STUDENT PROJECT What lives within and on a plant: our understanding from genome NGS data

Yalan Li[1], Kanae Nishii[2], Nathan Kelso[3], Sadie Barber[4], Louise Galloway[5], Michael Möller[6] & Joanne Taylor[7]

Abstract

Next-generation sequencing (NGS) can generate gigabytes of genome data. Unlike Sanger sequencing, NGS generates a 'read' from a single DNA molecule, reflecting directly the starting DNA, including non-target organisms such as symbionts and pathogens. Non-target organism sequences are usually discarded during genome assembly as contaminants; these are potentially a great source of information for understanding the microbiome surrounding the plant. The present study explores bioinformatically the identification of the non-target organisms from genome NGS datasets of two cultivated Gesneriaceae species. The datasets were generated using different NGS technologies: one is from *Streptocarpus rexii* (Bowie ex Hook.) Lindl., sequenced using Oxford Nanopore Technologies long-read sequencing, and the second from *Aeschynanthus angustifolius* (Blume) Steud., sequenced using Illumina short-read sequencing. The reads were first assembled and then analysed using BlobTools to identify the contaminants. For *S. rexii*, Actinomycetota and Basidiomycota occupied the highest ratio among genome contaminants, followed by Arthropoda, Ascomycota and Acidobacteriota. In *A. angustifolius*, the highest contaminant class was Pseudomonadota and the second Actinomycetota, followed by Basidiomycota and Chordata. Arthropoda included mealybugs which were also observed in the glasshouse. The differences in contaminant composition between *S. rexii* and *A. angustifolius* may be linked to the relatively short-lived leaves of the former and the long-lived ones of the latter. This pilot study demonstrates that, in principle, this method is suitable to detect and identify associated organisms, and the pipelines designed here greatly facilitated this process. This approach might be useful in a horticultural setting for the assessment of plant material in quarantine or biosecure conditions and may be able to detect pathogens prior to plants showing symptoms. It also has potentially more widespread applications for studying plant–microbiome interactions.

[1] Yalan Li graduated with an MSc in Biodiversity and Taxonomy of Plants from the Royal Botanic Garden Edinburgh and the University of Edinburgh in 2024.

Address: 20A Inverleith Row, Edinburgh EH3 5LR, UK.

[2] Kanae Nishii is Senior Technician in the Science Department at the Royal Botanic Garden Edinburgh and Research Associate at Kanagawa University, Japan.

Address: as above.

[3] Nathan Kelso is Glasshouse Horticulturist in the Glasshouse Department at the Royal Botanic Garden Edinburgh.
Address: as above.

[4] Sadie Barber is Research Collections Manager in the Glasshouse Department at the Royal Botanic Garden Edinburgh.
Address: as above.

[5] Louise Galloway is a Supervisor in the Glasshouse Department at the Royal Botanic Garden Edinburgh.
Address: as above.

[6] Michael Möller is Molecular Systematist and Cytologist in the Science Department at the Royal Botanic Garden Edinburgh.

Address: as above.

[7] Joanne Taylor is Research Associate: Mycologist in the Science Department at the Royal Botanic Garden Edinburgh.
Address: as above.

DOI 10.24823/Sibbaldia.2026.2130

Introduction

Plants are associated with a microbiome including both endosymbionts within the tissues, such as fungi, bacteria and viruses, and epiphytic organisms on the plant tissue surfaces, which might include fungi and bacteria as well as other organisms such as invertebrates and algae (Kennedy & Southwood, 1984; Liu *et al.*, 2023; Sohrabi *et al.*, 2023). The association that living plant tissues have with these organisms ranges from pathogenic (including parasitic but not necessarily pathogenic) to beneficial, or even neutral (Saikkonen *et al.*, 1998; Sohrabi *et al.*, 2023; Thomas *et al.*, 2024). Pathogenic organisms can often exist latently within or on plant tissues without causing symptoms of disease (Carroll, 1988). These are of particular biosecurity concern as they could be transported with what is believed to be healthy plant tissue (Marsberg *et al.*, 2017) and contribute to the spread of pests and pathogens globally through the movement of plant material (Mehl *et al.*, 2017; Franić *et al.*, 2023).

In a horticultural setting, such as botanic gardens holding globally important plant collections, the detection of plant-associated organisms is vital to prevent the introduction and spread of harmful organisms. In previous decades this was mainly done by inspection and isolation directly from plant tissue in dedicated quarantine units (Waller *et al.*, 2001; BGCI, 2022), but in recent years molecular techniques have been increasingly employed. These might include various methods involving polymerase chain reaction (PCR), including quantitative PCR (qPCR) and reverse transcription PCR (e.g. for RNA viruses), the use of probes (*in situ* hybridisation), immunohistochemistry and next-generation sequencing (NGS)-based techniques (Piombo *et al.*, 2021; Venbrux *et al.*, 2023). NGS, also known as 'high-throughput sequencing' (Saini *et al.*, 2023), can be used for genomic studies to sequence a genome of a single organism, but also for metabarcoding studies where DNA in environmental samples is sequenced (usually providing amplicons of the same gene loci for all the organisms present) to determine the identity of the organisms comprising the community (Pérez-Cobas *et al.*, 2020; Piombo *et al.*, 2021).

NGS includes a variety of contemporary sequencing technologies with four main steps: DNA isolation, target sequence enrichment, sequencing on NGS platforms and bioinformatic analysis (Valencia *et al.*, 2013). All NGS technologies can carry out parallel sequencing of millions of fragments of DNA (Behjati & Tarpey, 2013). 'Reads' are the nucleotide sequences resulting from sequencing DNA fragments. These are typically 25–150 base pairs (bp) in length in one direction for Illumina short-read sequencing (SRS) methods (Lin *et al.*, 2012), or up to >10 kbp for long-read sequencing (LRS) methods (Hu *et al.*, 2021; Pucker *et al.*, 2022). At present, SRS is dominated by Illumina and LRS by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) platforms. Overall, LRS is preferred since it generates longer contiguous assemblies, whereas Illumina SRS results in many shorter scaffolds (i.e. assemblies of reads) but is more cost-effective per bp.

A greatly overlooked source of metadata for analysis of the plant microbiome is the 'contamination' part of plant genome DNA raw data with reads from non-target organisms. Genome contamination refers to the unintentional inclusion of sequences from other organisms or the misclassification of sequences in public databases (Lupo *et al.*, 2021). Contamination may happen at various

DOI 10.24823/Sibbaldia.2026.2130

stages of a genome assembly project from other organisms existing in the surrounding environment or from endosymbionts (Lu & Salzberg, 2018). Researchers routinely observe 'contamination' of DNA samples with the genomes of other species including symbionts, pathogens and parasites (Laetsch & Blaxter, 2017). Some researchers have purposely studied these contaminants suggesting that they can provide incidental data that could be followed up with further investigation (Galanti *et al.*, 2024) or constitute complete studies where numerous genome datasets are available (Sangiovanni *et al.*, 2019; Roman-Reyna *et al.*, 2020; Gathercole *et al.*, 2021).

Laetsch & Blaxter (2017) provided an in-depth and detailed processing of NGS reads or scaffolds in BlobTools, indicating that it can assist in elementary partitioning of data, visualisation of genome assemblies and screening of ultimate assemblies for potential contaminants.

The present study is an exploratory investigation into the quality and quantity of contaminants in genomes. Gesneriaceae genomes were used as test cases. They were generated using a combination of NGS techniques to explore and establish workflows for the isolation and identification of contaminants and visualised using BlobTools. Within the Gesneriaceae, several genomic datasets are available, such as for *Streptocarpus rexii* (Bowie ex Hook.) Lindl., a short-lived herbaceous plant, published by Nishii *et al.* (2022), and for *Aeschynanthus angustifolius* (Blume) Steud., a perennial long-living epiphyte (Nishii *et al.*, unpublished). Analysing the contaminants in these genomes can help determine the diversity of plant pests and pathogens in the Royal Botanic Garden Edinburgh (RBGE) horticultural glasshouses, which can contribute to targeted pest control and plant protection efforts. It can also help researchers understand the presence and types of endogenous symbionts in Gesneriaceae.

Materials and methods

In this study, a separate pipeline was designed for each of the two datasets tailored to the sequencing technology: LRS (ONT) and SRS (Illumina) with paired-end reads (Figs 1 & 2).

The analyses were performed on the CropDiversity high-performance cluster (HPC) at the James Hutton Institute (Dundee, UK), where Slurm is employed as the job scheduler.[8] All detailed commands are provided in Li (2024).

LRS pipeline for Streptocarpus rexii *ONT data*

Data source: The contaminant raw data were part of an ONT genome sequencing long-read dataset from a previous study reported by Nishii *et al.* (2022). In brief, sequencing and nucleotide basecalling were carried out at Edinburgh Genomics, where two sequencing libraries were generated using SQK-LSK109 kit (ONT, Littlemore, Oxford, UK) and sequenced on a PromethION Flow cell FLO-PRO002 with pore type R9.4.1, and with basecalling by Guppy v.3.0.5.

