". In: *Landbauforschung Volkenrode* 65 (Oct. 2015), pp. 77–100. DOI: 10.3220/LBF1444216736000.

[82] Nicola M Cook et al. "High frequency of fungicide resistance-associated mutations in the wheat yellow rust pathogen Puccinia striiformis f. sp. tritici". In: *Pest Management Science* 77.7 (2021), pp. 3358–3371. DOI: 10.1002/ps.6380.

[83] *Copernicus: CAMS monitors reoccurring Saharan dust transport across the Atlantic during an extraordinary year of the dust cycle | Copernicus.* (Accessed 15/08/25). URL: https://atmosphere.copernicus.eu/copernicus-cams-monitors-reoccurring-saharan-dust-transport-across-atlantic-during-extraordinary.

[84] Isabel Corkley, Bart Fraaije, and Nichola Hawkins. "Fungicide resistance management: Maximizing the effective life of plant protection products". In: *Plant Pathology* 71.1 (2022), pp. 150–169. DOI: 10.1111/ppa.13467.

[85] Pilar Corredor-Moreno and Diane G. O. Saunders. "Expecting the unexpected: factors influencing the emergence of fungal and oomycete plant pathogens". In: *The New Phytologist* 225.1 (Jan. 2020), pp. 118–125. DOI: 10.1111/nph.16007.

[86] Jonas Henrique Costa et al. "*Penicillium digitatum* infection mechanisms in citrus: What do we know so far?" In: *Fungal Biology*. The Fungal Threat to Food Security 123.8 (Aug. 2019), pp. 584–593. DOI: 10.1016/j.funbio.2019.05.004.

[87] Mateus Cruz et al. "Smart Strawberry Farming Using Edge Computing and IoT". In: *Sensors* 22.15 (Jan. 2022), p. 5866. DOI: 10.3390/s22155866.

[88] Neftaly Cruz-Mireles et al. "The Biology of Invasive Growth by the Rice Blast Fungus Magnaporthe oryzae". In: *Methods in Molecular Biology (Clifton, N.J.)* 2356 (2021), pp. 19–40. DOI: 10.1007/978-1-0716-1613-0_2.

[89] Piotr Cuber et al. "Comparing the accuracy and efficiency of third generation sequencing technologies, Oxford Nanopore Technologies, and Pacific Biosciences, for DNA barcode sequencing applications". In: *Ecological Genetics and Genomics* 28 (Sept. 2023), p. 100181. DOI: 10.1016/j.egg.2023.100181.

[90] Ralph Dean et al. "The Top 10 fungal pathogens in molecular plant pathology". In: *Molecular Plant Pathology* 13.4 (2012), pp. 414–430. DOI: 10.1111/j.1364-3703.2011.00783.x.

[91] Andrea Delgado et al. "Control of White Rot Caused by Sclerotinia sclerotiorum in Strawberry Using Arbuscular Mycorrhizae and Plant-Growth-Promoting Bacteria". In: *Sustainability* 15.4 (Jan. 2023), p. 2901. DOI: 10.3390/su15042901.

[92] Aurelien Dommergue et al. "Methods to Investigate the Global Atmospheric Microbiome". In: *Frontiers in Microbiology* 10 (2019). DOI: 10.3389/fmicb.2019.00243.

[93] Gavin M. Douglas et al. “PICRUSt2 for prediction of metagenome functions”. In: *Nature Biotechnology* 38.6 (June 2020), pp. 685–688. DOI: 10.1038/s41587-020-0548-6.

[94] Daniela I. Drautz-Moses et al. “Vertical stratification of the air microbiome in the lower troposphere”. In: *Proceedings of the National Academy of Sciences* 119.7 (Feb. 2022), e2117293119. DOI: 10.1073/pnas.2117293119.

[95] Brandon Dunbar et al. “Culture-Independent Microbial Air Profiling using a Spaceflight-Compatible Nanopore Sequencing Method”. In: Saint Paul, MN.

[96] Jeremiah K. S. Dung et al. “Spatial Patterns of Ergot and Quantification of Sclerotia in Perennial Ryegrass Seed Fields in Eastern Oregon”. In: *Plant Disease* 100.6 (June 2016), pp. 1110–1117. DOI: 10.1094/PDIS-08-14-0787-RE.

[97] Azza Elemam, Joseph Rahimian, and William Mandell. “Infection with Panresistant Klebsiella pneumoniae: A Report of 2 Cases and a Brief Review of the Literature”. In: *Clinical Infectious Diseases* 49.2 (July 2009), pp. 271–274. DOI: 10.1086/600042.

[98] Ana Paula Mendes Emygdio et al. “One year of temporal characterization of fungal spore concentration in São Paulo metropolitan area, Brazil”. In: *Journal of Aerosol Science* 115 (Jan. 2018), pp. 121–132. DOI: 10.1016/j.jaerosci.2017.07.003.

[99] *Ergot in UK cereal crops and changing legislation | AHDB*. (Accessed 31/07/25). URL: https://ahdb.org.uk/news/ergot-in-uk-cereal-crops-and-changing-legislation.

[100] C. A. Eyre and M. Garbelotto. “Detection, Diversity, and Population Dynamics of Waterborne Phytophthora ramorum Populations”. In: *Phytopathology®* 105.1 (Jan. 2015), pp. 57–68. DOI: 10.1094/PHYTO-07-13-0196-R.

[101] M. L. Fall et al. “Spatiotemporal variation in airborne sporangia of hytophthora infestans: characterization and initiatives towards improving potato late blight risk estimation”. In: *Plant Pathology* 64.1 (2015), pp. 178–190. DOI: 10.1111/ppa.12235.

[102] Mamadou Lamine Fall et al. “Infection Efficiency of Four Phytophthora infestans Clonal Lineages and DNA-Based Quantification of Sporangia”. In: *PLOS ONE* 10.8 (Aug. 2015), e0136312. DOI: 10.1371/journal.pone.0136312.

[103] FAO. *FAOSTAT statistics database: Food balance sheets*. (Accessed 01/08/25). 2020. URL: https://www.fao.org/faostat/en/#home.

[104] E. Feliziani and G. Romanazzi. “Postharvest decay of strawberry fruit: Etiology, epidemiology, and disease management”. In: *Journal of Berry Research* 6.1 (Jan. 2016), pp. 47–63. DOI: 10.3233/JBR-150113.

[105] Gang Feng et al. “Metagenomic analysis of microbial community and function involved in cd-contaminated soil”. In: *BMC Microbiology* 18.1 (Feb. 2018), p. 11. DOI: 10.1186/s12866-018-1152-5.

[106] Robert M. W. Ferguson et al. “Bioaerosol biomonitoring: Sampling optimization for molecular microbial ecology”. In: *Molecular Ecology Resources* 19.3 (2019), pp. 672–690. DOI: 10.1111/1755-0998.13002.

[107] Robert M. W. Ferguson et al. “Size fractionation of bioaerosol emissions from green-waste composting”. In: *Environment International* 147 (Feb. 2021), p. 106327. DOI: 10.1016/j.envint.2020.106327.

[108] Ricardo B. Ferreira et al. “Fungal Pathogens: The Battle for Plant Infection”. In: *Critical Reviews in Plant Sciences* 25.6 (Dec. 2006), pp. 505–524. DOI: 10.1080/07352680601054610.

[109] Sabine Fillinger et al. “Genetic Analysis of Fenhexamid-Resistant Field Isolates of the Phytopathogenic Fungus Botrytis cinerea”. In: *Antimicrobial Agents and Chemotherapy* 52.11 (Nov. 2008), pp. 3933–3940. DOI: 10.1128/AAC.00615-08.

[110] Sabine Fillinger et al. “Functional and Structural Comparison of Pyrrolnitrin- and Iprodione-Induced Modifications in the Class III Histidine-Kinase Bos1 of Botrytis cinerea”. In: *PLOS ONE* 7.8 (Aug. 2012), e42520. DOI: 10.1371/journal.pone.0042520.

[111] Silvia Folloni et al. “Detection of airborne genetically modified maize pollen by real-time PCR”. In: *Molecular Ecology Resources* 12.5 (2012), pp. 810–821. DOI: 10.1111/j.1755-0998.2012.03168.x.

[112] Samuel P. Forry et al. “Variability and bias in microbiome metagenomic sequencing: an interlaboratory study comparing experimental protocols”. In: *Scientific Reports* 14.1 (Apr. 2024), p. 9785. DOI: 10.1038/s41598-024-57981-4.

[113] Sara Franco Ortega et al. “Monitoring and Surveillance of Aerial Mycobiota of Rice Paddy through DNA Metabarcoding and qPCR”. In: *Journal of Fungi* 6.4 (Dec. 2020), p. 372. DOI: 10.3390/jof6040372.

[114] Eric A. Franzosa et al. “Species-level functional profiling of metagenomes and metatranscriptomes”. In: *Nature methods* 15.11 (Nov. 2018), pp. 962–968. DOI: 10.1038/s41592-018-0176-y.

[115] J. Fröhlich-Nowoisky et al. “Diversity and seasonal dynamics of airborne archaea”. In: *Biogeosciences* 11.21 (Nov. 2014), pp. 6067–6079. DOI: 10.5194/bg-11-6067-2014.

[116] Limin Fu et al. “CD-HIT: accelerated for clustering the next-generation sequencing data”. In: *Bioinformatics* 28.23 (Dec. 2012), pp. 3150–3152. DOI: 10.1093/bioinformatics/bts565.

[117] *Fusarium Head Blight of Wheat*. Tech. rep. (Accessed 31/07/25). Crop Protection Network, Mar. 2019.

[118] David M. Gadoury et al. “Initiation, Development, and Survival of Cleistothecia of Podosphaera aphanis and Their Role in the Epidemiology of Strawberry Powdery Mildew”. In: *Phytopathology®* 100.3 (Mar. 2010), pp. 246–251. DOI: 10.1094/PHYTO-100-3-0246.

[119] Georgios Gaitanis et al. “The Malassezia Genus in Skin and Systemic Diseases”. In: *Clinical Microbiology Reviews* 25.1 (Jan. 2012), pp. 106–141. DOI: 10.1128/cmr.00021-11.

[120] Y. Gao et al. “Identification and Characterization of the SnTox6-Snn6 Interaction in the Parastagonospora nodorum–Wheat Pathosystem”. In: *Molecular Plant-Microbe Interactions®* 28.5 (May 2015), pp. 615–625. DOI: 10.1094/MPMI-12-14-0396-R.

[121] Antoine Géry et al. “How to Assess Fungal Contamination of Indoor Air in Dwellings of Patients with Cystic Fibrosis?” In: *Mycopathologia* 190.4 (July 2025), p. 62. DOI: 10.1007/s11046-025-00968-0.

[122] Jay S. Ghurye, Victoria Cepeda-Espinoza, and Mihai Pop. "Metagenomic Assembly: Overview, Challenges and Applications". In: *The Yale Journal of Biology and Medicine* 89.3 (Sept. 2016), pp. 353–362.

[123] Olivia Ginn et al. "Detection and Quantification of Enteric Pathogens in Aerosols Near Open Wastewater Canals in Cities with Poor Sanitation". In: *Environmental Science & Technology* 55.21 (Nov. 2021), pp. 14758–14771. DOI: 10.1021/acs.est.1c05060.

[124] Michael Giolai et al. "Measuring air metagenomic diversity in an agricultural ecosystem". In: *Current Biology* 0.0 (Aug. 2024). DOI: 10.1016/j.cub.2024.07.030.

[125] Dean A. Glawe. "The Powdery Mildews: A Review of the World's Most Familiar (Yet Poorly Known) Plant Pathogens". In: *Annual Review of Phytopathology* 46.Volume 46, 2008 (Sept. 2008), pp. 27–51. DOI: 10.1146/annurev.phyto.46.081407.104740.

[126] Gregory B. Gloor et al. "Microbiome Datasets Are Compositional: And This Is Not Optional". In: *Frontiers in Microbiology* 8 (Nov. 2017). DOI: 10.3389/fmicb.2017.02224.

[127] Cinta Gomez-Silvan et al. "A comparison of methods used to unveil the genetic and metabolic pool in the built environment". In: *Microbiome* 6.1 (Apr. 2018), p. 71. DOI: 10.1186/s40168-018-0453-0.

[128] M. Gonianakis et al. "Airborne Ascomycotina on the island of Crete: Seasonal patternsbased on an 8-year volumetric survey". In: *Aerobiologia* 21.1 (Mar. 2005), pp. 69–74. DOI: 10.1007/s10453-004-5881-6.

[129] Estefanía González-Fernández et al. "Botrytis cinerea Airborne Conidia and Their Germination Ability Assessed by Immunological Methods in a NW Spain Vineyard". In: *Agronomy* 11.7 (July 2021), p. 1441. DOI: 10.3390/agronomy11071441.

[130] Susan Gould et al. "Air and surface sampling for monkeypox virus in a UK hospital: an observational study". In: *The Lancet Microbe* 3.12 (Dec. 2022), e904–e911. DOI: 10.1016/S2666-5247(22)00257-9.

[131] Anja Grabke, Dolores Fernández-Ortuño, and Guido Schnabel. "Fenhexamid Resistance in Botrytis cinerea from Strawberry Fields in the Carolinas Is Associated with Four Target Gene Mutations". In: *Plant Disease* 97.2 (Feb. 2013), pp. 271–276. DOI: 10.1094/PDIS-06-12-0587-RE.

[132] L. L. Granke et al. "Dispersal and Movement Mechanisms of Phytophthora capsici Sporangia". In: *Phytopathology®* 99.11 (Nov. 2009), pp. 1258–1264. DOI: 10.1094/PHYTO-99-11-1258.

[133] Łukasz Grewling et al. "Bioaerosols on the atmospheric super highway: An example of long distance transport of *Alternaria* spores from the Pannonian Plain to Poland". In: *Science of The Total Environment* 819 (May 2022), p. 153148. DOI: 10.1016/j.scitotenv.2022.153148.

[134] Lisa J. Griffiths et al. "Comparison of DNA extraction methods for Aspergillus fumigatus using real-time PCR". In: *Journal of Medical Microbiology* 55.9 (2006), pp. 1187–1191. DOI: 10.1099/jmm.0.46510-0.

[135] Agnieszka Grinn-Gofroń et al. "Airborne fungal spore load and season timing in the Central and Eastern Black Sea region of Turkey explained by climate conditions and land use". In: *Agricultural and Forest Meteorology* 295 (Dec. 2020), p. 108191. DOI: 10.1016/j.agrformet.2020.108191.

[136] T. J. Gulya. "A Seedling Bioassay to Detect the Presence of Plasmopara Halstedii in Soil". In: *Advances in Downy Mildew Research — Volume 2*. Ed. by Peter Spencer-Phillips and Michael Jeger. Dordrecht: Springer Netherlands, 2004, pp. 233–240.

[137] Elena S. Gusareva et al. "Microbial communities in the tropical air ecosystem follow a precise diel cycle". In: *Proceedings of the National Academy of Sciences* 116.46 (Nov. 2019). Publisher: Proceedings of the National Academy of Sciences, pp. 23299–23308. DOI: 10.1073/pnas.1908493116.

[138] Elena S. Gusareva et al. "Taxonomic composition and seasonal dynamics of the air microbiome in West Siberia". In: *Scientific Reports* 10.1 (Dec. 2020), p. 21515. DOI: 10.1038/s41598-020-78604-8.

[139] Ali Hakimzadeh et al. "A pile of pipelines: An overview of the bioinformatics software for metabarcoding data analyses". In: *Molecular Ecology Resources* 24.5 (2024), e13847. DOI: 10.1111/1755-0998.13847.

[140] Avice Hall, Xiaolei Jin, and Jolyon Dodgson. *Control of strawberry powdery mildew under protection | AHDB*. Tech. rep. 29/16. University of Hertfordshire: Horticultural Development Council (HDC), 2016.

[141] Lingxi Han et al. "Deciphering the diversity, composition, function, and network complexity of the soil microbial community after repeated exposure to a fungicide boscalid". In: *Environmental Pollution* 312 (Nov. 2022), p. 120060. DOI: 10.1016/j.envpol.2022.120060.

[142] M. C. Hanson et al. "Climate change impact on fungi in the atmospheric microbiome". In: *Science of The Total Environment* 830 (July 2022), p. 154491. DOI: 10.1016/j.scitotenv.2022.154491.

[143] Emad Hashish et al. "Mycobacterium marinum infection in fish and man: epidemiology, pathophysiology and management; a review". In: *Veterinary Quarterly* 38.1 (Jan. 2018), pp. 35–46. DOI: 10.1080/01652176.2018.1447171.

[144] Thomas H. A. Haverkamp et al. "Detection and characterization of Campylobacter in air samples from poultry houses using shot-gun metagenomics – a pilot study". In: *BMC Microbiology* 24.1 (Dec. 2024), pp. 1–15. DOI: 10.1186/s12866-024-03563-3.

[145] N. D. Havis et al. "Spore dispersal patterns of the ascomycete fungus Ramularia collo-cygni and their influence on disease epidemics". In: *Aerobiologia* 39.2 (June 2023), pp. 213–226. DOI: 10.1007/s10453-023-09787-6.

[146] Neil D. Havis et al. "Ramularia collo-cygni—An Emerging Pathogen of Barley Crops". In: *Phytopathology®* 105.7 (July 2015), pp. 895–904. DOI: 10.1094/PHYTO-11-14-0337-FI.

[147] Peng He et al. "Characteristics of and variation in airborne ARGs among urban hospitals and adjacent urban and suburban communities: A metagenomic approach". In: *Environment International* 139 (June 2020), p. 105625. DOI: 10.1016/j.envint.2020.105625.