LRS pipeline: The raw data reads were cleaned using Chopper v.0.7.0 (De Coster & Rademakers, 2023) to remove reads shorter than 5000 bp and of quality less than 8, and Porechop v.0.2.4 (rrwick, 2018) to remove sequencing adapters and barcodes from the reads using the default settings. The quality and read length of the cleaned data

[8] https://slurm.schedmd.com

DOI 10.24823/Sibbaldia.2026.2130

were assessed using Nanoplot v.1.38.0 (De Coster *et al.*, 2018). Genome assembly was performed using Wtdbg2 (Ruan & Li, 2020) followed by sequence optimisation with Wtpoa-cns to generate fasta files. Genome assembly statistics were generated using QUAST v.4.6.3 (Gurevich *et al.*, 2013). To enable parallelisation for speeding up the analyses, the assembly was then divided into 200 sub-files using SeqKit (Shen *et al.*, 2016). BLAST v.2.2.29 (Jain *et al.*, 2015) was used to assign taxonomic identities to the scaffolds. The cleaned reads were mapped onto the genome scaffolds to obtain coverages using Minimap2 v.2.17 (Li, 2018). Samtools v.1.9 (Li *et al.*, 2009) was used to sort and index the reads. The assembled genome scaffolds, coverage data and BLAST result files were fed into BlobTools v.1.1.1 (Laetsch & Blaxter, 2017), which was then used to visualise the results and to generate blobDB tables.

Combined with National Centre for Biotechnology Information (NCBI) data, the three tables (BLAST output files, coverage and blobDB) were collated and the data visualised using ggplot2 v.3.3.2 (Wickham, 2016) in R v.4.4.1 (R Core Team, 2024). The pipeline is illustrated in Fig. 1.

SRS pipeline for Aeschynanthus angustifolius *Illumina data*

Data source: The SRS data were previously obtained for a study by Paton (2023). In brief, the genomic DNA library for *Aeschynanthus angustifolius* was generated using the NEBNext Ultra II DNA library prep kit (New England Biolabs, Ipswich, MA, USA) at RBGE. 150 bp paired-end sequencing on the Illumina Novaseq platform (Illumina, San Diego, CA, USA) was carried out at the Exeter Sequencing Facility (University of Exeter, Exeter, UK).

SRS pipeline: FastQC v.0.12.1 (Andrews, 2024) was employed to assess the quality of sequencing reads, after which Trimmomatic v.0.39 (Bolger *et al.*, 2014) was used to trim poor-quality reads and adapters, resulting in both paired and unpaired data. After trimming, the quality of the paired reads was re-evaluated using FastQC. Next, the genome was assembled using SOAPdenovo2 (Luo *et al.*, 2012), and the assembly quality was assessed with QUAST. As for the LRS pipeline, to speed up the analysis the assembly file was then divided into 400 sub-files using Seqkit; these were analysed using BLAST, and the output files were merged using the Linux cat command. In addition, an index was created for the assembly file; this was aligned with the paired reads using BWA-MEM, producing the SAM file. This file was then converted to a BAM file using Samtools, which also sorted and indexed the paired reads. BlobTools was utilised to organise and analyse the assembly data, BLAST results and BAM file, generating a JSON formatted file and a coverage table. Blobplot was then applied to create visual plots and the blobDB table. The data from three tables (BLAST output files, coverage and blobDB) were arranged, with NCBI data, into a comprehensive list. Finally, visualisation plots for this list were created in R using ggplot2. Fig. 2 illustrates the SRS pipeline used in this study.

Quantification and identification of contaminant diversity

To assign an identity to the contaminant scaffolds, BLAST database searches were carried out. The BLAST algorithm calculates bit-scores and *e*-values using a proprietary algorithm.[9] BLAST results are ordered primarily by bitscore and secondarily by

[9] https://blast.ncbi.nlm.nih.gov, www.metagenomics.wiki/tools/blast/evalue

DOI 10.24823/Sibbaldia.2026.2130

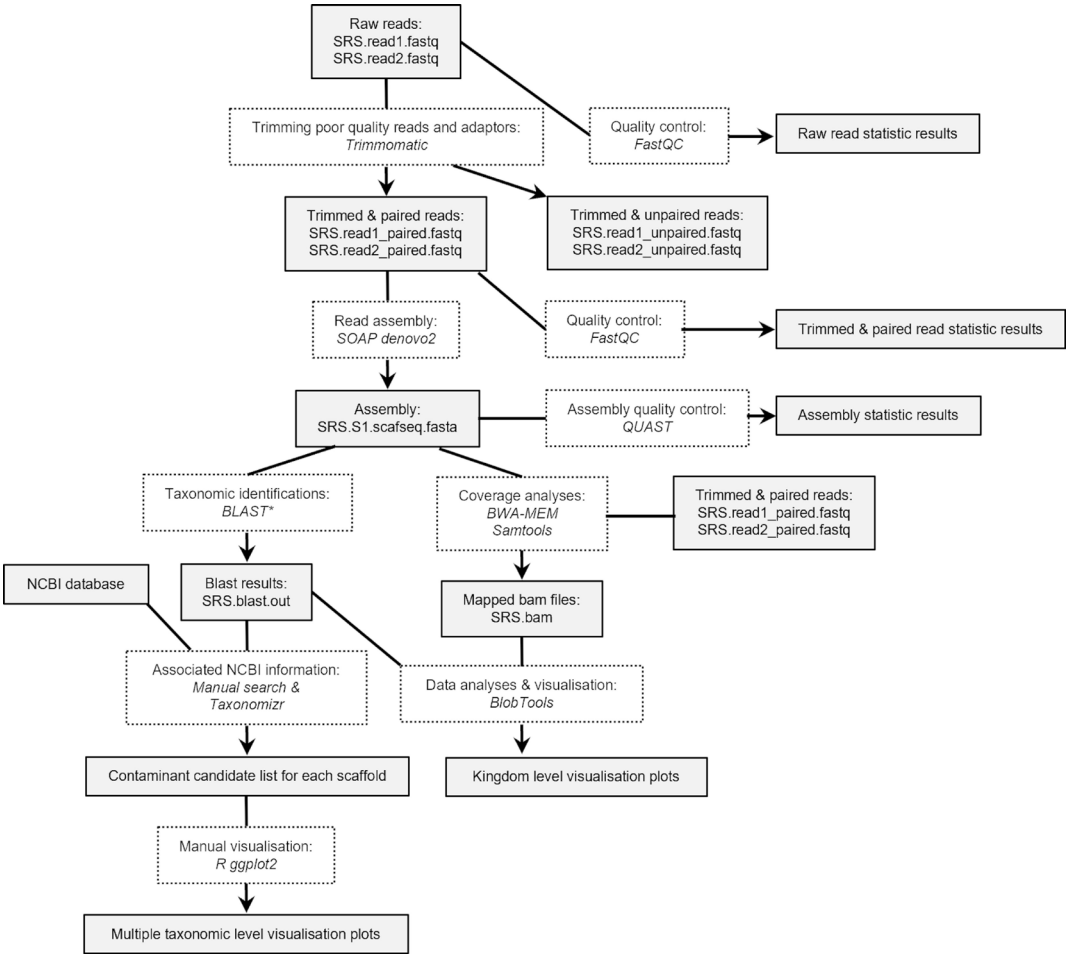

Fig. 1 Pipeline for analysing contaminants in long-read sequencing data of *Streptocarpus rexii*. Direction of workflow is indicated by arrows. Data and results are shown in shaded boxes and the bioinformatic steps and programs are in unshaded, dashed boxes.

e-value. A higher bit score or lower *e*-value indicates greater similarity between the query and subject sequences. The BLAST result values of the top 10 hits for each scaffold were plotted using the R ggplot2 package (Wickham, 2016), and Pearson correlation analyses were carried out using the R smplot2 package (Min & Zhou, 2021). The first hit of the BLAST searches for each scaffold identified as a contaminant among the genome assemblies were quantified as the number of scaffolds at phylum and genus level within bacteria, fungi and other domains, and tabulated or plotted in Microsoft Excel.

For the selection of candidates to ascertain contaminants to species level we selected a threshold of >95% identity and >250 bp alignment lengths. This was conservative, as in previous work 97% identity was used for identification (e.g. Pappalardo *et al*., 2025). Where more than one scaffold resulted, the one with the highest bitscore was retained.