[148] Thomas Heaven et al. "A Genomic Resource for the Strawberry Powdery Mildew Pathogen Podosphaera aphanis". In: *Phytopathology®* 113.2 (Feb. 2023), pp. 355–359. DOI: 10.1094/PHYTO-03-22-0091-A.

[149] Darren Heavens et al. "How low can you go? Driving down the DNA input requirements for nanopore sequencing". In: *bioRxiv* (Oct. 2021). DOI: 10.1101/2021.10.15.464554.

[150] William Hemstrom et al. "Next-generation data filtering in the genomics era". In: *Nature Reviews Genetics* 25.11 (Nov. 2024), pp. 750–767. DOI: 10.1038/s41576-024-00738-6.

[151] Alexander Herbig et al. "MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman". In: (Apr. 2016). DOI: https://doi.org/10.1101/050559.

[152] Fernando Hernández Trejo et al. "Airborne ascospores in Mérida (SW Spain) and the effect of rain and other meteorological parameters on their concentration". In: *Aerobiologia* 28.1 (Mar. 2012), pp. 13–26. DOI: 10.1007/s10453-011-9207-1.

[153] P. D. Hildebrand. "Weather Variables in Relation to an Epidemic of Onion Downy Mildew". In: *Phytopathology* 72.2 (1982), p. 219. DOI: 10.1094/Phyto-72-219.

[154] Frieder Hofmann, Mathias Otto, and Werner Wosniok. "Maize pollen deposition in relation to distance from the nearest pollen source under common cultivation - results of 10 years of monitoring (2001 to 2010)". In: *Environmental Sciences Europe* 26.1 (Oct. 2014), p. 24. DOI: 10.1186/s12302-014-0024-3.

[155] Andrew J. Hoisington et al. "Impact of sampler selection on the characterization of the indoor microbiome via high-throughput sequencing". In: *Building and Environment* Complete.80 (2014), pp. 274–282. DOI: 10.1016/j.buildenv.2014.04.021.

[156] M.J.L. de Hoon et al. "Open source clustering software". In: *Bioinformatics* 20.9 (June 2004), pp. 1453–1454. DOI: 10.1093/bioinformatics/bth078.

[157] R. van Hout et al. "The influence of local meteorological conditions on the circadian rhythm of corn (*Zea mays* L.) pollen emission". In: *Agricultural and Forest Meteorology* 148.6 (June 2008), pp. 1078–1092. DOI: 10.1016/j.agrformet.2008.02.009.

[158] Mogens S Hovmøller, Tine Thach, and Annemarie F Justesen. "Global dispersal and diversity of rust fungi in the context of plant health". In: *Current Opinion in Microbiology* 71 (Feb. 2023), p. 102243. DOI: 10.1016/j.mib.2022.102243.

[159] X. R. Hu et al. "Rapid on-site evaluation of the development of resistance to quinone outside inhibitors in Botrytis cinerea". In: *Scientific Reports* 7.1 (Oct. 2017), p. 13861. DOI: 10.1038/s41598-017-13317-z.

[160] Zhichao Hu et al. "Temporal discrepancy of airborne total bacteria and pathogenic bacteria between day and night". In: *Environmental Research* 186 (July 2020), p. 109540. DOI: 10.1016/j.envres.2020.109540.

[161] Li Hua et al. "Pathogenic mechanisms and control strategies of Botrytis cinerea causing post-harvest decay in fruits and vegetables". In: *Food Quality and Safety* 2.3 (Aug. 2018), pp. 111–119. DOI: 10.1093/fqsafe/fyy016.

[162] Mingwei Huang and Christina M. Hull. "Sporulation: how to survive on planet Earth (and beyond)". In: *Current Genetics* 63.5 (Oct. 2017), pp. 831–838. DOI: 10.1007/s00294-017-0694-7.

[163] Shu Huang et al. "Overview of biological ice nucleating particles in the atmosphere". In: *Environment International* 146 (Jan. 2021), p. 106197. DOI: 10.1016/j.envint.2020.106197.

[164] J. Alex Huffman et al. "Real-time sensing of bioaerosols: Review and current perspectives". In: *Aerosol Science and Technology* 54.5 (May 2020), pp. 465–495. DOI: 10.1080/02786826.2019.1664724.

[165] F. M. Humpherson-Jones and Kathleen Phelps. "Climatic factors influencing spore production in Alternaria brassicae and Alternaria brassicicola". In: *Annals of Applied Biology* 114.3 (1989), pp. 449–458. DOI: 10.1111/j.1744-7348.1989.tb03360.x.

[166] Daniel H. Huson et al. "MEGAN analysis of metagenomic data". In: *Genome Research* 17.3 (Mar. 2007), pp. 377–386. DOI: 10.1101/gr.5969107.

[167] R. J. Ingram, H. D. Ludwig, and H. Scherm. "Epidemiology of Exobasidium Leaf and Fruit Spot of Rabbiteye Blueberry: Pathogen Overwintering, Primary Infection, and Disease Progression on Leaves and Fruit". In: *Plant Disease* 103.6 (June 2019), pp. 1293–1301. DOI: 10.1094/PDIS-09-18-1534-RE.

[168] Robert Irving and Erika Wedgwood. *Strawberry crown rot | AHDB*. Tech. rep. 12/07. (Accessed 21/07/25). East Malling Research: Horticultural Development Council (HDC), 2007.

[169] M. Tofazzal Islam et al. "Wheat blast: a new threat to food security". In: *Phytopathology Research* 2.1 (Sept. 2020), p. 28. DOI: 10.1186/s42483-020-00067-6.

[170] S. Iwasaki et al. "Analysis of conidiogenesis and lifelong conidial production from single conidiophores of Podosphaera aphanis on strawberry leaves using digital microscopic and electrostatic techniques". In: *Australasian Plant Pathology* 50.5 (Sept. 2021), pp. 571–587. DOI: 10.1007/s13313-021-00794-0.

[171] Crystal Jaing et al. "Sierra Nevada sweep: metagenomic measurements of bioaerosols vertically distributed across the troposphere". In: *Scientific Reports* 10 (July 2020), p. 12399. DOI: 10.1038/s41598-020-69188-4.

[172] Gwang Il Jang, Chung Yeon Hwang, and Byung Cheol Cho. "Effects of heavy rainfall on the composition of airborne bacterial communities". In: *Frontiers of Environmental Science & Engineering* 12.2 (Nov. 2017), p. 12. DOI: 10.1007/s11783-018-1008-0.

[173] Tünde Jankovics et al. "New Insights into the Life Cycle of the Wheat Powdery Mildew: Direct Observation of Ascosporic Infection in Blumeria graminis f. sp. tritici". In: *Phytopathology®* 105.6 (June 2015), pp. 797–804. DOI: 10.1094/PHYTO-10-14-0268-R.

[174] Nathalie Jarosz et al. "Field measurements of airborne concentration and deposition rate of maize pollen". In: *Agricultural and Forest Meteorology* 119.1 (Oct. 2003), pp. 37–51. DOI: 10.1016/S0168-1923(03)00118-7.

[175] W. R. Jarvis. "The dispersal of spores of *Botrytis cinerea* fr. in a raspberry plantation". In: *Transactions of the British Mycological Society* 45.4 (Dec. 1962), pp. 549–559. DOI: 10.1016/S0007-1536(62)80015-1.

[176] Birgit Jensen et al. "Characterization of microbial communities and fungal metabolites on field grown strawberries from organic and conventional production". In: *International Journal of Food Microbiology* 160.3 (Jan. 2013), pp. 313–322. DOI: 10.1016/j.ijfoodmicro.2012.11.005.

[177] S.-Y. Jeong and T.g. Kim. "Comparison of five membrane filters to collect bioaerosols for airborne microbiome analysis". In: *Journal of Applied Microbiology* 131.2 (2021), pp. 780–790. DOI: 10.1111/jam.14972.

[178] Baofeng Jia et al. "CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database". In: *Nucleic Acids Research* 45.D1 (Jan. 2017), pp. D566–D573. DOI: 10.1093/nar/gkw1004.

[179] Chao Jiang et al. "Decoding personal biotic and abiotic airborne exposome". In: *Nature Protocols* 16.2 (Feb. 2021), pp. 1129–1151. DOI: 10.1038/s41596-020-00451-8.

[180] Wenjun Jiang et al. "Optimized DNA extraction and metagenomic sequencing of airborne microbial communities". In: *Nature Protocols* 10.5 (May 2015), pp. 768–779. DOI: 10.1038/nprot.2015.046.

[181] Xiaoqing Jiang et al. "Global Meta-analysis of Airborne Bacterial Communities and Associations with Anthropogenic Activities". In: *Environmental Science & Technology* 56.14 (July 2022), pp. 9891–9902. DOI: 10.1021/acs.est.1c07923.

[182] Xiaolei Jin et al. "The role of chasmothecia in the initiation of epidemics of powdery mildew (Podospheara aphanis) and the role of silicon in controlling the epidemics on strawberry". In: Oct. 2013.

[183] Mark D. Johnson et al. "Airborne eDNA Reflects Human Activity and Seasonal Changes on a Landscape Scale". In: *Frontiers in Environmental Science* 8 (2021), p. 276. DOI: 10.3389/fenvs.2020.563431.

[184] MD Jones and J.S. Brooks. *Effectiveness of Distance and Border Rows in Preventing Outcrossing in Corn*. Technical Bulletin T-38. (Accessed 17/07/25). Stillwater, OK: Oklahoma Agricultural Experiment Station, 1950, p. 18.

[185] Annemarie Fejer Justesen et al. "Hidden in plain sight: a molecular field survey of three wheat leaf blotch fungal diseases in North-Western Europe shows co-infection is widespread". In: *European Journal of Plant Pathology* 160.4 (Aug. 2021), pp. 949–962. DOI: 10.1007/s10658-021-02298-5.

[186] Lucy K.Mehra et al. "Septoria nodorum blotch of wheat". In: *Septoria nodorum blotch of wheat*. Plant Disease Profiles 19 (Jan. 2019).

[187] Agata Kaczmarek et al. "Influence of soil temperature on Globodera rostochiensis and Globodera pallida". In: *Phytopathologia Mediterranea* 53.3 (Dec. 2014), pp. 396–405. DOI: 10.14601/Phytopathol_Mediterr-13512.

[188] Kraiwuth Kallawicha et al. "Ambient Fungal Spore Concentration in a Subtropical Metropolis: Temporal Distributions and Meteorological Determinants". In: *Aerosol and Air Quality Research* 17.8 (2017), pp. 2051–2063. DOI: 10.4209/aaqr.2016.10.0450.

[189] James Kaminski et al. "High-Specificity Targeted Functional Profiling in Microbial Communities with ShortBRED". In: *PLOS Computational Biology* 11.12 (Dec. 2015). Ed. by William Stafford Noble, e1004557. DOI: 10.1371/journal.pcbi.1004557.

[190] Dongwan D. Kang et al. "MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities". In: *PeerJ* 3 (Aug. 2015), e1165. DOI: 10.7717/peerj.1165.

[191] Petteri Karisto, Frédéric Suffert, and Alexey Mikaberidze. "Measuring Splash Dispersal of a Major Wheat Pathogen in the Field". In: *PhytoFrontiers™* 2.1 (Feb. 2022), pp. 30–40. DOI: 10.1094/PHYTOFR-05-21-0039-R.

[192] Patrick J. Keeling et al. "The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing". In: *PLOS Biology* 12.6 (June 2014), e1001889. DOI: 10.1371/journal.pbio.1001889.

[193] Ryan P. Kelly, Andrew Olaf Shelton, and Ramón Gallego. "Understanding PCR Processes to Draw Meaningful Conclusions from Environmental DNA Studies". In: *Scientific Reports* 9.1 (Aug. 2019), p. 12133. DOI: 10.1038/s41598-019-48546-x.

[194] Napassawan Khammayom et al. "Impact of environmental factors on energy balance of greenhouse for strawberry cultivation". In: *Case Studies in Thermal Engineering* 33 (May 2022), p. 101945. DOI: 10.1016/j.csite.2022.101945.

[195] Chankyung Kim, Monnat Pongpanich, and Thantrira Porntaveetus. "Unraveling metagenomics through long-read sequencing: a comprehensive review". In: *Journal of Translational Medicine* 22.1 (Jan. 2024), p. 111. DOI: 10.1186/s12967-024-04917-1.

[196] Daehwan Kim et al. "Centrifuge: rapid and sensitive classification of metagenomic sequences". In: *Genome Research* 26.12 (Dec. 2016), p. 1721. DOI: 10.1101/gr.210641.116.

[197] Paula King et al. "Longitudinal Metagenomic Analysis of Hospital Air Identifies Clinically Relevant Microbes". In: *PLOS ONE* 11.8 (Aug. 2016), e0160124. DOI: 10.1371/journal.pone.0160124.

[198] Nathan Kleczewski et al. *An Overview of Stagonospora Nodorum Leaf and Glume Blotch*. Tech. rep. (Accessed 31/07/25). United States: Crop Protection Netework, Sept. 2020.

[199] Nathan Kleczewski et al. *An Overview of Stripe Rust of Wheat*. Tech. rep. (Accessed 31/07/25). United States: Crop Protection Netework, Sept. 2020.

[200] Magdalena Klimek-Ochab et al. "Comparative study of fungal cell disruption—scope and limitations of the methods". In: *Folia Microbiologica* 56.5 (Sept. 2011), pp. 469–475. DOI: 10.1007/s12223-011-0069-2.

[201] Leda N Kobziar et al. “Wildland fire smoke alters the composition, diversity, and potential atmospheric function of microbial life in the aerobiome”. In: *ISME Communications* 2.1 (Dec. 2022), p. 8. DOI: 10.1038/s43705-022-00089-5.

[202] Mikhail Kolmogorov et al. “metaFlye: scalable long-read metagenome assembly using repeat graphs”. In: *Nature Methods* 17.11 (Nov. 2020), pp. 1103–1110. DOI: 10.1038/s41592-020-00971-x.

[203] L. F. F. Kox et al. “Diagnostic Values and Utility of Immunological, Morphological, and Molecular Methods for In Planta Detection of Phytophthora ramorum”. In: *Phytopathology®* 97.9 (Sept. 2007), pp. 1119–1129. DOI: 10.1094/PHYTO-97-9-1119.

[204] Matthias Kretschmer et al. “Fungicide-Driven Evolution and Molecular Basis of Multidrug Resistance in Field Populations of the Grey Mould Fungus Botrytis cinerea”. In: *PLOS Pathogens* 5.12 (Dec. 2009), e1000696. DOI: 10.1371/journal.ppat.1000696.

[205] Teruhiko Kuroyanagi et al. “*Botrytis cinerea* identifies host plants via the recognition of antifungal capsidiol to induce expression of a specific detoxification gene”. In: *PNAS Nexus* 1.5 (Nov. 2022). Ed. by Karen E Nelson, pgac274. DOI: 10.1093/pnasnexus/pgac274.

[206] Daniele Lagomarsino Oneto et al. “Timing of fungal spore release dictates survival during atmospheric transport”. In: *Proceedings of the National Academy of Sciences* 117.10 (Mar. 2020), pp. 5134–5143. DOI: 10.1073/pnas.1913752117.

[207] Naama Lang-Yona et al. “Species Richness, rRNA Gene Abundance, and Seasonal Dynamics of Airborne Plant-Pathogenic Oomycetes”. In: *Frontiers in Microbiology* 9 (Nov. 2018). DOI: 10.3389/fmicb.2018.02673.

[208] Ben Langmead et al. “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome”. In: *Genome Biology* 10.3 (Mar. 2009), R25. DOI: 10.1186/gb-2009-10-3-r25.

[209] B. Laurent et al. “The VISA network: a collaborative project between research institutes and vineyard owners to create the first epidemiological monitoring network of downy mildew epidemic based on aerial spore capture”. In: *BIO Web of Conferences* 50 (2022), p. 04007. DOI: 10.1051/bioconf/20225004007.

[210] John A. Lednicky et al. “Collection of SARS-CoV-2 Virus from the Air of a Clinic within a University Student Health Care Center and Analyses of the Viral Genomic Sequence”. In: *Aerosol and Air Quality Research* 20.6 (2020), pp. 1167–1171. DOI: 10.4209/aaqr.2020.05.0202.

[211] A. K. Lees et al. “Real-Time PCR and LAMP Assays for the Detection of Spores of Alternaria solani and Sporangia of Phytophthora infestans to Inform Disease Risk Forecasting”. In: *Plant Disease* 103.12 (Dec. 2019), pp. 3172–3180. DOI: 10.1094/PDIS-04-19-0765-RE.

[212] Richard M. Leggett et al. “NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries”. In: *Bioinformatics* 30.4 (Feb. 2014), pp. 566–568. DOI: 10.1093/bioinformatics/btt702.

[213] Michele Scardine Corrêa de Lemos et al. "Aspergillus in the Indoor Air of Critical Areas of a Tertiary Hospital in Brazil". In: *Journal of Fungi* 10.8 (Aug. 2024), p. 538. DOI: 10.3390/jof10080538.

[214] Pierre Leroux et al. "Exploring Mechanisms of Resistance to Respiratory Inhibitors in Field Strains of *Botrytis cinerea* , the Causal Agent of Gray Mold". In: *Applied and Environmental Microbiology* 76.19 (Oct. 2010), pp. 6615–6630. DOI: 10.1128/AEM.00931-10.

[215] M. H. Y. Leung et al. "Characterization of the public transit air microbiome and resistome reveals geographical specificity". In: *Microbiome* 9.1 (May 2021), p. 112. DOI: 10.1186/s40168-021-01044-7.

[216] Dinghua Li et al. "MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph". In: *Bioinformatics* 31.10 (May 2015), pp. 1674–1676. DOI: 10.1093/bioinformatics/btv033.

[217] Heng Li. "Minimap2: pairwise alignment for nucleotide sequences". In: *Bioinformatics* 34.18 (Sept. 2018), pp. 3094–3100. DOI: 10.1093/bioinformatics/bty191.

[218] Heng Li. *minimap2/FAQ.md at master · lh3/minimap2*. (Accessed 18/09/25). URL: https://github.com/lh3/minimap2/blob/master/FAQ.md.

[219] Heng Li and Richard Durbin. "Fast and accurate long-read alignment with Burrows-Wheeler transform". In: *Bioinformatics (Oxford, England)* 26.5 (Mar. 2010), pp. 589–595. DOI: 10.1093/bioinformatics/btp698.

[220] Kejun Li et al. "Comparison of the biological content of air samples collected at ground level and at higher elevation". In: *Aerobiologia* 26.3 (Sept. 2010), pp. 233–244. DOI: 10.1007/s10453-010-9159-x.

[221] Linyun Li et al. "Municipal Solid Waste Treatment System Increases Ambient Airborne Bacteria and Antibiotic Resistance Genes". In: *Environmental Science & Technology* 54.7 (Apr. 2020), pp. 3900–3908. DOI: 10.1021/acs.est.9b07641.

[222] Joanne E. Littlefair et al. "Air-quality networks collect environmental DNA with the potential to measure biodiversity at continental scales". In: *Current Biology* 33.11 (June 2023), R426–R428. DOI: 10.1016/j.cub.2023.04.036.

[223] *Loose smut in cereals (the importance of clean or fungicide-treated seed) | AHDB*. (Accessed 31/07/25). URL: https://ahdb.org.uk/knowledge-library/loose-smut-in-cereals-the-importance-of-clean-or-fungicide-treated-seed.

[224] *Loose Smut of Wheat*. Reports on Plant Diseases 112. (Accessed 31/07/25). Integrated Pest Management: University of Illinois, May 1990.

[225] Liying Low et al. "Evaluation of full-length nanopore 16S sequencing for detection of pathogens in microbial keratitis". In: *PeerJ* 9 (Feb. 2021), e10778. DOI: 10.7717/peerj.10778.

[226] Irvan Luhung et al. "Experimental parameters defining ultra-low biomass bioaerosol analysis". In: *npj Biofilms and Microbiomes* 7.1 (Dec. 2021), p. 37. DOI: 10.1038/s41522-021-00209-4.

[227] Kevin C. Lutz et al. "A Survey of Statistical Methods for Microbiome Data Analysis". In: *Frontiers in Applied Mathematics and Statistics* 8 (June 2022). DOI: 10.3389/fams.2022.884810.

[228] Christina Lynggaard et al. "Airborne environmental DNA for terrestrial vertebrate community monitoring". In: *Current Biology* (Jan. 2022). DOI: 10.1016/j.cub.2021.12.014.

[229] Guilong Ma et al. "Fungicides alter the distribution and diversity of bacterial and fungal communities in ginseng fields". In: *Bioengineered* 12.1 (Jan. 2021), pp. 8043–8056. DOI: 10.1080/21655979.2021.1982277.

[230] L. V. Madden. "Epidemiology and Control of Leather Rot of Strawberries". In: *Plant Disease* 75.5 (1991), p. 439. DOI: 10.1094/PD-75-0439.

[231] Walter F. Mahaffee et al. "Catching Spores: Linking Epidemiology, Pathogen Biology, and Physics to Ground-Based Airborne Inoculum Monitoring". In: *Plant Disease* 107.1 (Jan. 2023), pp. 13–33. DOI: 10.1094/PDIS-11-21-2570-FE.

[232] Juliana Nicolau Maia et al. "Gray mold in strawberries in the Paraná state of Brazil is caused by Botrytis cinerea and its isolates exhibit multiple-fungicide resistance". In: *Crop Protection* 140 (Feb. 2021), p. 105415. DOI: 10.1016/j.cropro.2020.105415.

[233] Teruya Maki et al. "Vertical distribution of airborne microorganisms over forest environments: A potential source of ice-nucleating bioaerosols". In: *Atmospheric Environment* 302 (June 2023), p. 119726. DOI: 10.1016/j.atmosenv.2023.119726.

[234] Vijini Mallawaarachchi et al. "Solving genomic puzzles: computational methods for metagenomic binning". In: *Briefings in Bioinformatics* 25.5 (Sept. 2024), bbae372. DOI: 10.1093/bib/bbae372.

[235] *Management of powdery mildew in cereals | AHDB*. (Accessed 31/07/25). URL: https://ahdb.org.uk/knowledge-library/management-of-powdery-mildew-in-cereals.

[236] *Management of tan spot disease in wheat, barley and rye | AHDB*. (Accessed 31/07/25). URL: https://ahdb.org.uk/knowledge-library/management-of-tan-spot-disease-in-wheat-barley-and-rye.

[237] Sydonia Manibusan and Gediminas Mainelis. "Passive bioaerosol samplers: A complementary tool for bioaerosol research. A review". In: *Journal of Aerosol Science* 163 (June 2022), p. 105992. DOI: 10.1016/j.jaerosci.2022.105992.

[238] Marcus V. Marin and Natalia A. Peres. "Improving the Toolbox to Manage Phytophthora Diseases of Strawberry: Searching for Chemical Alternatives". In: *Plant Health Progress* 22.3 (Jan. 2021), pp. 294–299. DOI: 10.1094/PHP-02-21-0034-FI.

[239] Marcus V. Marin et al. "Resistance to Mefenoxam of Phytophthora cactorum and Phytophthora nicotianae Causing Crown and Leather Rot in Florida Strawberry". In: *Plant Disease* 105.11 (Nov. 2021), pp. 3490–3495. DOI: 10.1094/PDIS-11-20-2474-RE.

[240] Marcus Vinicius Marin and Natalia A. Peres. "First Report of Sclerotinia sclerotiorum Causing Strawberry Fruit Rot in Florida". In: *Plant Disease* (Aug. 2020). DOI: 10.1094/PDIS-04-20-0879-PDN.

[241] Marcel Martin. "Cutadapt removes adapter sequences from high-throughput sequencing reads". In: *EMBnet.journal* 17.1 (May 2011), pp. 10–12. DOI: 10.14806/ej.17.1.200.

[242] Federico Massi et al. "Fungicide Resistance Evolution and Detection in Plant Pathogens: Plasmopara viticola as a Case Study". In: *Microorganisms* 9.1 (Jan. 2021), p. 119. DOI: 10.3390/microorganisms9010119.

[243] Eva Mayol et al. "Long-range transport of airborne microbes over the global tropical and subtropical ocean". In: *Nature Communications* 8.1 (Aug. 2017), p. 201. DOI: 10.1038/s41467-017-00110-9.

[244] Hamza Mbareche et al. "Recovery of Fungal Cells from Air Samples: a Tale of Loss and Gain". In: *Applied and Environmental Microbiology* 85.9 (Apr. 2019), e02941–18. DOI: 10.1128/AEM.02941-18.

[245] Graham R. D. McGrann et al. "The genome of the emerging barley pathogen Ramularia collo-cygni". In: *BMC Genomics* 17.1 (Aug. 2016), p. 584. DOI: 10.1186/s12864-016-2928-3.

[246] Paul J. McMurdie and Susan Holmes. "phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data". In: *PLOS ONE* 8.4 (Apr. 2013), e61217. DOI: 10.1371/journal.pone.0061217.

[247] N. Mehmood et al. "First Report of Strawberry Leaf Spot Caused by Alternaria alternata in Pakistan". In: *Plant Disease* 102.4 (Apr. 2018), pp. 820–820. DOI: 10.1094/PDIS-09-17-1464-PDN.

[248] Romia Memon, Javed H. Niazi, and Anjum Qureshi. "Biosensors for airborne pathogenic fungal spores detection: a review". In: *Nanoscale* (July 2024). DOI: 10.1039/D4NR01175A.

[249] J. C. Mertely, S. J. MacKenzie, and D. E. Legard. "Timing of Fungicide Applications for Botrytis cinerea Based on Development Stage of Strawberry Flowers and Fruit". In: *Plant Disease* 86.9 (Sept. 2002), pp. 1019–1024. DOI: 10.1094/PDIS.2002.86.9.1019.

[250] Esra Mescioglu et al. "Efficiency of bioaerosol samplers: a comparison study". In: *Aerobiologia* 37.3 (Sept. 2021), pp. 447–459. DOI: 10.1007/s10453-020-09686-0.

[251] Kimberly L. Métris and Jérémy Métris. "Aircraft surveys for air eDNA: probing biodiversity in the sky". In: *PeerJ* 11 (Apr. 2023), e15171. DOI: 10.7717/peerj.15171.

[252] Fernando Meyer et al. "Critical Assessment of Metagenome Interpretation: the second round of challenges". In: *Nature Methods* 19.4 (Apr. 2022), pp. 429–440. DOI: 10.1038/s41592-022-01431-4.

[253] Duccio Migliorini et al. "Temporal patterns of airborne Phytophthora spp. in a woody plant nursery area detected using real-time PCR". In: *Aerobiologia* 35.2 (June 2019), pp. 201–214. DOI: 10.1007/s10453-018-09551-1.

[254] T. C. Miller et al. "Effects of Temperature and Water Vapor Pressure on Conidial Germination and Lesion Expansion of Sphaerotheca macularis f. sp. fragariae". In: *Plant Disease* 87.5 (May 2003), pp. 484–492. DOI: 10.1094/PDIS.2003.87.5.484.

[255] E. Moralejo, J. A. García Muñoz, and E. Descals. "Insights into Phytophthora ramorum sporulation: epidemiological and evolutionary implications". In: *EPPO Bulletin* 36.2 (2006), pp. 383–388. DOI: 10.1111/j.1365-2338.2006.01016.x.

[256] Andreas Mosbach et al. "Anilinopyrimidine Resistance in Botrytis cinerea Is Linked to Mitochondrial Function". In: *Frontiers in Microbiology* 8 (Nov. 2017). DOI: 10.3389/fmicb.2017.02361.

[257] Maurice O. Moss. "Mycotoxin review – 2. Fusarium". In: *Mycologist* 16.4 (Nov. 2002), pp. 158–161. DOI: 10.1017/S0269915X02004135.

[258] Govindan Muthukumar et al. "Early detection and quantification of airborne inocula of *Plasmopara viticola* causing grapevine downy mildew using impaction spore trap". In: *Physiological and Molecular Plant Pathology* 139 (Sept. 2025), p. 102842. DOI: 10.1016/j.pmpp.2025.102842.

[259] Daniel J. Nasko et al. "RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification". In: *Genome Biology* 19.1 (Oct. 2018), p. 165. DOI: 10.1186/s13059-018-1554-6.

[260] Nhu H. Nguyen et al. "FUNGuild: An open annotation tool for parsing fungal community datasets by ecological guild". In: *Fungal Ecology* 20 (Apr. 2016), pp. 241–248. DOI: 10.1016/j.funeco.2015.06.006.

[261] Ying Ni et al. "Benchmarking of Nanopore R10.4 and R9.4.1 flow cells in single-cell whole-genome amplification and whole-genome shotgun sequencing". In: *Computational and Structural Biotechnology Journal* 21 (Jan. 2023), pp. 2352–2364. DOI: 10.1016/j.csbj.2023.03.038.

[262] Mogens Nicolaisen et al. "Fungal Communities Including Plant Pathogens in Near Surface Air Are Similar across Northwestern Europe". In: *Frontiers in Microbiology* 8 (2017). DOI: https://doi.org/10.3389/fmicb.2017.01729.

[263] Rolf Henrik Nilsson et al. "The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications". In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D259–D264. DOI: 10.1093/nar/gky1022.

[264] Mutong Niu et al. "Influence of rainfall on fungal aerobiota in the urban atmosphere over Tianjin, China: A case study". In: *Atmospheric Environment: X* 12 (Dec. 2021), p. 100137. DOI: 10.1016/j.aeaoa.2021.100137.

[265] Nnaemeka Emmanuel Nnadi and Dee A. Carter. "Climate change and the emergence of fungal pathogens". In: *PLOS Pathogens* 17.4 (Apr. 2021), e1009503. DOI: 10.1371/journal.ppat.1009503.

[266] Sergey Nurk et al. "metaSPAdes: a new versatile metagenomic assembler". In: *Genome Research* 27.5 (May 2017), pp. 824–834. DOI: 10.1101/gr.213959.116.

[267] Matthew O'Leary. *What is wheat blast?* (Accessed 31/07/25). Dec. 2019. URL: https://www.cimmyt.org/news/what-is-wheat-blast/.

[268] Bilal Ökmen et al. "The Ustilago hordei–Barley Interaction is a Versatile System for Characterization of Fungal Effectors". In: *Journal of Fungi* 7.2 (Feb. 2021), p. 86. DOI: 10.3390/jof7020086.

[269] Jari Oksanen et al. "vegan: Community Ecology Package". In: (Aug. 2024). DOI: 10.32614/CRAN.package.vegan.

[270] Laura Olivares Boldú. *What is Oxford Nanopore Technology (ONT) sequencing?* (Accessed 17/09/25). URL: https://www.yourgenome.org/theme/what-is-oxford-nanopore-technology-ont-sequencing/.

[271] M. Oliveira et al. "Seasonal and intradiurnal variation of allergenic fungal spores in urban and rural areas of the North of Portugal". In: *Aerobiologia* 25.2 (June 2009), pp. 85–98. DOI: 10.1007/s10453-009-9112-z.

[272] Brian D. Ondov et al. "Mash: fast genome and metagenome distance estimation using MinHash". In: *Genome Biology* 17.1 (June 2016), p. 132. DOI: 10.1186/s13059-016-0997-x.

[273] OpenAI. *ChatGPT*. Large Language Model. URL: https://chat.openai.com/.

[274] Otso Ovaskainen et al. "Global Spore Sampling Project: A global, standardized dataset of airborne fungal DNA". In: *Scientific Data* 11.1 (May 2024), p. 561. DOI: 10.1038/s41597-024-03410-0.

[275] Michael G. Palmer and Gerald J. Holmes. "Fungicide Sensitivity in Strawberry Powdery Mildew Caused by Podosphaera aphanis in California". In: *Plant Disease* 105.9 (Sept. 2021), pp. 2601–2605. DOI: 10.1094/PDIS-12-20-2604-RE.

[276] Amy Y. Pan. "Statistical analysis of microbiome data: The challenge of sparsity". In: *Current Opinion in Endocrine and Metabolic Research* 19 (Aug. 2021), pp. 35–40. DOI: 10.1016/j.coemr.2021.05.005.

[277] Maohua Pan et al. "Collection of Viable Aerosolized Influenza Virus and Other Respiratory Viruses in a Student Health Care Center through Water-Based Condensation Growth". In: *mSphere* 2.5 (Oct. 2017), e00251–17. DOI: 10.1128/mSphere.00251-17.

[278] Franck Panabières et al. "Phytophthora nicotianae diseases worldwide: new knowledge of a long-recognised pathogen". In: *Phytopathologia Mediterranea* 55.1 (2016), pp. 20–40.

[279] Christopher J. Pantazis et al. "Development of an efficient transformation method by Agrobacterium tumefaciens and high throughput spray assay to identify transgenic plants for woodland strawberry (Fragaria vesca) using NPTII selection". In: *Plant Cell Reports* 32.3 (Mar. 2013), pp. 329–337. DOI: 10.1007/s00299-012-1366-1.

[280] Alma E. Parada, David M. Needham, and Jed A. Fuhrman. "Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples". In: *Environmental Microbiology* 18.5 (2016), pp. 1403–1414. DOI: 10.1111/1462-2920.13023.

[281] Ned Peel et al. "MARTi: a real-time analysis and visualisation tool for nanopore metagenomics". In: *bioRxiv* (Feb. 2025). DOI: 10.1101/2025.02.14.638261.

[282] Wentao Peng et al. "Metagenome complexity and template length are the main causes of bias in PCR-based bacteria community analysis". In: *Journal of Basic Microbiology* 58.11 (2018), pp. 987–997. DOI: 10.1002/jobm.201800265.

[283] Danilo Pereira, Bruce A McDonald, and Daniel Croll. "The Genetic Architecture of Emerging Fungicide Resistance in Populations of a Global Wheat Pathogen". In: *Genome Biology and Evolution* 12.12 (Dec. 2020), pp. 2231–2244. DOI: 10.1093/gbe/evaa203.

[284] Simón Pérez Martínez, Rod Snowdon, and Jörn Pons-Kühnemann. "Variability of Cuban and International Populations of Alternaria solani from Different Hosts and Localities: AFLP Genetic Analysis". In: *European Journal of Plant Pathology* 110.4 (Apr. 2004), pp. 399–409. DOI: 10.1023/B:EJPP.0000021071.65146.c0.

[285] O. S. Peries. "Studies on strawberry mildew, caused by Sphaerotheca macularis (Wallr. ex Fries) Jaczewski*". In: *Annals of Applied Biology* 50.2 (1962), pp. 211–224. DOI: 10.1111/j.1744-7348.1962.tb06004.x.

[286] Stefan Petrasch et al. "Grey mould of strawberry, a devastating disease caused by the ubiquitous necrotrophic fungal pathogen Botrytis cinerea". In: *Molecular Plant Pathology* 20.6 (Apr. 2019), pp. 877–892. DOI: 10.1111/mpp.12794.