DOI 10.24823/Sibbaldia.2026.2130

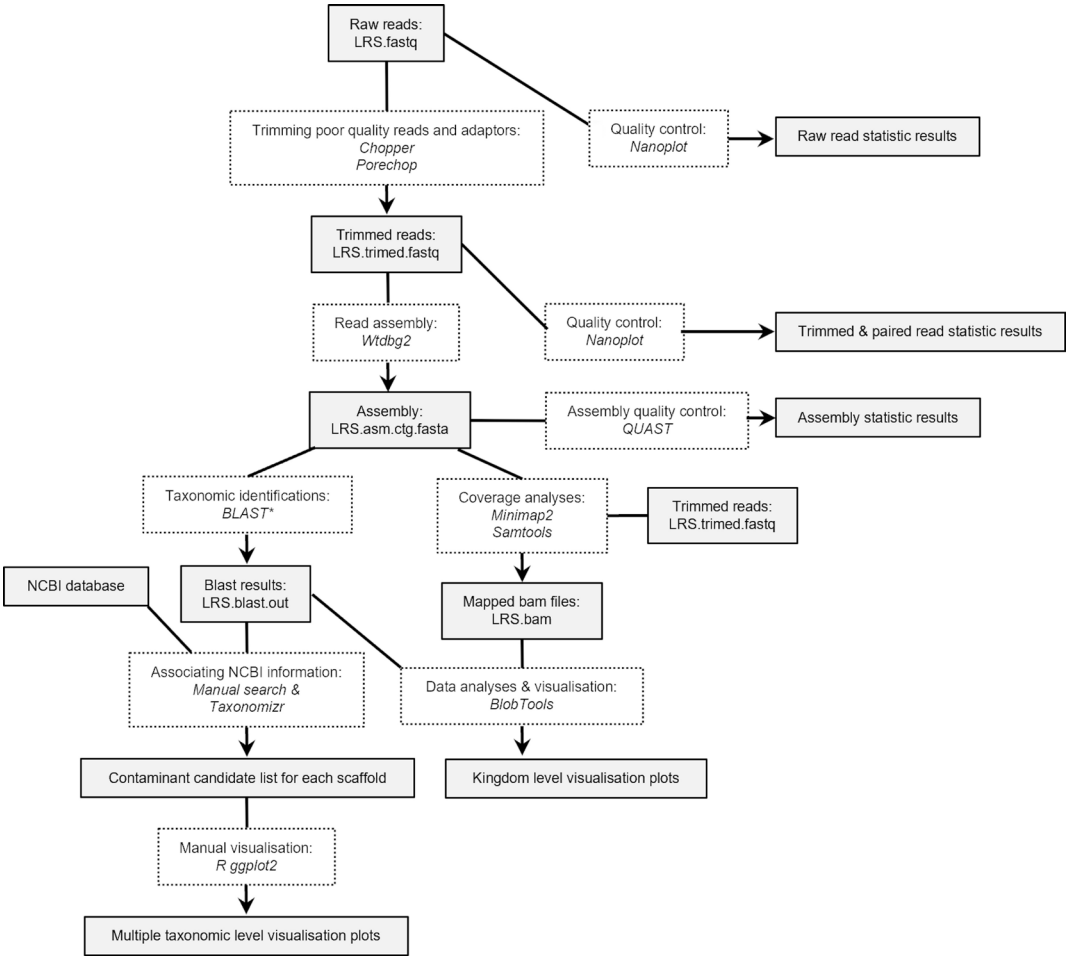

Fig. 2 Pipeline for analysing contaminants in short-read sequencing data of *Aeschynanthus angustifolius*. Direction of workflow is indicated by arrows. Data and results are shown in shaded boxes, and main analytical steps and programs used are in unshaded, dashed boxes.

Glasshouse pests and pathogens surveys

Sample collection: Every two weeks, contaminants (visible pests, pathogens and cohabitants of *Streptocarpus rexii*) were collected from *S. rexii* plants cultivated in the living collection at RBGE.

Pathogen investigation: In a preliminary study to investigate the fungal leaf symbionts and cohabitants of *Streptocarpus rexii* and for comparison with the BLAST results, six leaf pieces of the interface between green and necrotic parts of two *S. rexii* leaves were excised. The genome investigation of *S. rexii* had used unsterilised leaf material, so both unsterilised (where fast-growing fungal contaminants might overwhelm the petri dish) and sterilised material was used. Samples were either directly plated onto 1% Malt Extract Agar (MEA; Fluka, Sigma Aldrich, Darmstadt, Germany) supplemented with 0.3 g/L streptomycin sulphate (Sigma Aldrich, Darmstadt, Germany), or surface sterilised before plating (one minute in 70% ethanol, five minutes in 1% sodium hypochlorite (NaOCl), and finally 30 seconds in 70%

ethanol). The plates were stored in Ziploc® bags at room temperature in laboratory conditions and checked regularly for two weeks for fungal growth.

Image acquisition: Images of pests were taken under a Zeiss Stemi 2000-C stereomicroscope (Jena, Germany) and captured using a Zeiss Labscope v.4.0. For algae and fungi, material was mounted in distilled water on microscope slides and images acquired using a Zeiss Axiophot compound microscope fitted with a Zeiss AxioCam MRc5, or a Leica DM2500 (Wetzlar, Germany) compound microscope fitted with a Leica CCD camera DFC450. Occasionally, images were also captured using the camera option on mobile phones. The pests were visually identified by M. Philips, entomologist and Herbarium Digitiser at RBGE, using identification keys and standard references (e.g. Blackman, 2010).[10]

Results

Genome assembly and contaminant detection

Information on read quality, genome assembly, quality control and contaminant detection with 70% BLAST identity threshold of the NGS data for the two analysed plant species are provided in Supplementary material A. In brief, the *Streptocarpus rexii* genome assembly resulted in 8,681 scaffolds of which 667 (7.7%) were contaminants (Supplementary material A tables S3 & S4). For *Aeschynanthus angustifolius*, the assembly had 3,942,265 scaffolds of which 180 (0.005%) were identified as contaminants (Supplementary material A tables S7 & S8). In the BLAST searches, there was a strong linear correlation between bitscore (a statistical measure of sequence similarity) and sequence alignment length (Fig. 3A & B). There was a weak correlation between bitscore and percentage identity for the *S. rexii* data, but not for the *A. angustifolius* data, probably due to their short scaffold lengths (Fig. 3C & D). Extreme outliers were scaffold ctg1050 with a bitscore of 154,000 and alignment length of 97,332 bp (Fig. 3A, C & E; green circle). The coverage analysis revealed a high presence of several scaffolds: ctg170 had a long alignment length of 6,200 bp and high read coverage of 2,500× (Fig. 3E & G; red circle), while scaffold 104330 had a long alignment length of 1,395 bp and a high read coverage of 1,036× (Fig. 3F & H; blue circle).

Streptocarpus rexii *LRS dataset*

Contaminant diversity: The 667 contaminant scaffolds found in the *Streptocarpus rexii* LRS dataset came from the phyla Acidobacteriota, Actinomycetota, Arthropoda, Ascomycota, Basidiomycota, Pseudomonadota and Streptophyta (Table 1). Bacteria (Acidobacteriota, Actinomycetota, Mycoplasmatota and Pseudomonadota) made up approximately two-thirds of the contaminants, while fungi (Ascomycota and Basidiomycota) represented around one quarter. Overall, Pseudomonadota was the dominant phylum with 269 scaffolds (40.3%), followed by Actinomycetota with 154 scaffolds (23.1%) and Basidiomycota with 108 scaffolds (16.2%). Less numerous were Ascomycota, Arthropoda and Acidobacteriota, with 62, 38 and 30 scaffolds respectively. There was only one scaffold from Mycoplasmatota.

Diversity at genus level among the contaminant scaffolds indicated some dominant genera within phyla. For example, in the kingdom Bacillati, phylum Actinomycetota, *Mycobacterium* Lehmann

[10] See also https://influentialpoints.com/Gallery/Aphid_genera.htm

DOI 10.24823/Sibbaldia.2026.2130

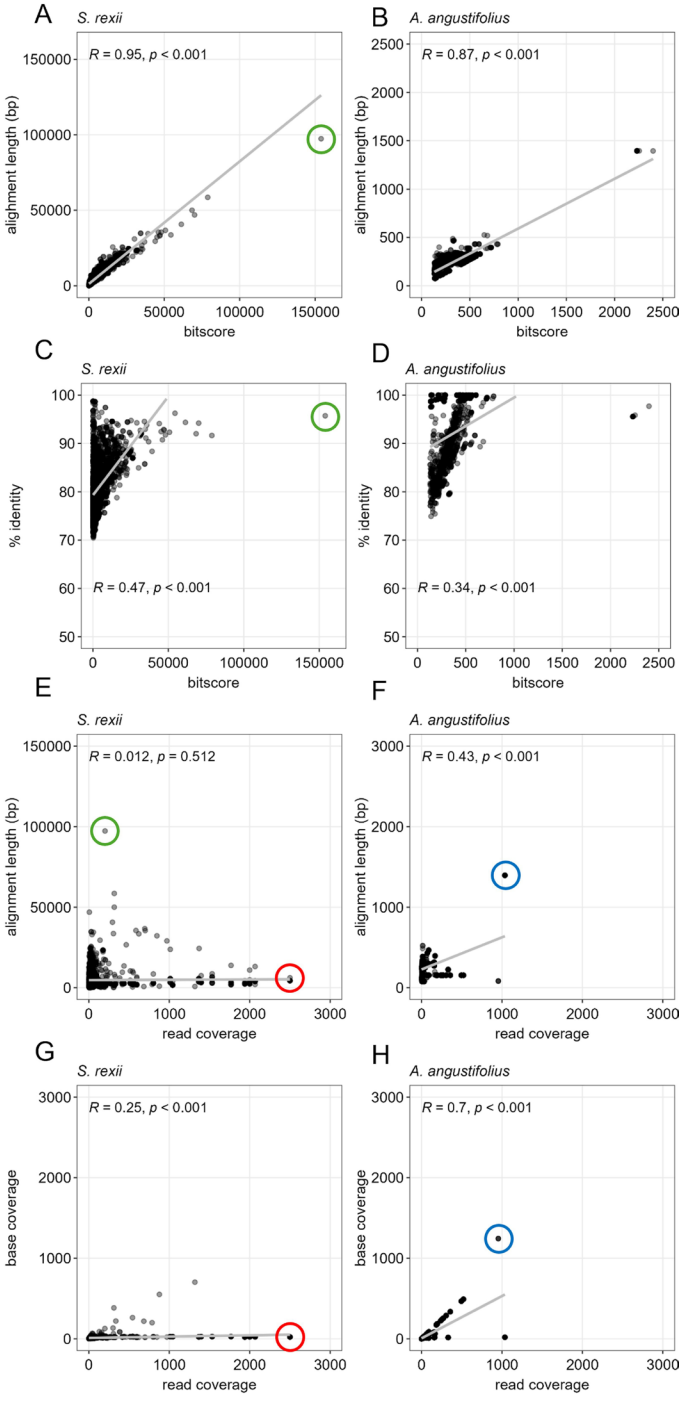

Fig. 3 Relationships between various BLAST selection criteria and coverage analyses for contaminant scaffolds in long-read sequencing (LRS) *Streptocarpus rexii* and short-read sequencing (SRS) *Aeschynanthus angustifolius* NGS datasets. **A & B** alignment length versus bitscore. **C & D** percentage identity versus bitscore. **E & F** read coverage versus alignment length. **G & H** read coverage versus base coverage. **A, C, E & G** LRS-based *S. rexii*. **B, D, F & H** SRS-based *A. angustifolius*. The grey line and the formula in the plots are results of the Pearson correlation analyses. Extreme outliers are circled: green = ctg1050; red = ctg170; blue = scaffold104330.