[287] Paola Pilo et al. "Comparison of microscopic and metagenomic approaches to identify cereal pathogens and track fungal spore release in the field". In: *Frontiers in Plant Science* 13 (Oct. 2022). DOI: 10.3389/fpls.2022.1039090.

[288] Robert Pinard et al. "Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing". In: *BMC Genomics* 7.1 (Aug. 2006), p. 216. DOI: 10.1186/1471-2164-7-216.

[289] Paola Pollegioni et al. "Variability of airborne microbiome at different urban sites across seasons: a case study in Rome". In: *Frontiers in Environmental Science* 11 (Sept. 2023). DOI: 10.3389/fenvs.2023.1213833.

[290] Marcel Polling et al. "Continuous daily sampling of airborne eDNA detects all vertebrate species identified by camera traps". In: *Environmental DNA* 6.4 (2024), e591. DOI: 10.1002/edn3.591.

[291] H. J. Potgieter and M. Alexander. "Susceptibility and Resistance of Several Fungi to Microbial Lysis". In: *Journal of Bacteriology* 91.4 (Apr. 1966), pp. 1526–1532. DOI: 10.1128/jb.91.4.1526-1532.1966.

[292] Alireza Poursafar et al. "Taxonomic study on Alternaria sections Infectoriae and Pseudoalternaria associated with black (sooty) head mold of wheat and barley in Iran". In: *Mycological Progress* 17.3 (Mar. 2018), pp. 343–356. DOI: 10.1007/s11557-017-1358-1.

[293] Powelson. "Initiation of Strawberry fruit rot caused by Botrytis cinerea." In: *Phytopathology* 50.7 (1960).

[294] Aaron J. Prussin, Ellen B. Garcia, and Linsey C. Marr. "Total Virus and Bacteria Concentrations in Indoor and Outdoor Air". In: *Environmental science & technology letters* 2.4 (2015), pp. 84–88. DOI: 10.1021/acs.estlett.5b00050.

[295] Aaron J. Prussin et al. "Seasonal dynamics of DNA and RNA viral bioaerosol communities in a daycare center". In: *Microbiome* 7.1 (Apr. 2019), p. 53. DOI: 10.1186/s40168-019-0672-z.

[296] Christoph Ptatscheck, Birgit Gansfort, and Walter Traunspurger. "The extent of wind-mediated dispersal of small metazoans, focusing nematodes". In: *Scientific Reports* 8.1 (May 2018), p. 6814. DOI: 10.1038/s41598-018-24747-8.

[297] Rohullah Qaderi et al. "Investigating the tolerance of different strawberry cultivars to Botrytis cinerea infection and its relation with fruit quality". In: *Journal of Berry Research* 14.2 (June 2024), pp. 89–103. DOI: 10.3233/JBR-230050.

[298] Nan Qin et al. "Longitudinal survey of microbiome associated with particulate matter in a megacity". In: *Genome Biology* 21.1 (Dec. 2020), p. 55. DOI: 10.1186/s13059-020-01964-x.

[299] Josh Quick. "The 'Three Peaks' faecal DNA extraction method for long-read sequencing". In: (Oct. 2019). DOI: dx.doi.org/10.17504/protocols.io.7rshm6e.

[300] Carolina Diaz Quijano et al. "KP4 to control Ustilago tritici in wheat: Enhanced greenhouse resistance to loose smut and changes in transcript abundance of pathogen related genes in infected KP4 plants". In: *Biotechnology Reports* 11 (Aug. 2016), pp. 90–98. DOI: 10.1016/j.btre.2016.08.002.

[301] Guru V. Radhakrishnan et al. "MARPLE, a point-of-care, strain-level disease diagnostics and surveillance tool for complex fungal pathogens". In: *BMC Biology* 17.1 (Aug. 2019), p. 65. DOI: 10.1186/s12915-019-0684-y.

[302] Bankapura Mariyappa Ravikumara et al. "Claviceps". In: *Compendium of Phytopathogenic Microbes in Agro-Ecology : Vol.1 Fungi*. Ed. by Natarajan Amaresan and Krishna Kumar. Cham: Springer Nature Switzerland, 2025, pp. 97–144.

[303] Gilbert S. Raynor, Eugene C. Ogden, and Janet V. Hayes. "Dispersion and Deposition of Corn Pollen from Experimental Sources". In: *Agronomy Journal* 64.4 (July 1972), pp. 420–427. DOI: 10.2134/agronj1972.00021962006400040004x.

[304] Peter C. Raynor et al. "Comparison of samplers collecting airborne influenza viruses: 1. Primarily impingers and cyclones". In: *PLOS ONE* 16.1 (Jan. 2021), e0244977. DOI: 10.1371/journal.pone.0244977.

[305] Katia Razzini et al. "SARS-CoV-2 RNA detection in the air and on surfaces in the COVID-19 ward of a hospital in Milan, Italy". In: *The Science of the Total Environment* 742 (Nov. 2020), p. 140540. DOI: 10.1016/j.scitotenv.2020.140540.

[306] Miguel A Redondo et al. "Vegetation type determines spore deposition within a forest–agricultural mosaic landscape". In: *FEMS Microbiology Ecology* 96.6 (June 2020), fiaa082. DOI: 10.1093/femsec/fiaa082.

[307] M. Remadi et al. "Effect of Temperature on Aggressivity of Tunisian Fusarium species Causing Potato (Solanum tuberosum L.) Tuber Dry Rot". In: *Journal of Agronomy* 5.2 (2006), pp. 350–355. DOI: 10.3923/ja.2006.350.355.

[308] Tim Reska et al. "Air monitoring by nanopore sequencing". In: *ISME Communications* 4.1 (Jan. 2024), ycae099. DOI: 10.1093/ismeco/ycae099.

[309] K M Reynolds et al. "Splash Dispersal of Phytophthoracactorum from Infected Strawberry Fruit by Simulated Canopy Drip". In: ().

[310] Mina Rho, Haixu Tang, and Yuzhen Ye. "FragGeneScan: predicting genes in short and error-prone reads". In: *Nucleic Acids Research* 38.20 (Nov. 2010), e191. DOI: 10.1093/nar/gkq747.

[311] Miles Richardson et al. "Concurrent measurement of microbiome and allergens in the air of bedrooms of allergy disease patients in the Chicago area". In: *Microbiome* 7.1 (June 2019), p. 82. DOI: 10.1186/s40168-019-0695-5.

[312] Stephen Ries M. *Gray Mold of Strawberry*. Tech. rep. 704. (Accessed 24/01/25). Nov. 1995.

[313] B. Ríos et al. “Diurnal variations of airborne pollen concentration and the effect of ambient temperature in three sites of Mexico City”. In: *International Journal of Biometeorology* 60.5 (May 2016), pp. 771–787. DOI: 10.1007/s00484-015-1061-3.

[314] Sander Y. A. Rodenburg et al. “Functional Analysis of Mating Type Genes and Transcriptome Analysis during Fruiting Body Development of Botrytis cinerea”. In: *mBio* 9.1 (Feb. 2018), e01939–17. DOI: 10.1128/mBio.01939-17.

[315] M. Rodolfi, E. Lorenzi, and A. M. Picco. “Study of the Occurrence of Greenhouse Microfungi in a Botanical Garden”. In: *Journal of Phytopathology* 151.11-12 (2003), pp. 591–599. DOI: 10.1046/j.0931-1785.2003.00771.x.

[316] Krista M. Ruppert, Richard J. Kline, and Md Saydur Rahman. “Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA”. In: *Global Ecology and Conservation* 17 (Jan. 2019), e00547. DOI: 10.1016/j.gecco.2019.e00547.

[317] Andrea Caroline Ruthes and Paul Dahlin. “The Impact of Management Strategies on the Development and Status of Potato Cyst Nematode Populations in Switzerland: An Overview from 1958 to Present”. In: *Plant Disease* 106.4 (Apr. 2022), pp. 1096–1104. DOI: 10.1094/PDIS-04-21-0800-SR.

[318] Jeffrey Sabina and John H. Leamon. “Bias in Whole Genome Amplification: Causes and Considerations”. In: *Whole Genome Amplification: Methods and Protocols*. Ed. by Thomas Kroneis. New York, NY: Springer New York, 2015, pp. 15–41.

[319] Susannah J. Salter et al. “Reagent and laboratory contamination can critically impact sequence-based microbiome analyses”. In: *BMC Biology* 12.1 (Nov. 2014), p. 87. DOI: 10.1186/s12915-014-0087-z.

[320] Jayesh B. Samtani et al. “The Status and Future of the Strawberry Industry in the United States”. In: *HortTechnology* 29.1 (Feb. 2019), pp. 11–24. DOI: 10.21273/HORTTECH04135-18.

[321] F. Sanger, S. Nicklen, and A. R. Coulson. “DNA sequencing with chain-terminating inhibitors”. In: *Proceedings of the National Academy of Sciences of the United States of America* 74.12 (Dec. 1977), pp. 5463–5467. DOI: 10.1073/pnas.74.12.5463.

[322] Daniel J. Sargent et al. “Identification of QTLs for powdery mildew (Podosphaera aphanis; syn. Sphaerotheca macularis f. sp. fragariae) susceptibility in cultivated strawberry (Fragaria ×ananassa)”. In: *PLOS ONE* 14.9 (Sept. 2019), e0222829. DOI: 10.1371/journal.pone.0222829.

[323] Diane G. O. Saunders, Zacharias A. Pretorius, and Mogens S. Hovmøller. “Tackling the re-emergence of wheat stem rust in Western Europe”. In: *Communications Biology* 2.1 (Feb. 2019), pp. 1–3. DOI: 10.1038/s42003-019-0294-9.

[324] Jana Ščevková and Jozef Kováč. “First fungal spore calendar for the atmosphere of Bratislava, Slovakia”. In: *Aerobiologia* 35.2 (June 2019), pp. 343–356. DOI: 10.1007/s10453-019-09564-4.

[325] Patrick D. Schloss et al. “Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system”. In: *PeerJ* 4 (Mar. 2016), e1869. DOI: 10.7717/peerj.1869.

[326] Mikkel Schubert, Stinus Lindgreen, and Ludovic Orlando. "AdapterRemoval v2: rapid adapter trimming, identification, and read merging". In: *BMC Research Notes* 9.1 (Feb. 2016), p. 88. DOI: 10.1186/s13104-016-1900-2.

[327] Silvia Scibetta et al. "Selection and Experimental Evaluation of Universal Primers to Study the Fungal Microbiome of Higher Plants". In: *Phytobiomes Journal* 2.4 (Jan. 2018), pp. 225–236. DOI: 10.1094/PBIOMES-02-18-0009-R.

[328] *Septoria nodorum disease symptoms in cereals | AHDB*. (Accessed 31/07/25). URL: https://ahdb.org.uk/knowledge-library/septoria-nodorum-disease-symptoms-in-cereals.

[329] *Septoria tritici in winter wheat | AHDB*. (Accessed 31/07/25). URL: https://ahdb.org.uk/knowledge-library/septoria-tritici-in-winter-wheat.

[330] Adam J. Sepulveda et al. "The Elephant in the Lab (and Field): Contamination in Aquatic Environmental DNA Studies". In: *Frontiers in Ecology and Evolution* 8 (Dec. 2020). DOI: 10.3389/fevo.2020.609973.

[331] N. Serrano-Silva and M. C. Calderón-Ezquerro. "Metagenomic survey of bacterial diversity in the atmosphere of Mexico City using different sampling methods". In: *Environmental Pollution (Barking, Essex: 1987)* 235 (Apr. 2018), pp. 20–29. DOI: 10.1016/j.envpol.2017.12.035.

[332] Sadikshya Sharma et al. "Genomic approaches for improving resistance to Phytophthora crown rot caused by P. cactorum in strawberry (Fragaria × ananassa)". In: *Frontiers in Agronomy* 4 (2022).

[333] Vineet K. Sharma et al. "Fast and Accurate Taxonomic Assignments of Metagenomic Sequences Using MetaBin". In: *PLOS ONE* 7.4 (Apr. 2012), e34030. DOI: 10.1371/journal.pone.0034030.

[334] Fangxia Shen et al. "Characteristics of biological particulate matters at urban and rural sites in the North China Plain". In: *Environmental Pollution* 253 (Oct. 2019), pp. 569–577. DOI: 10.1016/j.envpol.2019.07.033.

[335] Jian-Cheng Shi et al. "Evaluation of host resistance and susceptibility to *Podosphaera aphanis* NWAU1 infection in 19 strawberry varieties". In: *Scientia Horticulturae* 315 (May 2023), p. 111977. DOI: 10.1016/j.scienta.2023.111977.

[336] Yuting Shi et al. "Fungal Aerosol Diversity Over the Northern South China Sea: The Influence of Land and Ocean". In: *Journal of Geophysical Research: Atmospheres* 127.6 (2022), e2021JD035213. DOI: 10.1029/2021JD035213.

[337] Jaemyung Shin et al. "A deep learning approach for RGB image-based powdery mildew disease detection on strawberry leaves". In: *Computers and Electronics in Agriculture* 183 (Apr. 2021), p. 106042. DOI: 10.1016/j.compag.2021.106042.

[338] Eva Egelyng Sigsgaard et al. "Population-level inferences from environmental DNA—Current status and future perspectives". In: *Evolutionary Applications* 13.2 (Nov. 2019), pp. 245–262. DOI: 10.1111/eva.12882.

[339] Priscilla Gomes da Silva et al. "Evidence of Air and Surface Contamination with SARS-CoV-2 in a Major Hospital in Portugal". In: *International Journal of Environmental Research and Public Health* 19.1 (Jan. 2022), p. 525. DOI: 10.3390/ijerph19010525.

[340] Justin D. Silverman et al. "Measuring and mitigating PCR bias in microbiota datasets". In: *PLOS Computational Biology* 17.7 (July 2021), e1009113. DOI: 10.1371/journal.pcbi.1009113.

[341] Brajesh K. Singh et al. "Climate change impacts on plant pathogens, food security and paths forward". In: *Nature Reviews Microbiology* 21.10 (Oct. 2023), pp. 640–656. DOI: 10.1038/s41579-023-00900-7.

[342] Jagdeep Singh et al. "Important wheat diseases in the US and their management in the 21st century". In: *Frontiers in Plant Science* 13 (Jan. 2023). DOI: 10.3389/fpls.2022.1010191.

[343] Pawan K. Singh et al. "Wheat Blast: A Disease Spreading by Intercontinental Jumps and Its Management Strategies". In: *Frontiers in Plant Science* 12 (July 2021). DOI: 10.3389/fpls.2021.710707.

[344] Ritu Singh et al. "Ascochyta rabiei: A threat to global chickpea production". In: *Molecular Plant Pathology* 23.9 (2022), pp. 1241–1261. DOI: 10.1111/mpp.13235.

[345] C. A. Skjøth et al. "Alternaria spores in the air across Europe: abundance, seasonality and relationships with climate, meteorology and local environment". In: *Aerobiologia* 32.1 (Mar. 2016), pp. 3–22. DOI: 10.1007/s10453-016-9426-6.

[346] Rebecca H. Smith et al. "Investigating the impact of database choice on the accuracy of metagenomic read classification for the rumen microbiome". In: *Animal Microbiome* 4.1 (Nov. 2022), p. 57. DOI: 10.1186/s42523-022-00207-7.

[347] Audrey Sombardier et al. "Sensitivity of Podosphaera aphanis isolates to DMI fungicides: distribution and reduced cross-sensitivity". In: *Pest Management Science* 66.1 (2010), pp. 35–43. DOI: 10.1002/ps.1827.

[348] Hokyung Song et al. "Airborne Bacterial and Eukaryotic Community Structure across the United Kingdom Revealed by High-Throughput Sequencing". In: *Atmosphere* 11.8 (Aug. 2020), p. 802. DOI: 10.3390/atmos11080802.

[349] Lu Song et al. "Comparison of Airborne Antibiotic Resistance Genes in the Chicken Farm during Winter and Summer". In: *Indoor Air* 2024.1 (Jan. 2024), p. 1707863. DOI: 10.1155/2024/1707863.

[350] Robert Starke et al. "Incomplete cell disruption of resistant microbes". In: *Scientific Reports* 9.1 (Apr. 2019), p. 5618. DOI: 10.1038/s41598-019-42188-9.

[351] A. F. Stein et al. "NOAA's HYSPLIT Atmospheric Transport and Dispersion Modeling System". In: *Bulletin of the American Meteorological Society* 96.12 (Dec. 2015), pp. 2059–2077. DOI: 10.1175/BAMS-D-14-00110.1.

[352] Danuta Stępalska and Jerzy Wołek. "Intradiurnal periodicity of fungal spore concentrations (Alternaria, Botrytis, Cladosporium, Didymella, Ganoderma) in Cracow, Poland". In: *Aerobiologia* 25.4 (Dec. 2009), pp. 333–340. DOI: 10.1007/s10453-009-9137-3.

[353] Chris C. Stowers and Erik M. Boczko. "Reliable cell disruption in yeast". In: *Yeast* 24.6 (2007), pp. 533–541. DOI: 10.1002/yea.1491.

[354] "Strawberry plant named 'PRIZE'". PP26193. (Accessed 08/05/25). Dec. 2015.

[355] Alexis R. Sullivan et al. "Airborne eDNA captures three decades of ecosystem biodiversity". In: *bioRxiv* (Dec. 2023). DOI: 10.1101/2023.12.06.569882.

[356] Xiaozhe Sun et al. "Characterization of Alternaria Species Associated with Black Spot of Strawberry in Dandong, China". In: *Agronomy* 13.4 (Apr. 2023), p. 1014. DOI: 10.3390/agronomy13041014.

[357] Fiona A. Symon et al. "A fungal spore calendar for England: Analysis of 13 years of daily concentrations". In: *Allergy* 80.2 (2025), pp. 617–620. DOI: 10.1111/all.16356.

[358] *Tan Spot of Wheat*. Tech. rep. (Accessed 31/07/25). Crop Protection Network, Mar. 2019.

[359] Neslihan Taş et al. "Metagenomic tools in microbial ecology research". In: *Current Opinion in Biotechnology* 67 (Feb. 2021), pp. 184–191. DOI: 10.1016/j.copbio.2021.01.019.

[360] Nick P. Taylor and Nik J. Cunniffe. "Modelling quantitative fungicide resistance and breakdown of resistant cultivars: Designing integrated disease management strategies for Septoria of winter wheat". In: *PLOS Computational Biology* 19.3 (Mar. 2023), e1010969. DOI: 10.1371/journal.pcbi.1010969.

[361] Laura C. Terrón-Camero et al. "Comparison of Metagenomics and Metatranscriptomics Tools: A Guide to Making the Right Choice". In: *Genes* 13.12 (Dec. 2022), p. 2280. DOI: 10.3390/genes13122280.

[362] *The life cycle of ergot and its impact on cereals and grasses | AHDB*. (Accessed 31/07/25). URL: https://ahdb.org.uk/knowledge-library/the-life-cycle-of-ergot-and-its-impact-on-cereals-and-grasses.

[363] Anne-Céline Thuillet et al. "Picturing plant biodiversity from airborne environmental DNA". In: *bioRxiv* (Jan. 2024). DOI: 10.1101/2024.01.11.571706.

[364] J. A. Tomlinson, I. Barker, and N. Boonham. "Faster, Simpler, More-Specific Methods for Improved Molecular Detection of Phytophthora ramorum in the Field". In: *American Society for Microbiology* (June 2007), pp. 4040–4047. DOI: https://doi.org/10.1128/AEM.00161-07.

[365] J.A. Tomlinson, M.J. Dickinson, and N. Boonham. "Detection of Botrytis cinerea by loop-mediated isothermal amplification". In: *Letters in Applied Microbiology* 51.6 (Dec. 2010), pp. 650–657. DOI: 10.1111/j.1472-765X.2010.02949.x.

[366] Denise Pereira Torres et al. "Silencing of RBOHF2 Causes Leaf Age–Dependent Accelerated Senescence, Salicylic Acid Accumulation, and Powdery Mildew Resistance in Barley". In: *Molecular Plant-Microbe Interactions®* 30.11 (Nov. 2017), pp. 906–918. DOI: 10.1094/MPMI-04-17-0088-R.

[367] Georgios Archimidis Tsalidis. "Human Health and Ecosystem Quality Benefits with Life Cycle Assessment Due to Fungicides Elimination in Agriculture". In: *Sustainability* 14.2 (Jan. 2022), p. 846. DOI: 10.3390/su14020846.

[368] Akira Tsuda, Frank S. Henry, and James P. Butler. "Particle Transport and Deposition: Basic Physics of Particle Kinetics". In: *Comprehensive Physiology*. John Wiley & Sons, Ltd, 2013, pp. 1437–1471.

[369] Judith A. Turner et al. “Changes in agronomic practices and incidence and severity of diseases in winter wheat in England and Wales between 1999 and 2019”. In: *Plant Pathology* 70.8 (2021), pp. 1759–1778. DOI: 10.1111/ppa.13433.

[370] Jun Uetake et al. “Seasonal Changes of Airborne Bacterial Communities Over Tokyo and Influence of Local Meteorology”. In: *Frontiers in Microbiology* 10 (2019). DOI: https://doi.org/10.3389/fmicb.2019.01572.

[371] Food and Agriculture Organisation of the United Nations. *Plant production and protection*. (Accessed 15/05/25). 2025. URL: https://www.fao.org/plant-production-protection/about/en.

[372] Martin Urban et al. “PHI-base: the pathogen-host interactions database”. In: *Nucleic Acids Research* 48.D1 (Jan. 2020), pp. D613–D620. DOI: 10.1093/nar/gkz904.

[373] Harika Urel et al. “Nanopore- and AI-empowered metagenomic viability inference”. In: *bioRxiv* (June 2024). DOI: 10.1101/2024.06.10.598221.

[374] Gherman V. Uritskiy, Jocelyne DiRuggiero, and James Taylor. “MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis”. In: *Microbiome* 6.1 (Sept. 2018), p. 158. DOI: 10.1186/s40168-018-0541-1.

[375] Evgeny V. Usachev et al. “Portable automatic bioaerosol sampling system for rapid on-site detection of targeted airborne microorganisms”. In: *Journal of environmental monitoring: JEM* 14.10 (Oct. 2012), pp. 2739–2745. DOI: 10.1039/c2em30317e.

[376] Ioannis Vagelas. “Effective Strategies for Managing Wheat Diseases: Mapping Academic Literature Utilizing VOSviewer and Insights from Our 15 Years of Research”. In: *Agrochemicals* 4.1 (Mar. 2025), p. 4. DOI: 10.3390/agrochemicals4010004.

[377] Pedro M. Valero-Mora. “ggplot2: Elegant Graphics for Data Analysis”. In: *Journal of Statistical Software* 35 (July 2010), pp. 1–3. DOI: 10.18637/jss.v035.b01.

[378] Hervé Van der Heyden et al. “Monitoring airborne inoculum for improved plant disease management. A review”. In: *Agronomy for Sustainable Development* 41.3 (May 2021), p. 40. DOI: 10.1007/s13593-021-00694-z.

[379] T. Veloukas et al. “Fitness and Competitive Ability of Botrytis cinerea Field Isolates with Dual Resistance to SDHI and QoI Fungicides, Associated with Several sdhB and the cytb G143A Mutations”. In: *Phytopathology®* 104.4 (Apr. 2014), pp. 347–356. DOI: 10.1094/PHYTO-07-13-0208-R.

[380] Tomáš Větrovský and Petr Baldrian. “The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses”. In: *PLOS ONE* 8.2 (Feb. 2013), e57923. DOI: 10.1371/journal.pone.0057923.

[381] Brian J. Viner, Mark E. Westgate, and Raymond W. Arritt. “A Model to Predict Diurnal Pollen Shed in Maize”. In: *Crop Science* 50.1 (2010), pp. 235–245. DOI: 10.2135/cropsci2008.11.0670.

[382] Roland W. S. Weber and Matthias Hahn. “Grey mould disease of strawberry in northern Germany: causal agents, fungicide resistance and management strategies”. In: *Applied Microbiology and Biotechnology* 103.4 (Feb. 2019), pp. 1589–1597. DOI: 10.1007/s00253-018-09590-1.

[383] Stephen N. Wegulo et al. "Management of Fusarium head blight of wheat and barley". In: *Crop Protection*. Ecology and management of Fusarium diseases 73 (July 2015), pp. 100–107. DOI: 10.1016/j.cropro.2015.02.025.

[384] Fransizka Wemheuer et al. "Tax4Fun2: prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences". In: *Environmental Microbiome volume* 15.11 (May 2020). DOI: https://doi.org/10.1186/s40793-020-00358-7.

[385] Tao Wen et al. "The best practice for microbiome analysis using R". In: *Protein & Cell* 14.10 (Oct. 2023), pp. 713–725. DOI: 10.1093/procel/pwad024.

[386] J.s. West and R.b.e. Kimber. "Innovations in air sampling to detect plant pathogens". In: *Annals of Applied Biology* 166.1 (2015), pp. 4–17. DOI: 10.1111/aab.12191.

[387] Laura S. Weyrich et al. "Laboratory contamination over time during low-biomass sample analysis". In: *Molecular Ecology Resources* 19.4 (2019), pp. 982–996. DOI: 10.1111/1755-0998.13011.

[388] *Wheat stem rust : USDA ARS*. (Accessed 31/07/25). URL: https://www.ars.usda.gov/midwest-area/stpaul/cereal-disease-lab/docs/cereal-rusts/wheat-stem-rust/.

[389] Liam Whitmore et al. "Inadvertent human genomic bycatch and intentional capture raise beneficial applications and ethical concerns with environmental DNA". In: *Nature Ecology & Evolution* 7.6 (June 2023), pp. 873–888. DOI: 10.1038/s41559-023-02056-2.

[390] Ryan R Wick. *ONT-only accuracy: 5 kHz and Dorado*. (Accessed 17/09/25). Oct. 2023. URL: https://zenodo.org/records/10038673.

[391] Ryan R Wick. *Yet another ONT accuracy test: Dorado v0.5.0*. (Accessed 17/09/25). Dec. 2023. URL: https://zenodo.org/records/10397818.

[392] Thies Marten Wieczorek et al. "Early detection of sugar beet pathogen Ramularia beticola in leaf and air samples using qPCR". In: *European Journal of Plant Pathology* 138.4 (Apr. 2014), pp. 775–785. DOI: 10.1007/s10658-013-0349-6.

[393] Brian Williamson et al. "Effect of humidity on infection of rose petals by dry-inoculated conidia of *Botrytis cinerea*". In: *Mycological Research* 99.11 (Nov. 1995), pp. 1303–1310. DOI: 10.1016/S0953-7562(09)81212-4.

[394] Brian Williamson et al. "Botrytis cinerea: the cause of grey mould disease". In: *Molecular Plant Pathology* 8.5 (2007), pp. 561–580. DOI: 10.1111/j.1364-3703.2007.00417.x.

[395] Derrick E. Wood and Steven L. Salzberg. "Kraken: ultrafast metagenomic sequence classification using exact alignments". In: *Genome Biology* 15.3 (Mar. 2014), R46. DOI: 10.1186/gb-2014-15-3-r46.

[396] Yu-Wei Wu et al. "MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm". In: *Microbiome* 2.1 (Aug. 2014), p. 26. DOI: 10.1186/2049-2618-2-26.

[397] C. L. Xiao et al. "Comparison of Epidemics of Botrytis Fruit Rot and Powdery Mildew of Strawberry in Large Plastic Tunnel and Field Production Systems". In: *Plant Disease* 85.8 (Aug. 2001), pp. 901–909. DOI: 10.1094/PDIS.2001.85.8.901.

[398] Yifei Xu et al. "Detection of Viral Pathogens With Multiplex Nanopore MinION Sequencing: Be Careful With Cross-Talk". In: *Frontiers in Microbiology* 9 (Sept. 2018). DOI: 10.3389/fmicb.2018.02225.

[399] Naomichi Yamamoto et al. "Particle-size distributions and seasonal diversity of allergenic and pathogenic fungi in outdoor air". In: *The ISME Journal* 6.10 (Oct. 2012), pp. 1801–1811. DOI: 10.1038/ismej.2012.30.

[400] Xu Yan et al. "Characteristics of airborne bacterial communities and antibiotic resistance genes under different air quality levels". In: *Environment International* 161 (Mar. 2022), p. 107127. DOI: 10.1016/j.envint.2022.107127.

[401] G Yang et al. "Greenhouses represent an important evolutionary niche for Alternaria alternata". In: *Microbial Ecology* 12.6 (May 2024). DOI: 10.1128/spectrum.00390-24.

[402] Lu Yang and Jun Chen. "Benchmarking differential abundance analysis methods for correlated microbiome sequencing data". In: *Briefings in Bioinformatics* 24.1 (Jan. 2023), bbac607. DOI: 10.1093/bib/bbac607.

[403] *Yellow rust symptoms and management in wheat | AHDB*. (Accessed 31/07/25). URL: https://ahdb.org.uk/yellowrust.

[404] Xiaole Yin et al. "ARGs-OAP v2.0 with an expanded SARG database and Hidden Markov Models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes". In: *Bioinformatics* 34.13 (July 2018), pp. 2263–2270. DOI: 10.1093/bioinformatics/bty053.

[405] Yanni Yin et al. "Fungicide Resistance: Progress in Understanding Mechanism, Monitoring, and Management". In: *Phytopathology®* 113.4 (Apr. 2023), pp. 707–718. DOI: 10.1094/PHYTO-10-22-0370-KD.

[406] Leilei Yu et al. "Postharvest control of *Penicillium expansum* in fruits: A review". In: *Food Bioscience* 36 (Aug. 2020), p. 100633. DOI: 10.1016/j.fbio.2020.100633.

[407] Qiangyi Yu et al. "A cultivated planet in 2010 – Part 2: The global gridded agricultural-production maps". In: *Earth System Science Data* 12.4 (Dec. 2020), pp. 3545–3572. DOI: 10.5194/essd-12-3545-2020.

[408] Abdulmujib G. Yusuf et al. "Optimizing greenhouse microclimate for plant pathology: challenges and cooling solutions for pathogen control in arid regions". In: *Frontiers in Plant Science* 16 (Feb. 2025), p. 1492760. DOI: 10.3389/fpls.2025.1492760.

[409] Rahat Zaheer et al. "Impact of sequencing depth on the characterization of the microbiome and resistome". In: *Scientific Reports* 8.1 (Apr. 2018), p. 5890. DOI: 10.1038/s41598-018-24280-8.

[410] E. a. V. Zauza et al. "Wind dispersal of Puccinia psidii urediniospores and progress of eucalypt rust". In: *Forest Pathology* 45.2 (2015), pp. 102–110. DOI: 10.1111/efp.12133.

[411] Jiang Zhang et al. "PEAR: a fast and accurate Illumina Paired-End reAd mergeR | Bioinformatics | Oxford Academic". In: *Bioinformatics* 30.5 (Mar. 2014), pp. 614–620. DOI: 10.1093/bioinformatics/btt593.

[412] Caiwang Zheng, Amr Abd-Elrahman, and Vance Whitaker. "Remote Sensing and Machine Learning in Crop Phenotyping and Management, with an Emphasis on Applications in Strawberry Farming". In: *Remote Sensing* 13.3 (Jan. 2021), p. 531. DOI: 10.3390/rs13030531.

[413] *ZymoBIOMICS Microbial Community Standard*. (Accessed 25/07/25). URL: https://zymoresearch.eu/products/zymobiomics-microbial-community-standard.

Appendix A

Supplementary Material

A.1 Chapter 3 - Fungicide Application Data

Table A.1: Table of fungicides applied at Wilkin & Sons during sampling period, alongside the target, active chemical, grouping, FRAC code and resistance risk.

Fungicide	**Target**	**Active Chemical**	**Chemical or Biological Group**	**FRAC Code**	**Resistance Risk**
Switch	Botrytis	Cyprodinil	Anilinopyrimidine	9	Medium
Switch	Botrytis	Fludioxonil	Phenylpyrroles	12	Low - Medium
Signum	Botrytis & Podosphaera	Boscalid	Pyridine-carboximide	7	Medium - High
Signum	Botrytis & Podosphaera	Pyraclostrobin	Methoxy-carbamates	11	High
Frupica SC	Botrytis	Mepanipyrim	Aminopyrimidines	9	Medium
Systhane 20 EW	Podosphaera	Myclobutanil	Triazoles	3	Medium
Karma	Podosphaera	Potassium hydrogen carbonate	Organic	N/A	N/A
Sonata	Podosphaera	*Bacillus pumilus* (QST2808)	Bacteria	N/A	N/A
Charm	Podosphaera	Difenoconazole	Triazole	3	Medium
Charm	Podosphaera	Fluxapyroxad	Pyrazole-4- carboxamides	7	Medium - High
Stroby WG	Podosphaera	Kresoxim-methyl	Oximino-acetates	11	High
Teldor	Botrytis	Fenhexamid	Hydroxyanilides	17	Low - Medium
Talius	Podosphaera	Proquinazid	Quinazolinone	13	Medium
Topas	Podosphaera	Penconazole	Triazole	3	Medium
Serenade ASO	Botrytis	*Bacillus subtilis* (QST 713)	Bacteria	BM 02	N/A
Nimrod	Podosphaera	Bupirimate	Hydroxy-(2-amino)-pyrimidines	8	Medium
Takumi SC	Podosphaera	Cyflufenamid	Phenyl-acetamide	U 06	N/A
Luna Sensation	Botrytis & Podosphaera	Fluopyram	Pyridinyl-ethyl-benzamides	7	Medium - High
Luna Sensation	Botrytis & Podosphaera	Trifloxystrobin	Oximino-acetates	11	High
Justice	Podosphaera	Proquinazid	Quinazolinone	13	Medium
Amistar Top	Podosphaera	Azoxystrobin	Methoxy-acrylates	11	High
Amistar Top	Podosphaera	Difenoconazole	Triazoles	3	Medium
Amylo X WG	Botrytis & Podosphaera	*Bacillus amyloliquefaciens* (plantarum D747)	Bacteria	BM 02	N/A
Amistar	Podosphaera	Azoxystrobin	Methoxy-acrylates	11	High
Scala	Botrytis	Pyrimethanil	Anilino-pyrimidines	9	Medium

Table A.2: Table of fungicides applied at Wilkin & Sons, including the location, date, volume and repeat.