DOI 10.24823/Sibbaldia.2026.2130

& Neumann and *Frondihabitans* Greene *et al.* were most numerous, with 63 and 35 scaffolds (16.8% and 9.4%) respectively (Fig. 4A); within phylum Pseudomonadota in kingdom Pseudomonadati, *Burkholderia* Yabuuchi *et al.* was the most dominant, with 60 scaffolds (16%) (Fig. 4B). Among Ascomycota and Basidiomycota fungi, *Exophiala* J.W. Carmich., with 38 scaffolds (22.4%), and *Meira* Boekhout *et al.* ex Denchev & T. Denchev, with 92 scaffolds (54.1%), were most prominent respectively (Fig. 4C & D).

Species-level identification of contaminants: Unique contaminants identified at species level among the 667 contaminant scaffolds with >95% identity and >250 bp alignment length in BLAST searches came from four scaffolds (Supplementary material B table S1). Their bitscore ranged from 1,410 to 54,431, alignment lengths from 842 bp to 33,641 bp, identity from 95.47% to 98.66% and read coverage from 49× to 385×. The highest bitscore was calculated for scaffold ctg976, identified as *Candidatus Mycobacterium wuenschmannii* Zeineldin *et al*., with 96.25% identity and an alignment length of 33,641 bp and 306× read coverage. The next highest was scaffold ctg3883, with bitscore 8,288, alignment length 4,897 bp and 49× read coverage, identified as *Spiroplasma phoeniceum* Saillard *et al*., a plant pathogen on periwinkle (*Catharanthus* G. Don). Scaffold ctg7708 had a bitscore of 5,830, an alignment length of 3,565 bp and a read coverage of 385×, and was identified as a 'black yeast' (Chaetothyriales M.E. Barr sp. MCRE12). The fourth was scaffold ctg5577, which received several BLAST hits, all from Arthropoda, representing several mealybug species; *Planococcus citri* Risso (citrus mealybug) had the highest bitscore of 3,862 and the longest alignment length of 2,448 bp (identity 95.47%). *Pseudococcus jackbeardsleyi* Gimpel & Miller (Jack Beardsley mealybug) had the highest identity with 98.66%, though with a much shorter alignment length of 822 bp for the same scaffold.

Aeschynanthus angustifolius *SRS dataset*

Contaminant diversity: The majority of the 180 scaffolds identified as contaminants in the *Aeschynanthus angustifolius* SRS dataset came from the bacteria phyla Actinomycetota (57 scaffolds = 31.7% of all scaffolds) and Pseudomonadota (63 = 35%) (Table 1). Less numerous were the fungus phylum Basidiomycota (12 = 6.7%) and Metazoan Chordata (11 = 6.1%).

The diversity at genus level among the 180 scaffolds indicated that the most scaffolds (58 scaffolds = 32.2%) came from genus *Sphingomonas* Yabuuchi *et al.* in phylum Pseudomonadota in kingdom Pseudomonadati (Fig. 5B). Among Bacillati, genus *Microbacterium* Orla-Jensen in phylum Actinomycetota was the most numerous (39 = 21.7%). Both genera were by far the most numerous in their respective kingdoms. Among fungi, the Basidiomycota genus *Cryptococcus* Vuill. was most often found, with three scaffolds (Fig. 5).

Species-level identification of contaminants: Contaminant identification at species level from scaffolds with over 95% identity and >250 bp alignment lengths in BLAST searches yielded 15 scaffolds (Supplementary material B table S2). The highest bitscores were obtained for scaffold 104330 which received hits for several Basidiomycota, with the highest of 2,398 for *Papiliotrema terrestris* Crestani *et al.* ex Xin Zhan Liu *et al.* (a biocontrol agent) with 97.71% identity and a 1,395 bp alignment length. It also received the highest read coverage of 1,036×. Five other organisms

DOI 10.24823/Sibbaldia.2026.2130

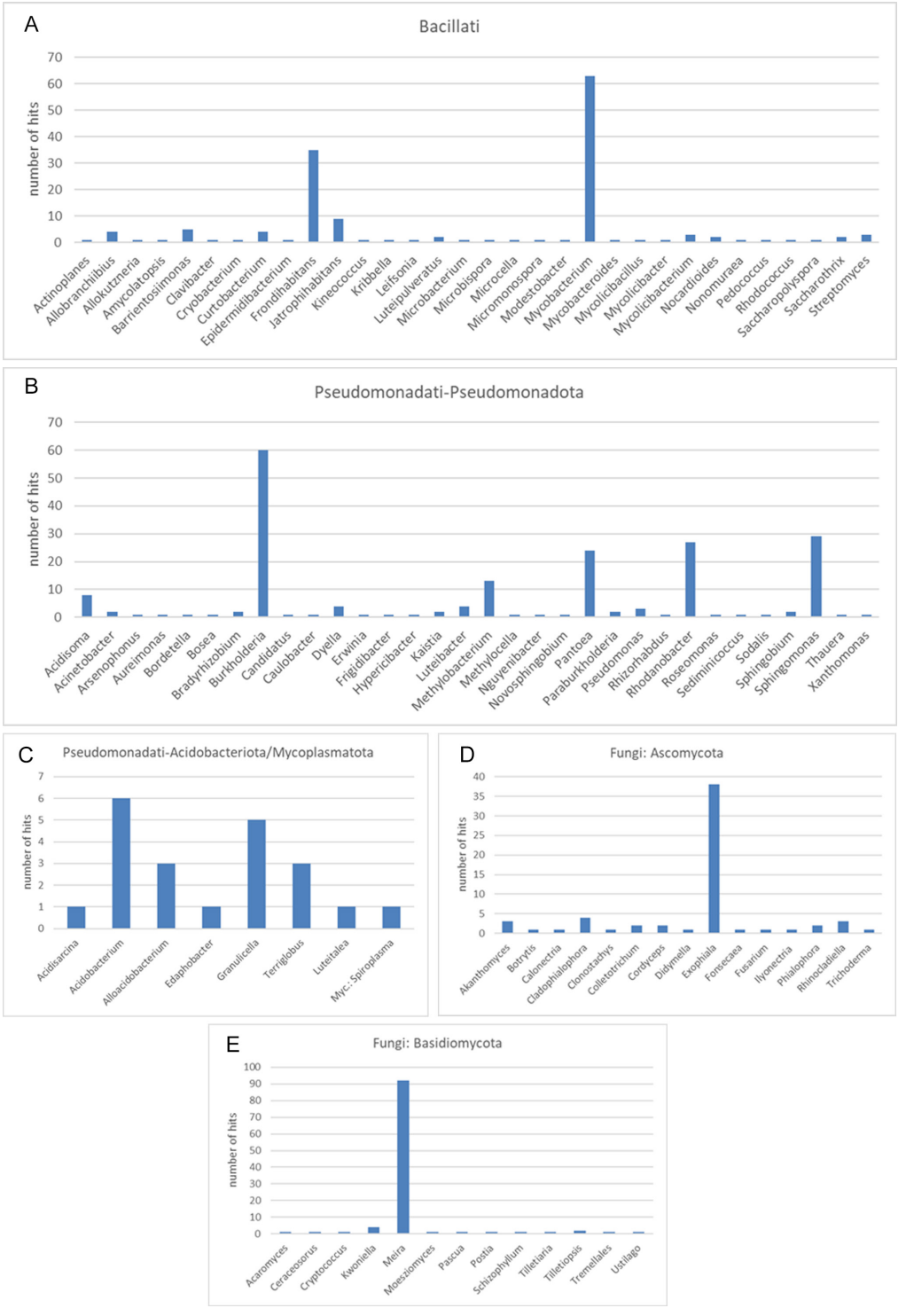

Fig. 4 Bar charts illustrating the diversity of contaminant scaffolds at genus level (at 70% confidence) in bacteria and fungi identified in *Streptocarpus rexii* long-read sequencing data. **A** Bacillati. **B** Pseudomonadota-Pseudomonadati. **C** Acidobacteriota/Mycoplasmatota. **D** Ascomycota. **E** Basidiomycota.