Location	Date	Fungicide	Volume	Units	Repeat
Greenhouse 1	04/03/2022	Rigel WP	3	kg	1
Greenhouse 1	19/03/2022	Switch	1	kg	1
Greenhouse 1	23/03/2022	Batavia	1	L	1
Greenhouse 1	05/04/2022	Signum	1.8	kg	1
Greenhouse 1	21/04/2022	Frupica SC	0.9	L	1
Greenhouse 1	27/04/2022	Systhane 20	0.3	L	1
Greenhouse 1	08/05/2022	Karma	2.5	kg	1
Greenhouse 1	13/05/2022	Sonata	10	L	1
Greenhouse 1	20/05/2022	Charm	0.6	L	1
Greenhouse 1	25/05/2022	Karma	2.5	kg	2
Greenhouse 1	03/06/2022	Charm	0.6	L	2
Greenhouse 1	20/06/2022	Stroby WG	0.3	kg	1
Greenhouse 1	20/06/2022	Switch	1	kg	2
Greenhouse 1	26/06/2022	Signum	1.8	kg	2
Greenhouse 1	05/07/2022	Teldor	1	kg	1
Greenhouse 1	05/07/2022	Talius	0.19	L	1
Greenhouse 1	06/07/2022	Topas	0.5	L	1
Greenhouse 1	12/07/2022	Serenade ASO	8	L	1
Greenhouse 1	12/07/2022	Nimrod	1	L	1
Greenhouse 1	21/07/2022	Topas	0.5	L	2
Greenhouse 1	28/07/2022	Luna Sensation	0.8	L	1
Greenhouse 1	05/08/2022	Sonata	10	L	2
Greenhouse 1	13/08/2022	Charm	0.6	L	3
Greenhouse 1	20/08/2022	Sonata	10	L	3
Greenhouse 2	05/04/2022	Signum	1.8	kg	1
Greenhouse 2	19/04/2022	Talius	0.19	L	1
Greenhouse 2	29/04/2022	Systhane 20 EW	0.3	L	1
Greenhouse 2	06/05/2022	Takumi SC	0.15	L	1
Greenhouse 2	13/05/2022	Sonata	10	L	1
Greenhouse 2	20/05/2022	Charm	0.6	L	1
Greenhouse 2	24/05/2022	Karma	2.5	kg	1
Greenhouse 2	02/06/2022	Nimrod	1	L	1
Greenhouse 2	02/06/2022	Serenade ASO	8	L	1
Greenhouse 2	11/06/2022	Justice	0.19	L	1
Greenhouse 2	21/06/2022	Amistar Top	1	L	1
Greenhouse 2	28/06/2022	Karma	3	kg	2
Greenhouse 2	04/07/2022	Charm	0.6	L	2
Greenhouse 2	06/07/2022	Sonata	10	L	2
Greenhouse 2	06/07/2022	Amylo X WG	2.5	kg	1
Greenhouse 2	08/07/2022	Karma	3	kg	3
Greenhouse 2	14/07/2022	Takumi SC	0.15	L	2
Greenhouse 2	21/07/2022	Sonata	10	L	3
Greenhouse 2	27/07/2022	Amylo X WG	2.5	kg	2
Greenhouse 2	03/08/2022	Charm	0.6	L	3
Field	05/05/2022	Switch	1.25	kg	1
Field	12/05/2022	Topas	0.625	L	1
Field	12/05/2022	Teldor	1.875	kg	1
Field	23/05/2022	Switch	1.25	kg	2
Field	23/05/2022	Topas	0.625	L	2
Field	30/05/2022	Frupica SC	1.125	L	1
Field	30/05/2022	Amistar	1.25	L	1
Field	03/06/2022	Signum	2.25	kg	1
Field	03/06/2022	Takumi SC	0.188	L	1
Field	10/06/2022	Scala	2.5	L	1
Field	16/06/2022	Amistar	1.25	L	1
Field	16/06/2022	Frupica SC	1.125	L	2

A.2 Chapter 5 - Wind Speed Data

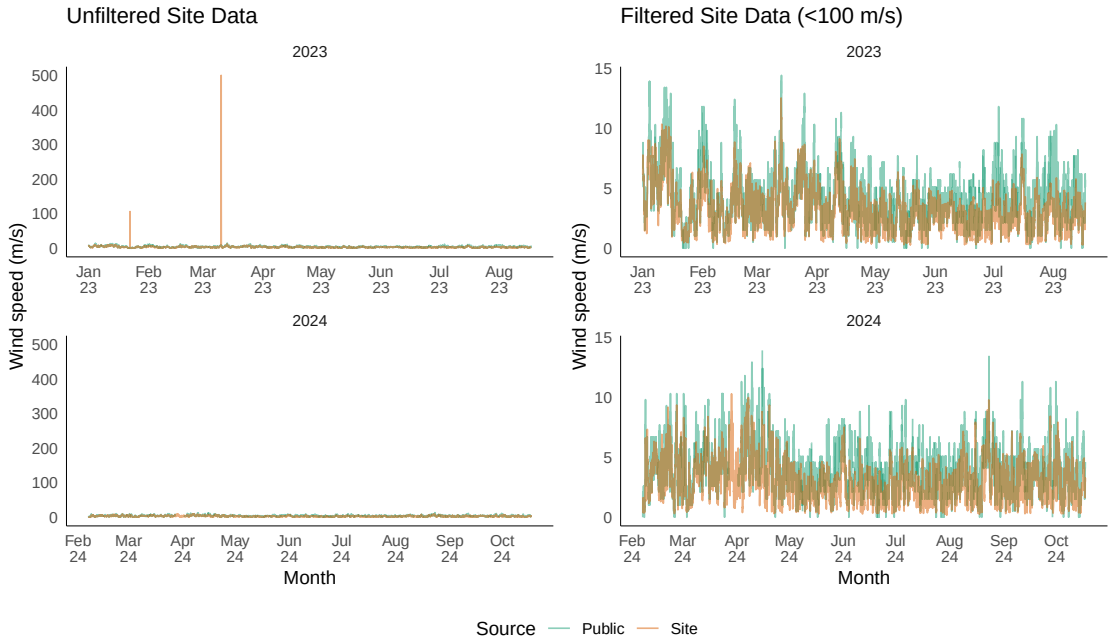


Figure A.1: Comparison of wind speeds recorded at the Church Farm weather station (Site, 10-minute measurements aggregated to hourly means) and the nearby Tibenham Airfield station (Public, reported hourly). The left panels show the raw site data including anomalously high observations (>100 m s^{-1}), while the right panels show the data after these values were removed. Facets indicate the sampling windows used in 2023 and 2024. Lines represent wind speed (m s^{-1}).

A.3 Chapter 7 - Pathogen Reference Database

Table A.3: Species in the reference database

Organism name	**TaxaID**	**Accession**	**Level**	***Organism name***	**TaxaID**	**Accession**	**Level**
Acidovorax citrulli	80869	GCF_022493915.1	Cmpl	*Lonsdalea populi*	1172565	GCF_015999465.1	Cmpl
Acinetobacter baumannii	470	GCF_009035845.1	Cmpl	*Lonsdalea quercina*	71657	GCF_900107885.1	Scf
Acinetobacter nosocomialis	106654	GCF_041021905.1	Cmpl	*Macrosiphum euphorbiae*	13131	GCA_949089665.1	Scf
Actinobacillus pleuropneumoniae	715	GCF_003290385.1	Cmpl	*Magnaporthe oryzae*	242507	GCF_000002495.2	Chr
Aeromonas hydrophila	380703	GCF_000014805.1	Cmpl	*Melampsora medusae*	258770	GCA_002157035.1	Scf
Aeromonas salmonicida	645	GCF_028355655.1	Cmpl	*Meloidogyne javanica*	6303	GCA_036172935.1	Scf
Agrobacterium tumefaciens	358	GCF_013318015.2	Cmpl	*Melon yellowing-associated virus*	255255	GCF_002817615.1	Cmpl
Agrobacterium vitis	373	GCF_001541345.2	Scf	*Metarhizium acridum*	92637	GCF_019434415.1	Scf
Alphacarmovirus calibrachoae	204928	GCF_000908175.1	Cmpl	*Metarhizium anisopliae*	5530	GCA_013305495.1	Ctg
Alstroemeria necrotic streak virus	693450	GCF_013086275.1	Cmpl	*Metarhizium robertsii*	655844	GCF_000187425.2	Scf
Alternaria alternata	5599	GCF_001642055.1	Scf	*Mint vein banding-associated virus*	265877	GCF_002820325.1	Chr

Abbreviations: Cmpl complete genome, Scf scaffold, Ctg contig, Chr chromosome

Organism name	TaxaID	Accession	Level	*Organism name*	TaxaID	Accession	Level
Alternaria brassicicola	29001	GCA_002796735.1	Ctg	*Monilinia fructicola*	38448	GCA_008692225.1	Ctg
Alternaria gaisen	167740	GCA_004156025.2	Ctg	*Monilinia polystroma*	255361	GCA_002909645.1	Scf
Alternaria longipes	160389	GCA_019059555.1	Chr	*Moniliophthora perniciosa*	554373	GCA_000183025.1	Scf
Alternaria panax	48097	GCA_019702505.1	Scf	*Muricauda olearia*	552546	GCF_004150225.1	Scf
Alternaria solani	48100	GCA_002952155.1	Cmpl	*Mycobacterium abscessus*	561007	GCF_000069185.1	Cmpl
American plum line pattern virus	134632	GCF_000850385.1	Cmpl	*Mycobacterium avium*	439334	GCF_022175585.2	Cmpl
Andean potato latent virus	73819	GCF_000906715.1	Cmpl	*Mycobacterium marinum*	1131442	GCF_000723425.2	Cmpl
Andean potato mild mosaic virus	1296569	GCF_000905815.1	Cmpl	*Mycobacterium tuberculosis*	83332	GCF_000195955.2	Cmpl
Andean potato mottle virus	12259	GCF_024750095.1	Cmpl	*Mycoplasma gallisepticum*	2096	GCF_900476085.1	Cmpl
Anisogramma anomala	529478	GCA_038502455.1	Scf	*Myzus persicae*	13164	GCF_001856785.1	Scf
Apple chlorotic leaf spot virus	12175	GCF_000848285.1	Cmpl	*Nectria haematococca*	984957	GCA_010015875.1	Scf
Apple dimple fruit viroid	73494	GCF_000854605.1	Cmpl	*Neisseria gonorrhoeae*	485	GCF_013030075.1	Cmpl
Apple fruit crinkle viroid	190808	GCF_000853885.1	Cmpl	*Neisseria meningitidis*	487	GCF_022869645.1	Cmpl
Apple mosaic virus	12319	GCF_000849545.1	Cmpl	*Neofabraea malicorticis*	108569	GCA_047495995.1	Chr
Apple necrotic mosaic virus	1779339	GCF_004117155.1	Cmpl	*Neofabraea vagabunda*	108571	GCA_045999785.1	Chr
Apple scar skin viroid	190971	GCA_031116735.1	Cmpl	*Neofusicoccum parvum*	310453	GCA_020912385.1	Cmpl
Apple stem grooving virus	33759	GCA_031106045.1	Cmpl	*Neonectria ditissima*	78410	GCA_001305505.1	Scf
Apple stem pitting virus	35350	GCF_000850465.1	Cmpl	*Neonectria neomacrospora*	78403	GCA_917563905.1	Cmpl
Apricot latent virus	75387	GCF_000888875.1	Cmpl	*Onion yellow dwarf virus*	43130	GCF_000862605.1	Cmpl
Arabis mosaic virus	12271	GCF_000855205.1	Cmpl	*Ophiognomonia clavigignenti-juglandacearum*	218668	GCA_003671545.1	Scf
Arion vulgaris	1028688	GCA_020796225.1	Chr	*Ophiostoma novo-ulmi*	170178	GCA_029298975.1	Ctg
Armillaria mellea	47429	GCA_030407055.1	Ctg	*Paenibacillus larvae*	147375	GCF_002951935.1	Cmpl
Ascochyta rabiei	5454	GCF_004011695.2	Cmpl	*Pantoea ananatis*	1095774	GCF_000233595.1	Cmpl
Aspergillus flavus	5059	GCF_009017415.1	Cmpl	*Pantoea stewartii*	66269	GCF_011044475.1	Cmpl
Aspergillus fumigatus	330879	GCF_000002655.1	Chr	*Pantoea stewartii subsp. stewartii*	660596	GCF_002082215.1	Cmpl
Aspergillus nidulans	227321	GCF_000011425.1	Chr	*Paracoccidioides brasiliensis*	502780	GCF_000150735.1	Scf
Avibacterium paragallinarum	728	GCF_011765605.1	Cmpl	*Parastagonospora nodorum*	321614	GCF_000146915.1	Scf
Bacillus anthracis	261594	GCF_000008445.1	Cmpl	*Passalora fulva*	5499	GCF_020509005.1	Cmpl
Bacillus cereus	226900	GCF_046524075.1	Ctg	*Passiflora chlorosis virus*	551003	GCF_002828725.1	Chr
Badnavirus betamaculaflavicannae	419782	GCF_002819365.1	Cmpl	*Pea necrotic yellow dwarf virus*	753670	GCF_000914235.1	Cmpl
Bean golden mosaic virus	10839	GCF_000841845.1	Cmpl	*Peach latent mosaic viroid*	12894	GCF_000850605.1	Cmpl
Beauveria bassiana	655819	GCF_000280675.1	Scf	*Peach mosaic virus*	183585	GCF_000883375.1	Cmpl

Abbreviations: Cmpl complete genome, Scf scaffold, Ctg contig, Chr chromosome

Organism name	TaxaID	Accession	Level	*Organism name*	TaxaID	Accession	Level
Beet curly top virus	268961	GCA_031119855.1	Cmpl	*Peach rosette mosaic virus*	65068	GCF_002029615.1	Chr
Begomovirus solanumvariati	10835	GCF_000837105.1	Cmpl	*Pear blister canker viroid*	12783	GCF_000855365.1	Cmpl
Bipolaris maydis	5016	GCF_028858645.1	Scf	*Pectobacterium aroidearum*	1201031	GCF_015689195.1	Cmpl
Bipolaris oryzae	930090	GCF_000523455.1	Scf	*Pectobacterium atrosepticum*	29471	GCF_000740965.1	Cmpl
Bipolaris sorokiniana	665912	GCF_000338995.1	Scf	*Pectobacterium carotovorum*	554	GCF_013488025.1	Cmpl
Bipolaris victoriae	930091	GCF_000527765.1	Scf	*Pectobacterium wasabiae*	1175631	GCF_001742185.1	Cmpl
Bipolaris zeicola	930089	GCF_000523435.1	Scf	*Penicillium digitatum*	36651	GCF_016767815.1	Cmpl
Black raspberry necrosis virus	367301	GCF_000867325.1	Cmpl	*Penicillium expansum*	27334	GCF_000769745.1	Ctg
Blackcurrant reversion virus	65743	GCF_000849565.1	Cmpl	*Pepino mosaic virus*	112229	GCF_000856965.1	Cmpl
Blastomyces dermatitidis	559297	GCF_000003525.1	Scf	*Pepper chat fruit viroid*	574040	GCF_000881835.1	Cmpl
Blueberry leaf mottle virus	38172	GCF_024750015.1	Cmpl	*Pepper vein yellows virus*	909827	GCF_004787375.1	Cmpl
Blueberry mosaic associated virus	1520332	GCF_000921335.2	Cmpl	*Peronospora belbahrii*	622444	GCA_920618645.1	Ctg
Blueberry red ringspot virus	172220	GCF_000837345.1	Cmpl	*Phomopsis vaccinii*	105482	GCA_039880775.1	Scf
Blueberry scorch virus	31722	GCF_000861365.1	Cmpl	*Photobacterium damselae*	85581	GCF_038086725.1	Cmpl
Blueberry shock virus	747056	GCF_000912635.1	Cmpl	*Photorhabdus luminescens*	171439	GCF_001083805.1	Ctg
Blumeria graminis	1689686	GCA_905067625.1	Chr	*Phyllosticta citricarpa*	55181	GCA_038025095.1	Ctg
Blunervirus solani	2762435	GCF_018595335.1	Cmpl	*Phymatotrichopsis omnivora*	231936	GCA_016880775.1	Scf
Bordetella bronchiseptica	518	GCF_900636925.1	Cmpl	*Phytophthora cactorum*	29920	GCA_016864655.1	Ctg
Bordetella pertussis	520	GCF_004008975.1	Cmpl	*Phytophthora cambivora*	53983	GCA_000443045.1	Scf
Borrelia burgdorferi	445984	GCF_000181575.2	Scf	*Phytophthora capsici*	4784	GCA_030324255.1	Chr
Borrelia hermsii	140	GCF_020422925.2	Cmpl	*Phytophthora cinnamomi*	4785	GCF_018691715.1	Scf
Botryosphaeria kuwatsukai	2021124	GCA_004016305.1	Scf	*Phytophthora citrophthora*	4793	GCA_031305395.1	Scf
Botryosphaeria laricina	121618	GCA_029906385.1	Chr	*Phytophthora cryptogea*	4786	GCA_000468175.2	Scf
Botrytis cinerea	332648	GCF_000143535.2	Cmpl	*Phytophthora foliorum*	415976	GCA_024679135.1	Scf
Botrytis elliptica	278938	GCA_024478385.1	Ctg	*Phytophthora fragariae*	53985	GCA_009729435.1	Ctg
Bremia lactucae	4779	GCF_004359215.1	Chr	*Phytophthora fragariaefolia*	1490495	GCA_030267785.1	Ctg
Brenneria salicis	714314	GCF_012932875.1	Ctg	*Phytophthora infestans*	403677	GCF_000142945.1	Scf
Bretziella fagacearum	1836592	GCA_002018255.1	Scf	*Phytophthora kernoviae*	325452	GCA_008080845.1	Ctg
Brucella abortus	1169205	GCF_000369945.1	Scf	*Phytophthora lateralis*	129355	GCA_000500205.2	Scf
Brucella melitensis	224914	GCF_000007125.1	Cmpl	*Phytophthora nicotianae var. parasitica*	761204	GCF_000247585.1	Scf

Abbreviations: Cmpl complete genome, Scf scaffold, Ctg contig, Chr chromosome

Organism name	TaxaID	Accession	Level	*Organism name*	TaxaID	Accession	Level
Brucella suis	204722	GCF_000007505.1	Cmpl	*Phytophthora parasitica*	761204	GCF_000247585.1	Scf
Burkholderia caryophylli	28094	GCF_034424545.1	Cmpl	*Phytophthora pinifolia*	538568	GCA_000500225.2	Scf
Burkholderia cenocepacia	95486	GCF_001718895.1	Cmpl	*Phytophthora pluvialis*	1330343	GCA_001314425.1	Scf
Burkholderia cepacia	292	GCF_009586235.1	Cmpl	*Phytophthora pseudosyringae*	221518	GCA_019155715.1	Scf
Burkholderia contaminans	488447	GCF_040215475.1	Cmpl	*Phytophthora ramorum*	164328	GCF_020800215.1	Scf
Burkholderia glumae	1176492	GCF_000960995.1	Cmpl	*Phytophthora rubi*	129364	GCA_000687305.2	Scf
Burkholderia mallei	13373	GCF_033956065.1	Cmpl	*Phytophthora sojae*	67593	GCF_000149755.1	Scf
Burkholderia pseudomallei	28450	GCF_030297255.1	Cmpl	*Phytophthora tentaculata*	129362	GCA_033557915.1	Ctg
Burkholderia thailandensis	271848	GCF_000012365.1	Cmpl	*Plantago asiatica mosaic virus*	28354	GCF_000856745.1	Cmpl
Bursaphelenchus xylophilus	6326	GCA_904066235.2	Scf	*Plasmodium berghei*	5823	GCF_900002375.2	Chr
Campylobacter jejuni	192222	GCF_000009085.1	Cmpl	*Plasmodium falciparum*	36329	GCF_000002765.6	Cmpl
Candida albicans	237561	GCF_000182965.3	Chr	*Plasmopara halstedii*	4781	GCF_900000015.1	Scf
Candida dubliniensis	573826	GCF_000026945.1	Cmpl	*Plasmopara obducens*	162140	GCA_003640625.1	Scf
Candida glabrata	5478	GCF_010111755.1	Chr	*Plenodomus tracheiphilus*	1408161	GCA_010093695.1	Scf
Candida parapsilosis	5480	GCF_000182765.1	Ctg	*Plum pox virus*	12211	GCA_002828785.1	Cmpl
Candida tropicalis	294747	GCF_000006335.3	Scf	*Podosphaera aphanis*	79252	GCA_022627015.2	Scf
Candidatus Liberibacter africanus	1277257	GCF_001021085.1	Cmpl	*Porphyromonas gingivalis*	431947	GCF_000010505.1	Cmpl
Candidatus Liberibacter americanus	1261131	GCF_000496595.1	Cmpl	*Potato black ringspot virus*	257464	GCF_000913055.1	Cmpl
Candidatus Liberibacter asiaticus	34021	GCF_000590865.3	Cmpl	*Potato spindle tuber viroid*	12892	GCF_000856265.1	Cmpl
Candidatus Liberibacter solanacearum	658172	GCF_000183665.1	Cmpl	*Potato virus B*	2340870	GCF_003033835.1	Cmpl
Candidatus Phytoplasma australiense	980422	GCF_000397185.1	Cmpl	*Potato virus H*	1046402	GCF_000899815.1	Cmpl
Candidatus Phytoplasma fraxini	35780	GCF_038024965.1	Cmpl	*Potato virus P*	329164	GCF_000871905.1	Cmpl
Candidatus Phytoplasma mali	37692	GCF_000026205.1	Cmpl	*Potato virus T*	36403	GCF_000879695.1	Cmpl
Candidatus Phytoplasma palmae	85624	GCF_046896035.1	Scf	*Potato yellow dwarf virus*	195060	GCF_000895555.1	Cmpl
Candidatus Phytoplasma phoenicium	198422	GCF_001189415.1	Ctg	*Potato yellow mosaic virus*	223311	GCA_031766275.1	Cmpl
Candidatus Phytoplasma pini	267362	GCF_007821455.1	Ctg	*Potato yellow vein virus*	103881	GCF_000852745.1	Cmpl
Candidatus Phytoplasma pruni	479893	GCF_013391955.1	Ctg	*Potato yellowing virus*	936004	GCA_031678425.1	Cmpl
Candidatus Phytoplasma prunorum	47565	GCF_036924415.2	Scf	*Potexvirus ecshostae*	214439	GCF_000880815.1	Cmpl

Abbreviations: Cmpl complete genome, Scf scaffold, Ctg contig, Chr chromosome

Organism name	TaxaID	Accession	Level	*Organism name*	TaxaID	Accession	Level
Candidatus Phytoplasma rubi	399025	GCF_026821955.1	Cmpl	*Proteus mirabilis*	529507	GCF_000069965.1	Cmpl
Candidatus Phytoplasma solani	69896	GCF_040126175.1	Cmpl	*Prune dwarf virus*	33760	GCF_000869765.1	Cmpl
Candidatus Phytoplasma vitis	131152	GCA_023934045.1	Cmpl	*Prunus necrotic ringspot virus*	37733	GCF_000851045.1	Cmpl
Carrot thin leaf virus	114922	GCF_000924455.1	Cmpl	*Pseudocercospora fijiensis*	383855	GCF_000340215.1	Scf
Ceratocystis fimbriata	5158	GCA_032173455.1	Ctg	*Pseudocercospora fuligena*	685502	GCA_014298035.1	Ctg
Ceratocystis harringtonii	312341	GCA_002018265.1	Scf	*Pseudocercospora pini-densiflorae*	1367541	GCA_000504365.2	Scf
Ceratocystis platani	88771	GCA_000978885.1	Ctg	*Pseudomonas aeruginosa*	208964	GCF_000006765.1	Cmpl
Cercospora apii	132184	GCA_022836995.1	Scf	*Pseudomonas avellanae*	46257	GCF_000452845.1	Ctg
Cercospora beticola	122368	GCF_033473495.1	Cmpl	*Pseudomonas cannabina pv. alisalensis*	757414	GCF_016599635.1	Cmpl
Cercospora kikuchii	84275	GCF_019650295.1	Ctg	*Pseudomonas cichorii*	36746	GCF_018343775.1	Cmpl
Cercospora nicotianae	29003	GCA_029490675.1	Ctg	*Pseudomonas fluorescens*	294	GCF_900215245.1	Chr
Cercospora zeae-maydis	135779	GCA_023512815.1	Ctg	*Pseudomonas fuscovaginae*	53407	GCF_900105475.1	Chr
Cherry green ring mottle virus	65467	GCF_000848245.1	Cmpl	*Pseudomonas savastanoi*	29438	GCF_020917325.1	Cmpl
Cherry leaf roll virus	12615	GCF_000893515.1	Cmpl	*Pseudomonas savastanoi pv. savastanoi*	360920	GCF_022026035.1	Cmpl
Cherry mottle leaf virus	131226	GCF_000848465.1	Cmpl	*Pseudomonas syringae*	317	GCF_018394375.1	Cmpl
Cherry necrotic rusty mottle virus	129143	GCF_000849465.1	Cmpl	*Pseudomonas syringae pv. actinidiae*	1108972	GCF_000344475.3	Cmpl
Cherry rasp leaf virus	202566	GCF_000859565.1	Cmpl	*Pseudomonas syringae pv. aesculi*	251722	GCF_029855025.1	Cmpl
Cherry rusty mottle associated virus	1312929	GCF_000907155.1	Cmpl	*Pseudomonas syringae pv. morsprunorum*	129138	GCF_002905685.2	Cmpl
Cherry twisted leaf associated virus	1424279	GCF_000921255.1	Cmpl	*Pseudomonas syringae pv. persicae*	237306	GCF_003700805.1	Scf
Chilli leaf curl virus	341713	GCA_002821925.1	Cmpl	*Pseudomonas syringae pv. syringae*	321	GCF_023277945.1	Cmpl
Chilli veinal mottle virus	52280	GCF_000860025.1	Cmpl	*Pseudomonas syringae pv. ulmi*	251720	GCF_001401165.1	Scf
Chlamydia muridarum	243161	GCF_000006685.1	Cmpl	*Pseudomonas tolaasii*	564423	GCF_002813445.1	Ctg
Chlamydia pneumoniae	182082	GCF_000007205.1	Cmpl	*Pseudomonas viridiflava*	33069	GCF_900184295.1	Cmpl
Chlamydia trachomatis	272561	GCF_000008725.1	Cmpl	*Pseudopyrenochaeta lycopersici*	285811	GCA_003313425.1	Ctg
Chondrostereum purpureum	58369	GCA_004354395.1	Ctg	*Puccinia graminis*	418459	GCF_000149925.1	Scf
Chrysanthemum stem necrosis virus	83871	GCF_001343765.1	Cmpl	*Puccinia horiana*	331382	GCA_001624995.1	Ctg
Chrysanthemum stunt viroid	12897	GCF_000853905.1	Cmpl	*Puccinia striiformis*	168172	GCF_021901695.1	Chr
Ciborinia camelliae	647257	GCA_025890175.1	Scf	*Pyrenophora teres*	97480	GCA_014334815.1	Cmpl

Abbreviations: Cmpl complete genome, Scf scaffold, Ctg contig, Chr chromosome

Organism name	TaxaID	Accession	Level	*Organism name*	TaxaID	Accession	Level
Citrobacter rodentium	1218085	GCF_021278985.1	Cmpl	*Pyrenophora tritici-repentis*	45151	GCF_003171515.1	Chr
Citrus bark cracking viroid	12898	GCF_000852785.1	Cmpl	*Pyricularia grisea*	148305	GCF_004355905.1	Chr
Citrus chlorotic spot dichorhavirus	1980624	GCF_004790215.1	Cmpl	*Pythium aphanidermatum*	1223555	GCA_000387445.2	Scf
Citrus exocortis viroid	457005	GCA_031122165.1	Cmpl	*Quambalaria eucalypti*	363177	GCA_004016185.1	Ctg
Citrus leaf blotch virus	557419	GCA_031123385.1	Cmpl	*Radopholus similis*	46012	GCA_013357305.1	Scf
Citrus mosaic virus	57325	GCA_031522385.1	Cmpl	*Raffaelea lauricola*	483707	GCA_050613675.1	Ctg
Citrus psorosis ophiovirus	73561	GCF_000855005.1	Cmpl	*Raffaelea quercivora*	637633	GCA_002778125.1	Ctg
Citrus tatter leaf virus	33759	GCA_031106045.1	Cmpl	*Ralstonia pickettii*	329	GCF_902374465.1	Scf
Citrus variegation virus	37127	GCF_000870745.1	Cmpl	*Ralstonia pseudosolanacearum*	1310165	GCF_024925465.1	Cmpl
Clavibacter michiganensis	28447	GCF_021216655.1	Cmpl	*Ralstonia solanacearum*	305	GCF_001587155.1	Cmpl
Clavibacter michiganensis subsp. insidiosus	33014	GCF_000958465.1	Cmpl	*Ralstonia syzygii subsp. celebesensis*	1310168	GCF_041734965.1	Cmpl
Clavibacter michiganensis subsp. michiganensis	33013	GCF_011995665.1	Cmpl	*Ralstonia syzygii subsp. indonesiensis*	1310167	GCF_037101365.1	Cmpl
Clavibacter sepedonicus	31964	GCF_000069225.1	Cmpl	*Ralstonia syzygii subsp. syzygii*	1310166	GCA_919592095.1	Ctg
Claviceps purpurea	5111	GCA_029405325.1	Chr	*Raspberry bushy dwarf virus*	12451	GCF_000851365.1	Cmpl
Clavispora lusitaniae	36911	GCF_014636115.1	Ctg	*Raspberry latent virus*	907191	GCF_000889355.1	Cmpl
Clostridioides difficile	1496	GCF_018885085.1	Cmpl	*Raspberry leaf mottle virus*	326941	GCF_000870265.1	Cmpl
Clostridium perfringens	1502	GCF_016027375.1	Cmpl	*Raspberry ringspot virus*	12809	GCF_000854425.1	Cmpl
Clover yellow mosaic virus	12177	GCF_000849905.1	Cmpl	*Raspberry vein chlorosis virus*	758677	GCF_013088605.1	Cmpl
Coccidioides immitis	246410	GCF_000149335.2	Scf	*Rathayibacter toxicus*	145458	GCF_014770185.1	Cmpl
Coccidioides posadasii	443226	GCF_018416015.2	Cmpl	*Rhizoctonia solani*	456999	GCF_016906535.1	Cmpl
Coconut cadang-cadang viroid	36453	GCF_000848145.1	Cmpl	*Rhizopus oryzae*	64495	GCA_048164915.1	Ctg
Colletotrichum acutatum	27357	GCF_030867785.1	Ctg	*Rhynchosporium commune*	2792576	GCA_046529785.1	Cmpl
Colletotrichum chlorophyti	708187	GCA_001937105.1	Scf	*Riemerella anatipestifer*	693978	GCF_000183155.1	Cmpl
Colletotrichum chrysophilum	1836956	GCF_026319265.1	Ctg	*Rose rosette virus*	1980433	GCF_000891875.6	Cmpl
Colletotrichum coccodes	27358	GCA_020466075.1	Ctg	*Rosellinia necatrix*	77044	GCA_021209875.1	Ctg
Colletotrichum fructicola	1213859	GCF_000319635.2	Ctg	*Rubus yellow net virus*	198310	GCF_000928335.1	Cmpl
Colletotrichum gloeosporioides	474922	GCF_011800055.1	Scf	*Saccharomyces cerevisiae*	559292	GCF_000146045.2	Cmpl
Colletotrichum graminicola	645133	GCF_000149035.1	Scf	*Salmonella enterica*	99287	GCF_000006945.2	Cmpl
Colletotrichum higginsianum	759273	GCF_001672515.1	Chr	*Satsuma dwarf virus*	47416	GCF_000860985.1	Cmpl
Colletotrichum karsti	1095194	GCF_011947395.1	Scf	*Sclerotinia sclerotiorum*	665079	GCF_000146945.2	Scf

Abbreviations: Cmpl complete genome, Scf scaffold, Ctg contig, Chr chromosome

Organism name	TaxaID	Accession	Level	*Organism name*	TaxaID	Accession	Level
Colletotrichum lentis	1585795	GCA_003386485.1	Scf	*Serratia marcescens*	615	GCF_030291735.1	Cmpl
Colletotrichum lindemuthianum	290576	GCA_001693025.3	Scf	*Setosphaeria turcica*	671987	GCF_000359705.1	Scf
Colletotrichum linicola	500171	GCA_043790985.1	Cmpl	*Shigella flexneri*	198214	GCF_000006925.2	Cmpl
Colletotrichum orbiculare	1213857	GCA_000350065.2	Scf	*Southern tomato virus*	591166	GCF_000883395.1	Cmpl
Colletotrichum scovillei	1209932	GCF_011075155.1	Ctg	*Sphaerulina musiva*	692275	GCF_000320565.1	Scf
Colletotrichum siamense	690259	GCF_013390195.1	Ctg	*Spiranthes Mosaic Virus 3*	290031	GCF_002829005.1	Chr
Colletotrichum theobromicola	912112	GCA_014705045.1	Scf	*Spiroplasma citri*	2133	GCF_001886855.1	Cmpl
Colletotrichum trifolii	5466	GCA_004367215.1	Scf	*Spongospora subterranea*	166267	GCA_049724395.1	Scf
Colletotrichum truncatum	5467	GCF_014235925.2	Ctg	*Squash leaf curl virus*	10829	GCF_000837705.1	Cmpl
Colombian datura virus	91613	GCF_000903975.1	Cmpl	*Squash vein yellowing virus*	397544	GCF_000879215.1	Cmpl
Columnea latent viroid	12901	GCF_000850525.1	Cmpl	*Stagonosporopsis chrysanthemi*	1200837	GCA_022560025.1	Scf
Coniferiporia sulphurascens	175648	GCA_002794785.1	Ctg	*Staphylococcus aureus*	93061	GCF_000013425.1	Cmpl
Coniferiporia weirii	135589	GCA_024195185.1	Scf	*Staphylococcus epidermidis*	1282	GCF_006094375.1	Cmpl
Corinectria fuckeliana	930093	GCA_019137255.1	Scf	*Staphylococcus lugdunensis*	28035	GCF_001558775.1	Cmpl
Corynebacterium diphtheriae	1717	GCF_001457455.1	Cmpl	*Staphylococcus saprophyticus*	342451	GCF_000010125.1	Cmpl
Cowpea mild mottle virus	67761	GCF_000888775.1	Cmpl	*Stemphylium vesicarium*	119933	GCA_004380135.1	Scf
Coxiella burnetii	227377	GCF_000007765.2	Cmpl	*Stenocarpella maydis*	238245	GCA_002270565.1	Scf
Cronartium comandrae	1301515	GCA_000464975.1	Scf	*Stenotrophomonas maltophilia*	40324	GCF_900186865.1	Cmpl
Cronartium harknessii	3061949	GCA_041381035.1	Scf	*Strawberry chlorotic fleck-associated virus*	399314	GCF_000869405.1	Cmpl
Cronartium quercuum	708437	GCA_015951145.1	Scf	*Strawberry crinkle virus*	135656	GCF_002815575.1	Chr
Cronartium quercuum f. sp. fusiforme	708437	GCA_015951145.1	Scf	*Strawberry latent ringspot virus*	28351	GCF_000857165.1	Cmpl
Cronobacter turicensis	413502	GCF_041222865.1	Cmpl	*Strawberry mild yellow edge virus*	12187	GCF_000855525.1	Cmpl
Cronobacter universalis	1074000	GCF_001277175.1	Cmpl	*Strawberry mottle virus*	167161	GCF_000850445.1	Cmpl
Cryphonectria parasitica	660469	GCF_011745365.1	Scf	*Strawberry necrotic shock virus*	243563	GCF_000869645.1	Cmpl
Cryptococcus gattii	367775	GCF_000185945.1	Chr	*Strawberry vein banding virus*	47903	GCF_000857365.1	Cmpl
Cryptococcus neoformans	235443	GCF_000149245.1	Chr	*Streptococcus agalactiae*	1311	GCF_001552035.1	Cmpl
Cucumber mosaic virus	12305	GCA_003110365.1	Cmpl	*Streptococcus equi*	40041	GCF_015689395.1	Cmpl
Cucumber vein yellowing virus	137475	GCF_000858065.1	Cmpl	*Streptococcus iniae*	1346	GCF_000831485.1	Cmpl
Cucurbit yellow stunting disorder virus	51330	GCF_000851805.1	Cmpl	*Streptococcus parauberis*	873447	GCF_000187935.1	Cmpl