DOI 10.24823/Sibbaldia.2026.2130

DOI 10.24823/Sibbaldia.2026.2130

Table 1 Number of scaffolds and proportion of contaminants detected in *Streptocarpus rexii* and *Aeschynanthus angustifolius* at phylum level.

Domain	Kingdom	Phylum	*Streptocarpus rexii*	Proportion per phylum (%)	Proportion overall (%)	*Aeschynanthus angustifolius*	Proportion per phylum (%)	Proportion overall (%)
Bacteria	Bacillati	Actinomycetota	154	33.8	23.0	57	45.6	31.6
Bacteria	Bacillati	Bacillota	0	0.0	0	2	1.6	1.1
Bacteria	Bacillati	Mycoplasmatota	1	0.2	0.2	0	0.0	0
Bacteria	Thermotogati	Deinococcota	0	0.0	0	3	2.4	1.7
Bacteria	Pseudomonadati	Acidobacteriota	30	6.6	4.5	0	0.0	0
Bacteria	Pseudomonadati	Pseudomonadota	269	59.1	40.3	63	50.4	35.0
Bacteria	unknown	unknown	1	0.2	0.2	0	0.0	0
			Sum 455	*Sum 100.0*	*Sum 68.2*	*Sum 125*	*Sum 100.0*	*Sum 69.4*
Eukaryota	Fungi	Ascomycota	62	36.5	9.2	2	14.3	1.1
Eukaryota	Fungi	Basidiomycota	108	63.5	16.2	12	85.7	6.6
			Sum 170	*Sum 100.0*	*Sum 25.4*	*Sum 14*	*Sum 100.0*	*Sum 7.7*
Eukaryota	Metazoa	Annelida	0	0.0	0	1	2.4	0.6
Eukaryota	Metazoa	Arthropoda	38	90.5	5.7	1	2.4	0.6
Eukaryota	Metazoa	Chordata	0	0.0	0	11	26.8	6.1
Eukaryota	unplaced	Gyrista	1	2.4	0.2	0	0.0	0
Viruses	Heungongvirae	Peploviricota	0	0.0	0	1	2.4	0.6
Viruses	Heungongvirae	Uroviricota	0	0.0	0	2	4.9	1.1
Others	Others	Others	3	7.1	0.5	25	61.0	13.9
			Sum 42	*Sum 100.0*	*Sum 6.4*	*Sum 41*	*Sum 100.0*	*Sum 22.9*
Total			667		*100.0*	180		*100.0*

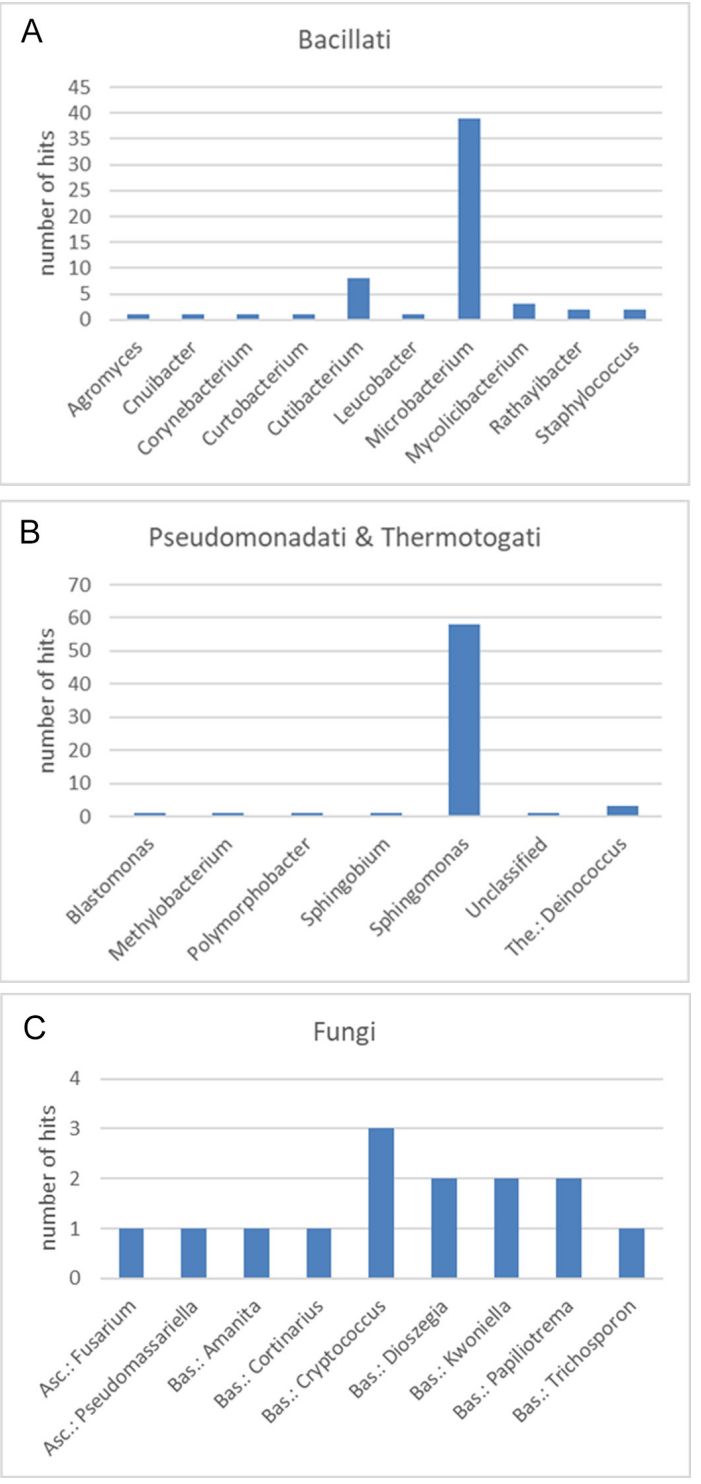

Fig. 5 Bar charts illustrating the diversity of contaminant scaffolds at genus level (at 70% confidence) in bacteria and fungi in *Aeschynanthus angustifolius* SRS data. **A** Bacillati. **B** Pseudomonadati & Thermotogati. **C** Fungi. Asc. = Ascomycota; Bas. = Basidiomycota; The. = Thermotogati.

DOI 10.24823/Sibbaldia.2026.2130

had similarly high values for this scaffold. The remaining scaffolds received much lower bitscores of less than 800, shorter alignment lengths (265–430 bp) and low read coverages (6–71×) but received up to 100% identity and included mostly bacteria, one Chordata hit and a virus.

Glasshouse pests and pathogens and their bioinformatic identification

A total of 17 Arthropods (15 Insecta and 2 Arachnida) were collected from *Streptocarpus rexii* leaves including a range of common glasshouse pests (Fig. 6). Among the BLAST hits (Supplementary material B table S1) only *Planococcus citri* overlapped, with 95.47% identity (Fig. 6B). Of the other Arthropods collected, such as aphids, thrips and whitefly, no BLAST hits were found.

Fungal colonies isolated from *Streptocarpus rexii* leaf material after one week (Fig. 7A & B) and two weeks (Fig. 7C & D) indicated the presence of *Trichoderma* Pers. observed by morphological identification of colonies from uncleaned specimens (Fig. 8A & B). Another colony produced appressoria-like structures but no spores (Fig. 8C) and thus identification to species level was not possible. No DNA-based identification was undertaken on any of the colonies which grew out of the leaf material. *Trichoderma* hits were found in the *S. rexii* BLAST list but with low sequence identities: four species were implicated from *S. rexii* LRS data, *T. asperellum* Samuels *et al.* (ctg1995: 81.94%; 310 bp alignment), *T. citrinoviride* Bissett (ctg5236: 81.74%; 219 bp alignment), *T. reesei* E.G. Simmons (ctg3518: 76.97%; 2,183 bp alignment) and *T. simmonsii* P. Chaverri *et al.* (ctg6420: 80.47%; 722 bp alignment). However, only the BLAST result of scaffold ctg3518, *T. reesei*, appeared as a first hit (Supplementary material D).

Discussion

The present study was exploratory, primarily to design bioinformatic pipelines for the detection and identification of contaminants through NGS data mining. In this study, genomic data from different sources and NGS approaches were used: genome data of *Streptocarpus rexii* obtained through ONT LRS and of *Aeschynanthus angustifolius* based on Illumina SRS. Overall, the ONT data resulted in longer scaffolds (Supplementary material A) and thus longer alignments in BLAST searches for contaminants (Fig. 3), presumably more reliable for species identification, whereas the shorter assembly scaffolds for the Illumina *A. angustifolius* reads resulted in more no-hits (Supplementary material A tables S4 & S8). The read coverage, which could be interpreted as the abundance of the organisms' DNA, was overall higher for the LRS data compared with the SRS data (Fig. 3C & D), with potential exceptions (see below). However, this can be misleading when hits represent multicopy gene regions, such as nuclear ribosomal DNA (rDNA) that occurs in hundreds of copies within a cell; for example, the five hits in scaffold ctg5577 for mealybugs all came from the 28S ribosomal DNA gene and show a high read coverage (>330×). Some further considerations for the interpretation of the results would be prudent and are discussed below.

Bioinformatics developments

With the emerging technologies for whole genome sequence acquisition, analytical bioinformatic tools are being developed rapidly alongside. BlobTools, with which we scanned across entire genomes of contaminants, has its advantages as the entire genome is theoretically available as queries. A disadvantage, however, could be an overestimation of contaminant species from

DOI 10.24823/Sibbaldia.2026.2130

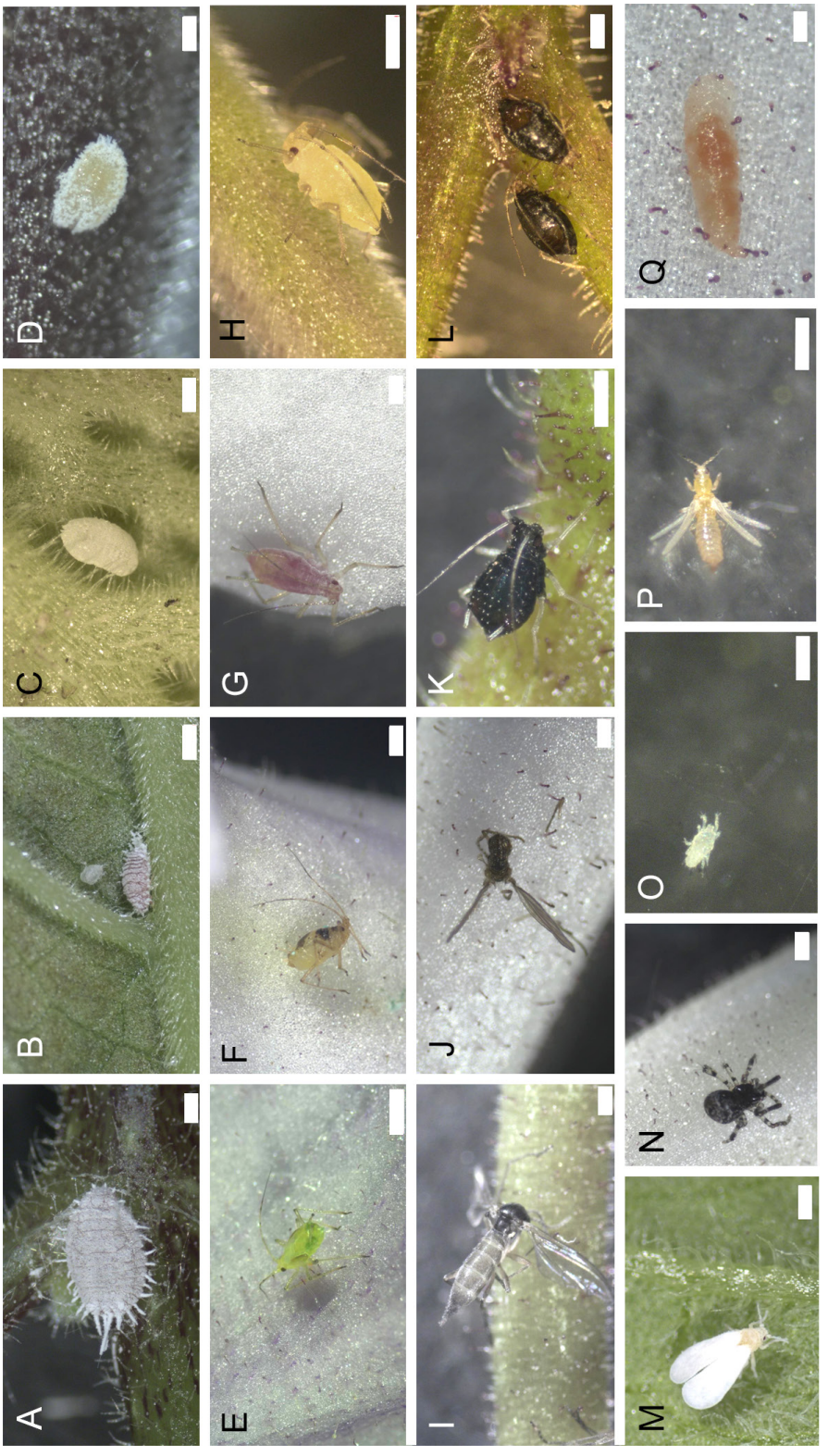

Fig. 6 Images of insects observed on *Streptocarpus rexii*. **A** *Planococcus* sp. **B** *Planococcus citri*. **C** cf. *Rhizoecus dianthii*. **D** cf. *Paracoccus marginatus*. **E** *Aulacorthum solani*. **F** *Neomyzus circumflexus*. **G** *Acyrthosiphon malvae*. **H** *Aulacorthum* sp. **I** *Sciaridae* sp. **J** *Diptera* sp. **K** *Idiopterus nephrelepidis* (black fern aphid). **L** *Aphis* sp. **M** *Trialeurodes vaporariorum* (glasshouse whitefly). **N** Unknown arachnid. **O** *Tetranychidae* sp. **P** *Thysanoptera* sp. **Q** *Aphidoletes aphidimyza* larva (predatory gall midge/aphid midge). Scales: A, C, F–N & P = 0.5 mm; B & E = 1 mm; D, O & Q = 0.2 mm.

DOI 10.24823/Sibbaldia.2026.2130

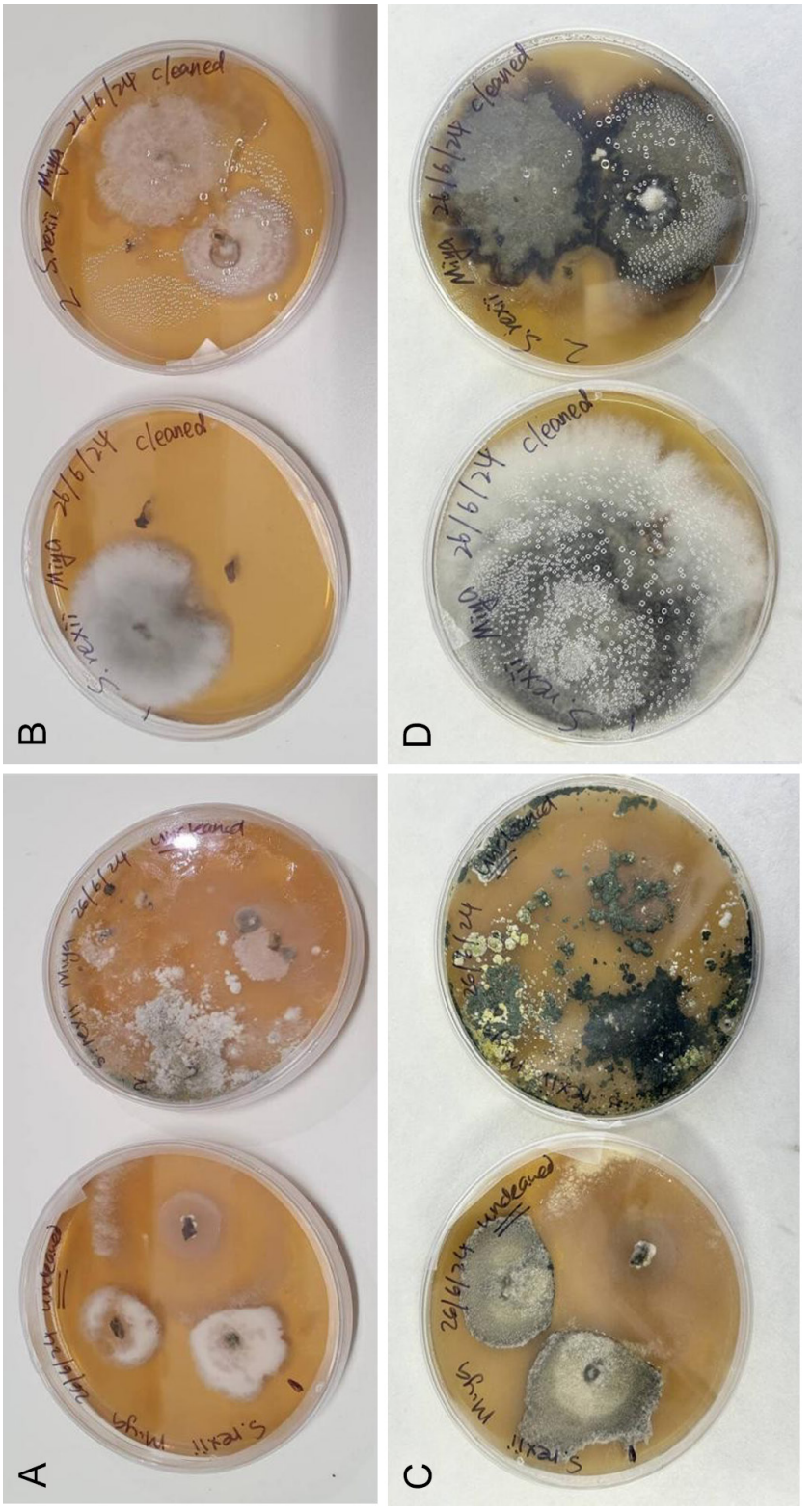

Fig. 7 Assessment of leaf samples of *Streptocarpus rexii* for fungal colonies. **A** & **B** after one week (A) without surface sterilisation and (B) with surface sterilisation. **C** & **D** after two weeks (C) without surface sterilisation and (D) with surface sterilisation.