Abbreviations: Cmpl complete genome, Scf scaffold, Ctg contig, Chr chromosome

Organism name	TaxaID	Accession	Level	*Organism name*	TaxaID	Accession	Level
Curtobacterium flaccumfaciens pv. flaccumfaciens	138532	GCF_046996355.2	Cmpl	*Streptococcus pneumoniae*	1313	GCF_001457635.1	Cmpl
Curtobacterium flaccumfaciens pv. poinsettiae	159612	GCF_047226435.1	Cmpl	*Streptococcus pyogenes*	1314	GCF_900475035.1	Cmpl
Curvularia lunata	5503	GCA_005212705.1	Scf	*Streptococcus suis*	568814	GCF_000026745.1	Cmpl
Cystobacter fuscus	1242864	GCF_000335475.2	Ctg	*Streptococcus uberis*	1349	GCF_900475595.1	Cmpl
Davidsoniella virescens	1580837	GCA_001513805.1	Scf	*Streptomyces scabiei*	680198	GCF_000091305.1	Cmpl
Diaporthe amygdali	1214568	GCF_026229845.1	Scf	*Streptomyces turgidiscabies*	85558	GCF_033794965.1	Scf
Diaporthe caulivora	60444	GCA_023703485.1	Ctg	*Stromatinia cepivora*	38492	GCA_014898415.1	Ctg
Dickeya dadantii	204038	GCF_003049785.1	Cmpl	*Sweet potato chlorotic stunt virus*	81931	GCF_000853225.1	Cmpl
Dickeya dianthicola	204039	GCF_003403135.1	Cmpl	*Sweet potato mild mottle virus*	41459	GCF_000860665.1	Cmpl
Dickeya solani	1225786	GCF_001644705.1	Cmpl	*Synchytrium endobioticum*	286115	GCA_006535955.1	Scf
Diplocarpon mali	946123	GCA_024741835.1	Scf	*Temperate fruit decay-associated virus*	1628899	GCF_001190495.1	Cmpl
Diplodia corticola	236234	GCF_001883845.1	Scf	*Tilletia indica*	43049	GCA_001689995.1	Scf
Dothistroma pini	1367539	GCA_002116355.1	Scf	*Tobacco mild green mosaic virus*	12241	GCF_000847545.1	Cmpl
Dothistroma septosporum	64363	GCA_002236755.2	Scf	*Tobacco ringspot virus*	12282	GCF_000856105.1	Cmpl
Edwardsiella ictaluri	67780	GCF_003074995.2	Cmpl	*Tomato apical stunt viroid*	458152	GCA_031122185.1	Cmpl
Edwardsiella tarda	636	GCF_019933175.1	Chr	*Tomato black ring virus*	12275	GCF_000853325.1	Cmpl
Eggplant mottled dwarf nucleorhabdovirus	488317	GCF_000927295.1	Cmpl	*Tomato brown rugose fruit virus*	1761477	GCF_001461485.1	Cmpl
Elsinoe australis	40998	GCA_007556505.1	Scf	*Tomato chlorosis virus*	67754	GCF_000864085.1	Cmpl
Elsinoe fawcettii	40997	GCA_007556565.1	Scf	*Tomato chlorotic dwarf viroid*	100785	GCF_000855405.1	Cmpl
Enterococcus faecalis	1169293	GCF_000393015.1	Scf	*Tomato chocolate virus*	626985	GCA_031522395.1	Cmpl
Enterococcus faecium	1352	GCF_009734005.1	Cmpl	*Tomato infectious chlorosis virus*	52135	GCF_000885955.1	Cmpl
Epichloe festucae	877507	GCA_003814445.1	Cmpl	*Tomato leaf curl New Delhi virus*	223347	GCF_000842925.1	Cmpl
Erwinia amylovora	552	GCF_043228865.1	Cmpl	*Tomato marchitez virus*	470166	GCF_000879575.1	Cmpl
Erwinia pyrifoliae	79967	GCF_002952315.1	Cmpl	*Tomato mild mottle virus*	178599	GCF_002987495.1	Cmpl
Escherichia coli	386585	GCF_000008865.2	Cmpl	*Tomato mosaic Havana virus*	223357	GCF_000837605.1	Cmpl
Euphorbia mosaic virus	429564	GCF_000869345.1	Cmpl	*Tomato mottle mosaic virus*	1391702	GCF_000911995.1	Cmpl
European mountain ash ringspot associated virus	1980426	GCF_000884235.1	Cmpl	*Tomato mottle Taino virus*	223358	GCF_000838745.1	Cmpl
Exophiala dermatitidis	858893	GCF_000230625.1	Scf	*Tomato planta macho viroid*	12888	GCF_000856425.1	Cmpl
Fig mosaic virus	1980427	GCF_001580335.1	Cmpl	*Tomato ringspot virus*	12280	GCF_000860465.1	Cmpl

Abbreviations: Cmpl complete genome, Scf scaffold, Ctg contig, Chr chromosome

Organism name	TaxaID	Accession	Level	*Organism name*	TaxaID	Accession	Level
Flavobacterium psychrophilum	96345	GCF_013343195.2	Cmpl	*Tomato severe rugose virus*	158463	GCF_000874065.1	Cmpl
Francisella tularensis	1450527	GCF_000833355.1	Cmpl	*Tomato spotted wilt virus*	3052585	GCF_000854725.1	Cmpl
Fusarium agapanthi	1803897	GCA_001654545.1	Scf	*Tomato torrado virus*	370833	GCF_000872965.1	Cmpl
Fusarium asiaticum	282267	GCA_025258505.1	Cmpl	*Tomato yellow leaf curl Sardinia virus*	398296	GCA_031121375.1	Cmpl
Fusarium circinatum	48490	GCA_024047395.1	Chr	*Tomato yellow leaf curl virus*	220944	GCA_002987165.1	Cmpl
Fusarium culmorum	5516	GCA_016952355.1	Cmpl	*Tomato yellow ring virus*	304859	GCF_013086795.1	Cmpl
Fusarium foetens	246455	GCA_013623845.1	Scf	*Toxoplasma gondii*	508771	GCF_000006565.2	Chr
Fusarium fujikuroi	1279085	GCF_900079805.1	Chr	*Treponema denticola*	243275	GCF_000008185.1	Cmpl
Fusarium graminearum	229533	GCF_000240135.3	Chr	*Trichoderma virens*	413071	GCF_000170995.1	Scf
Fusarium miscanthi	75915	GCA_014898875.1	Scf	*Trichophyton mentagrophytes*	523103	GCA_047301425.1	Chr
Fusarium oxysporum	660027	GCF_013085055.1	Cmpl	*Trichophyton rubrum*	559305	GCF_000151425.1	Scf
Fusarium oxysporum f. sp. albedinis	72712	GCA_032206225.1	Chr	*Trypanosoma brucei*	185431	GCF_000002445.2	Chr
Fusarium oxysporum f. sp. basilici	121232	GCA_025203125.1	Ctg	*Trypanosoma cruzi*	5693	GCF_000209065.1	Scf
Fusarium oxysporum f. sp. cubense	61366	GCA_027920445.1	Cmpl	*Ustilaginoidea virens*	1159556	GCF_000687475.1	Cmpl
Fusarium oxysporum f. sp. lactucae	299031	GCA_045786885.1	Scf	*Ustilago hordei*	120017	GCF_900519145.1	Ctg
Fusarium pseudograminearum	1028729	GCF_000303195.2	Chr	*Ustilago maydis*	5270	GCF_000328475.2	Chr
Fusarium sambucinum	5128	GCA_050947815.1	Cmpl	*Venturia inaequalis*	5025	GCA_003689225.1	Ctg
Fusarium solani	169388	GCF_023522795.1	Ctg	*Venturia nashicola*	86259	GCA_004522665.1	Ctg
Fusarium sporotrichioides	5514	GCA_019054645.1	Scf	*Verticillium albo-atrum*	27335	GCA_049907295.1	Chr
Fusarium verticillioides	334819	GCF_000149555.1	Chr	*Verticillium dahliae*	498257	GCF_000150675.1	Scf
Fusarium virguliforme	232082	GCA_020883615.1	Ctg	*Verticillium longisporum*	100787	GCA_019188255.1	Scf
Geosmithia morbida	1094350	GCF_012550715.1	Scf	*Verticillium nonalfalfae*	1051616	GCF_003724135.2	Scf
Glaesserella parasuis	738	GCF_017352235.1	Cmpl	*Vibrio aestuarianus*	28171	GCF_028858425.1	Ctg
Globodera pallida	36090	GCA_965641665.1	Scf	*Vibrio anguillarum*	55601	GCF_003390515.1	Chr
Globodera rostochiensis	31243	GCA_018350325.1	Scf	*Vibrio campbellii*	680	GCF_002906475.1	Cmpl
Gnomoniopsis smithogilvyi	1191159	GCA_027946515.1	Scf	*Vibrio cholerae*	666	GCF_008369605.1	Cmpl
Gooseberry vein banding associated virus	157270	GCF_000899795.1	Cmpl	*Vibrio harveyi*	669	GCF_030060435.1	Cmpl
Grapevine fanleaf virus	12274	GCF_000860305.1	Cmpl	*Vibrio parahaemolyticus*	223926	GCF_000196095.1	Cmpl
Grapevine fleck virus	103722	GCF_000859005.1	Cmpl	*Vibrio tasmaniensis*	212663	GCF_024347635.1	Cmpl
Grapevine Pinot gris virus	1051792	GCF_000894735.2	Cmpl	*Vibrio vulnificus*	1219061	GCF_002224265.1	Cmpl

Abbreviations: Cmpl complete genome, Scf scaffold, Ctg contig, Chr chromosome

Organism name	TaxaID	Accession	Level	*Organism name*	TaxaID	Accession	Level
Grapevine red blotch-associated virus	1395525	GCA_031155745.1	Cmpl	*Watermelon silver mottle orthotospovirus*	3052571	GCF_000858945.1	Cmpl
Grapevine roditis leaf discoloration-associated virus	1471299	GCF_001019755.1	Cmpl	*Wheat dwarf virus*	452663	GCA_031122015.1	Cmpl
Grapevine Syrah virus 1	630199	GCF_000882775.1	Cmpl	*Xanthomonas albilineans*	29447	GCF_009931595.1	Cmpl
Grapevine vein clearing virus	1050407	GCF_000891255.1	Cmpl	*Xanthomonas arboricola pv. corylina*	487821	GCA_030284545.1	Cmpl
Grosmannia clavigera	655863	GCF_000143105.1	Scf	*Xanthomonas arboricola pv. juglandis*	195709	GCF_905367715.1	Cmpl
Groundnut bud necrosis virus	198612	GCA_031551155.1	Cmpl	*Xanthomonas arboricola pv. pruni*	69929	GCF_041464225.1	Cmpl
Groundnut ringspot virus	12675	GCF_003972785.1	Cmpl	*Xanthomonas axonopodis*	487832	GCF_041519315.1	Cmpl
Haemophilus ducreyi	730	GCF_001647695.1	Cmpl	*Xanthomonas axonopodis pv. allii*	1437449	GCF_000730305.1	Chr
Haemophilus influenzae	727	GCF_020736045.1	Cmpl	*Xanthomonas axonopodis pv. citri*	611301	GCF_000961215.1	Cmpl
Helicobacter pylori	210	GCF_025998455.1	Cmpl	*Xanthomonas axonopodis pv. dieffenbachiae*	92828	GCF_008639345.1	Cmpl
Heterobasidion annosum	13563	GCA_001457955.1	Scf	*Xanthomonas axonopodis pv. phaseoli*	317013	GCF_002759115.2	Cmpl
Heterobasidion irregulare	747525	GCF_000320585.1	Scf	*Xanthomonas axonopodis pv. poinsettiicola*	375353	GCF_041484315.1	Cmpl
Heterobasidion occidentale	942053	GCA_039111635.1	Scf	*Xanthomonas campestris*	359385	GCF_013388375.1	Cmpl
Heterobasidion parviporum	207832	GCA_002994785.1	Scf	*Xanthomonas campestris pv. fici*	487866	GCF_041463755.1	Cmpl
Heterodera glycines	51029	GCA_004148225.2	Chr	*Xanthomonas citri*	611301	GCF_000961215.1	Cmpl
Histoplasma capsulatum	447093	GCF_000150115.1	Scf	*Xanthomonas dyei*	743699	GCF_032698475.1	Cmpl
Hop stunt viroid	12893	GCF_000847785.1	Cmpl	*Xanthomonas euvesicatoria*	359387	GCF_017724035.1	Cmpl
Hyaloperonospora arabidopsidis	272952	GCA_001414525.2	Ctg	*Xanthomonas fragariae*	48664	GCF_900183975.1	Cmpl
Hymenoscyphus fraxineus	746836	GCA_900184765.1	Scf	*Xanthomonas fuscans subsp. fuscans*	366649	GCF_004000475.1	Cmpl
Impatiens necrotic spot virus	11612	GCF_000852025.1	Cmpl	*Xanthomonas gardneri*	2754056	GCF_001908775.1	Cmpl
Iresine viroid 1	53193	GCF_000852825.1	Cmpl	*Xanthomonas hortorum*	56454	GCF_002285515.1	Cmpl
Iris yellow spot virus	60456	GCF_001611645.5	Cmpl	*Xanthomonas hydrangeae*	2775159	GCF_032697525.1	Cmpl
Kabatiella zeae	1592451	GCA_048593495.1	Ctg	*Xanthomonas oryzae*	129394	GCF_008370835.2	Cmpl
Kingella kingae	504	GCF_900475905.1	Cmpl	*Xanthomonas oryzae pv. oryzae*	360094	GCF_000019585.2	Cmpl
Klebsiella pneumoniae	1125630	GCF_000240185.1	Cmpl	*Xanthomonas oryzae pv. oryzicola*	129394	GCF_008370835.2	Cmpl

Abbreviations: Cmpl complete genome, Scf scaffold, Ctg contig, Chr chromosome

Organism name	TaxaID	Accession	Level	*Organism name*	TaxaID	Accession	Level
Lactococcus lactis	1360	GCF_003176835.1	Cmpl	*Xanthomonas perforans*	442694	GCF_013112235.1	Ctg
Lecanicillium fungicola	93591	GCA_051527305.1	Scf	*Xanthomonas translucens*	487909	GCF_017301775.1	Cmpl
Lecanosticta acicola	111012	GCA_963674455.1	Ctg	*Xanthomonas translucens pv. translucens*	134875	GCF_017301695.1	Cmpl
Leek yellow stripe virus	42004	GCF_000858985.1	Cmpl	*Xanthomonas vesicatoria*	925775	GCF_001908725.1	Cmpl
Legionella pneumophila	446	GCF_001941585.1	Cmpl	*Xenorhabdus nematophila*	406817	GCF_000252955.1	Chr
Leishmania infantum	435258	GCF_000002875.2	Chr	*Xylella fastidiosa*	2371	GCF_028891345.1	Cmpl
Leishmania major	347515	GCF_000002725.2	Cmpl	*Xylella taiwanensis*	1444770	GCF_013177435.1	Cmpl
Leishmania mexicana	929439	GCF_000234665.1	Chr	*Xylophilus ampelinus*	54067	GCF_003217575.1	Scf
Leptosphaeria maculans	5022	GCF_022343315.1	Scf	*Yersinia enterocolitica*	930944	GCF_000253175.1	Cmpl
Leptospira interrogans	44275	GCF_002073495.2	Cmpl	*Yersinia pestis*	1035377	GCF_000222975.1	Cmpl
Lettuce infectious yellows virus	31713	GCF_000850585.1	Cmpl	*Yersinia pseudotuberculosis*	633	GCF_900637475.1	Cmpl
Listeria monocytogenes	169963	GCF_000196035.1	Cmpl	*Yersinia ruckeri*	29486	GCF_017498685.1	Cmpl
Little cherry virus 1	217686	GCF_000862325.1	Cmpl	*Zymoseptoria tritici*	336722	GCF_000219625.1	Chr
Little cherry virus 2	154339	GCF_000855865.1	Cmpl				

Abbreviations: Cmpl complete genome, Scf scaffold, Ctg contig, Chr chromosome