DOI 10.24823/Sibbaldia.2026.2130

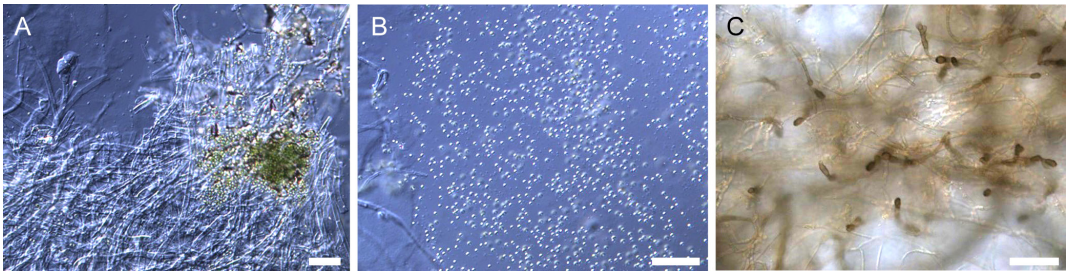

Fig. 8 Images of fungal structures from two-week-old colonies isolated from leaves of *Streptocarpus rexii* from samples (A & B) without surface sterilisation and (C) with surface sterilisation. **A** *Trichoderma* hyphae and conidia. **B** *Trichoderma* conidia. **C** Appressoria-like structures in culture. Scales: A–C = 50 µm.

several unique scaffolds. One way around this could be to utilise specific genome regions as in Bard *et al.* (2025), who targeted the ITS ribosomal DNA reads in genome data. BlobTools has also been developed into BlobToolKit, which now includes tools for identifying and isolating contaminant data in draft and publicly available assemblies (Challis *et al.*, 2020). The use of this toolkit would perhaps speed up the workflow and should be investigated in the future.

Sampling limitations

Our study used existing genome datasets that were obtained with the aim of assembling plant genomes, rather than to specifically identify and quantify contaminants in RBGE glasshouses. As such, the leaf material collected for these NGS studies was superficially healthy and clean, and as such largely 'free' of external contaminants. They were also taken from only two plants from separate glasshouses and thus do not represent RBGE's entire glasshouse area. Therefore, our results are a small snapshot of the non-plant biological diversity harboured in RBGE glasshouses. This will largely explain the mismatch between, for example, identified (Supplementary material B) and observed pests and pathogens (Figs 6–8). Nonetheless, the developed pipelines are applicable to any NGS data and are thus useful. Furthermore, we identified bacterial and fungal contaminants not observed visually in the glasshouses, making this a useful exercise (see below).

Considerations for contaminant identification

For identification purposes, alignment length and level of identity in BLAST searches may provide an indication of identification accuracy. Mealybugs were visually identified in our glasshouse survey, and among the implicated mealybugs through our bioinformatics pipeline, we found one, *Planococcus citri*, with high identity (95.47%) and long alignment length (2,448 bp, ctg5577, *Streptocarpus rexii*). Overall, the longer a BLAST alignment, the more certain an identification may be. The reason for the lower identity for this species (95.47%) in BLAST compared with the other species hits (97.27–98.66%) may lie in the fact that its query scaffold was longer (genome assembly OX465509.1, GenBank) than the rDNA sequence of the other species in GenBank (~830 bp) and may have included less conserved regions. It could also be explained by differences in the origin of the mealybugs in GenBank and in RBGE's glasshouses and reflect the species' population-level diversity,

as it was shown that within mealybugs 28S ribosomal RNA similarity can be as low as 93.2% (Abd-Rabou *et al*., 2012).

A critical point is the threshold for species identification which, we guess, depends on the genetic variation within a species separating it from the next closest one and which could be as low as 0.05% and as high as 5% nucleotide diversity, depending on gene, gene region and the organism's life cycle (Buckler IV & Thornsberry, 2002; Qin *et al*., 2021). Sequencing errors are another potential source of uncertainty of contaminant identification from the genome scaffolds, despite NGS sequencing accuracy improving over time. Sequencing accuracy tended to be higher in SRS compared to LRS, however PacBio has developed HiFi LRS and ONT have developed a new chemistry (R10) for more accurate basecalling (i.e. calling nucleotide data from raw ionic current data) (Dohm *et al*., 2020; Kanwar *et al*., 2021, cited in CD Genomics, 2025).

Furthermore, as the taxon coverage of GenBank is incomplete, searches of available accessions will potentially not include the query taxon and, if this is absent, would instead return results for the nearest match. In contrast, some gene regions are known to be invariable between species (e.g. ITS: the internal transcribed spacer of the ribosomal DNA region is identical for several species of *Streptocarpus* (Bowie ex Hook.) Lindl. (Möller & Cronk, 2001)). This is also demonstrated in our study where, for example, scaffold 20564817 received hits from several species all with 100% identity (Supplementary material E). In such cases it is difficult to ascertain the identity of the contaminant. Ideally, a complete database for comparison is desirable and can only be compiled over time.

A final point to consider in the interpretation of the results is the source of contamination, which could be natural (i.e. pre-existing) or could represent 'hitchhikers' during cultivation (Hernández-Tasco *et al*., 2023) or contaminations post-sampling of the leaves (RBGE, internal and/or external sources). Depending on the source, the contaminants will have differences in their biology which may aid identification.

However, despite these limitations some assumptions can be made on the identities of the contaminants that were found in our exploratory study.

Quantity versus diversity

It is difficult to differentiate exactly between quantity among the identified contaminants due to uncertainties in identification to species level (e.g. identity and alignment length cutoff points; genetic depth of species; database limits due to possible misidentification and mix-ups). Therefore, we chose a compromise and presented diversity at the genus level, whereby a high score did not equate to high diversity but could also be due to several scaffolds identifying the same species (for example, the *Streptocarpus rexii* contaminant *Microbacterium testaceum* (Komagata & Iizuka) Takeuchi & Hatano increased the genus score, although it was the top hit in 19 scaffolds) (Supplementary material D). However, read or base coverage for the same scaffold may be a better indicator of the quantitative level of the presence of a contaminant, and these varied greatly among the scaffolds (Fig. 3C & D). One should be thus careful equating quantity with diversity.

Contaminant domains of Aeschynanthus angustifolius *and* Streptocarpus rexii

Despite the great differences in the total number of contaminant scaffolds found in

DOI 10.24823/Sibbaldia.2026.2130

the two plant species (Table 1), the most frequently detected contaminant domain was bacteria, with ~70% of hits (Table 1). In this domain, phylum Pseudomonadota comprised the major part, with 35% to ~40% of the contaminants from *Aeschynanthus angustifolius* and *Streptocarpus rexii* respectively, followed by Actinomycetota, which accounted for ~23% to ~32%. It was interesting to note that irrespective of the NGS approach used, the proportional contaminant profiles were similar at the domain level (Table 1). Similar results were found in previous studies on different plants, whether herbaceous (e.g. *Sinningia* Nees, Hernández-Tasco *et al.*, 2023 or *Thlaspi* L., Galanti *et al.*, 2024) or woody (e.g. *Quercus* L., Gathercole *et al.*, 2021) (Table 2). An even higher proportion – two-thirds – of Pseudomonadonta was found by Roman-Reyna *et al*. (2020). Also, the order of predominance within domains seems conserved among bacteria, with phylum Pseudomonadota showing higher proportions than phylum Actinomycetota.

Differences between Aeschynanthus angustifolius *and* Streptocarpus rexii

While within the bacteria domain the two plant species showed similar proportions of phyla Actinomycetota and Pseudomonadota, only *Streptocarpus rexii* had contaminants from phylum Acidobacteriota (Table 1), which are common in soil. The differences may be explained by the terrestrial habit of *S. rexii* as opposed to the epiphytic habit of *Aeschynanthus angustifolius*. The proportion of fungi phyla also differed between the two plant species, with *S. rexii* having twice as many Basidiomycota than Ascomycota contaminants, while *A. angustifolia*, though with fewer overall, had around seven times more Basidiomycota contaminants.

This unusual finding of the higher proportion of Basidiomycota for both plants (Table 1) is not usually observed (Vorholt, 2012; Rungjindamai & Jones, 2024) in metabarcoding or metagenomic studies of phyllosphere fungi (e.g. Schönrogge *et al.*, 2022; Hernández-Tasco *et al.*, 2023). However, close scrutiny of the data shows that the most frequent Basidiomycota genus is *Meira*, a yeast isolated from leaves and also mite-associated (Denchev & Denchev, 2021), occurring only on *Streptocarpus rexii*, which is also the only host of Cheyletidae (mites) in this study. The Ascomycota genus *Exophiala* J.W. Carmich. also occurs on *S. rexii*, but these are common environmental fungi found on soil and organic matter (Thitla *et al.*, 2022) and would possibly be found on unsterilised leaves of herbaceous plants that are in contact with the soil, as is the case with *S. rexii.* Both genera contribute to the proportionally high levels of fungi for this host when compared with other studies (e.g. Table 2). There are low levels of fungi associated with *A. angustifolius* (in comparison with other studies, Table 2), and of these many are Basidiomycota yeasts occurring on the phylloplane (*Cryptococcus*, *Dioszegia* Zsolt, *Kwoniella* Statzell-Tallman & Fell, *Papiliotrema* J.P. Samp. *et al.*).

The environment in which the plant occurs will be a contributing factor to the microbial community recorded. When whole-leaf microbiota of herbaceous plants in the natural environment (e.g. Ramakrishnan *et al.*, 2024; Yang *et al.*, 2024) are compared with those in glasshouse environments there can be increased levels of Basidiomycota in the latter (e.g. Geyer *et al.*, 2024). Interestingly, the most common fungal genus recorded by Geyer (2024) was *Trichoderma*, also recorded in the present study of glasshouse-raised plants. In a metagenomics study that involved greenhouse-grown peppers,

DOI 10.24823/Sibbaldia.2026.2130

DOI 10.24823/Sibbaldia.2026.2130

Table 2 Comparison of next-generation sequencing contaminants from the present and selected studies (%).

Organism			**Reference**				**Present study**	**Present study**
			Gathercole	Roman-Reyna *et al.* (2021)	Galanti *et al.* (2020)*	Hernández-Tasco *et al.* (2024)⁺		
Phylum (synonyms)	**Phylum**	**Kingdom**	*Quercus*	*Oryza*	*Thlaspi*	*Sinningia*	*Streptocarpus*	*Aeschynanthus*
Archaea			<1		0.15			
Bacteria – Euryarchaeota	= Methanobacteriota	Methanobacteriati		2.8				
Bacteria – Thermofilum	= Thermoproteota	Crenarchaeota		~1.4				
Bacteria – Acidobacteria	= Acidobacteriota (gram–)	Pseudomonadati	<1	none	0.31	0	3.9	0
Bacteria – Bacteroidetes	= Bacteroidota (gram–)	Pseudomonadati	12.13	1.5	3.62	1.9 (0.1–8.5)		
Bacteria – Proteobacteria	= Pseudomonadota	Pseudomonadati	44.55	67.3	54.76	77.0 (58–88)	40.7	35.0
Bacteria – Actinobacteria	Actinomycetota (gram+)	Bacillati	33.90	8.6	12.69	16.9 (3.4–31)	23.2	32.2
Bacteria – Spirochaetes	= Spirochaetota (gram–)	Pseudomonadati		1.4	0.04	0		
Bacteria – Firmicutes	= Bacillota	Bacillati	1.28	10.1	7.14	4.1 (0.8–16.6)	0	1.1
Bacteria – Tenericutes	= Mycoplasmatota	Bacillati		7.2	0.02		0.1	0
Bacteria – Deinococcus–Thermus	= Deinococcota	Thermotogati					0	1.7
Bacteria – Cyanobacteria	= Cyanobacteriota (gram–)			5.6	0.82			
Bacteria – other			<1					
Sum bacteria			***91.86***	***105.9°***	***79.55***	***[99.9[a]]***	***67.8***	***70.0***
Fungi – Ascomycota			6.20		8.24	37.9 (4.8–90)	9.6	1.1
Fungi – Basidiomycota			1.02		0.22	7.1 (0.2–22)	16.3	8.3
Fungi – other			<1					
Sum fungi			***7.22***	***nd***	***8.46***	***[45.0[b]]***	***25.9***	***9.4***
Arthropoda					8.25		5.7	1.7
Chordata					1.08		0	6.1
Sum metazoa			***nd***	***nd***	***9.33***	***[nd]***	***5.7***	***7.8***

* calculated from Supplementary file 2

⁺ calculated from proportion data in text (rounded); overall: 60.9% fungi and 39.1% bacteria

° their numbers in Fig.1

[a] sum within bacteria

[b] sum within fungi

nd = no data

however, Basidiomycota were infrequent and *Trichoderma* was not recorded (Jo *et al.*, 2021), demonstrating that host is also an influence.

At the genus level, most contaminant genera were unique to the two plants with only six out of 82 bacteria, and three out of 35 fungal genera shared, and none of Metazoa (Supplementary material C). These differences may be related to leaf longevity, with *Streptocarpus rexii* having short-lived leaves and *Aeschynanthus angustifolius* having long-lived leaves. For example, Chordata contaminants found in *A. angustifolius* might have been deposited externally on the leaves or introduced during leaf handling, DNA extraction or DNA sequencing. It is possible that the presence of more Chordata contamination in *A. angustifolius* is due to its longer leaf life. On the other hand, *S. rexii* has a short leaf lifespan, so there is less time for it to be contaminated with Chordata.

Selected contaminants

Our study identified some contaminants with high bitscores or of a specific nature. For example, ctg1050 with the highest bit score (154,000) and longest alignment length (97,332 bp) detected in the *Streptocarpus rexii* genome data (Supplementary material D) was identified as *Candidatus Mycobacterium wuenschmannii*, a slow-growing non-tuberculous mycobacterium isolated from the livers of Amazonian milk frogs, which may infect animals and humans (Zeineldin *et al.*, 2023). The same organism was identified for ctg976 with a high bitscore of 54,431 and a high identity of 96.25% (Supplementary material D table S1), and is thus unlikely to be an exact match to this mycobacterium. Our contaminant therefore probably represents an organism not yet included in the GenBank database.

A high-identity contaminant (ctg2980: 98.31% identity in 5S ribosomal DNA) was the fungus *Exobasidium vaccinii* (Fuckel) Woronin (red leaf disease), which may cause galls on *Vaccinium* L. and *Rhododendron* L. spp. (Piątek *et al.*, 2012). Since a plant of *Vaccinium cuneifolium* Miq. was growing in the same glasshouse at RBGE where *Streptocarpus rexii* was cultivated, it may explain the presence of a closely related species of this genus of biotrophic, host-specific fungi (as spores) on *S. rexii* leaves.

In the *Streptocarpus rexii* genome data, a high bitscore (3,862) for the presence of the citrus mealybug was found for scaffold ctg5577 (discussed above). This is a pest ubiquitous to most RBGE glasshouses but was not detected in the genome of *Aeschynanthus angustifolius*. Here, the leathery glabrous leaves of this species will have allowed the easy detection of the pest when sampling the leaf, which was much more difficult to avoid for the densely pilose leaves of *S. rexii*.

Spiroplasma phoeniceum Saillard *et al*. was found to have relatively high key indicators (ctg3883: 97.43% identity; 4,897 bp alignment length) among the contaminants of *Streptocarpus rexii*. This microaerophilic plant pathogen was identified as the causal agent of yellowing disease affecting periwinkle, *Catharanthus* G.Don (Saillard *et al.*, 1987), a plant that is widely grown as an ornamental and is in a tropical glasshouse at the RBGE nursery.

An ascomycete in the order Chaetothyriales was found in *Streptocarpus rexii* with a high bitscore, particularly based on a long alignment (3,565 bp) and high read coverage (385×), with the highest identity (96.44%) among the hits for ctg7708. The nearest match in GenBank was to a strain found in the carton nests of ants (Vasse *et al.*,

2017). This suggests the presence of a black yeast-related ascomycete with an unknown biology and identity.

The Basidiomycota yeast *Meira* occurs on leaves and can be mite-associated (Denchev & Denchev, 2021), and occurred only on *Streptocarpus rexii* based on over 90 scaffolds (Fig. 4E, Supplementary material D) exhibiting high values for both bit score and alignment length. The majority of percentage identity was around 85%, with some reaching >90%. Base coverage and read coverage values averaged roughly 7× and 15× respectively. *Meira* could well have originated from mite faeces, as mites (*Tetranychidae* sp.) were also reported only from this plant (Fig. 6).

Among contaminants in the *Aeschynanthus angustifolius* data (Supplementary material B table S2), the BLAST result with the highest bit score (2,398), alignment length (1,395 bp) and read coverage (1,036×) was for scaffold104330 that suggested the presence of *Papiliotrema terrestris*, with the GenBank submission (LR697097.1) originating from a soil-based yeast originally isolated from an epiphytic microbial community on apple fruits (Palmieri *et al.*, 2021). That organism has also been used as a biocontrol agent (Miccoli *et al.*, 2020). Whereas biocontrol agents have been used in the past on *Aeschynanthus* at RBGE (such as entomopathogenic *Akanthomyces muscarius* (Petch) Spatafora *et al.* Ve6; Ascomycota, Hypocreales), the unrelated *Papiliotrema* (Basidiomycota, Tremellales) had not been employed, and its presence is either a coincidence or represents a closely related species.

Within the *Aeschynanthus angustifolius* genome, most other scaffolds had rather low bitscores, alignment depth and/or read/base coverage, and their identification to species level is doubtful. Interestingly, some contaminants detected here were found in previous studies, such as the bacterium *Cutibacterium acnes* (Gilchrist) Scholz & Kilian (previously *Propionibacterium acnes* (Gilchrist) Douglas & Gunter) that Sangiovanni *et al.* (2019) found as a major contaminant in all their samples, as well as *Staphylococcus warneri* Kloos & Schleifer and *Streptococcus pneumoniae* (Klein) Chester / *Staphylococcus aureus* Rosenbach (one scaffold). Perhaps these are human-associated and more likely to be found and to accumulate on the long-lived leaves of this plant that is handled by people over longer periods of time.

Conclusions

Our approach was opportunistic, utilising existing genome datasets that included contaminants as 'bycatch' and that were primarily used to develop the bioinformatic pipelines for isolating and identifying contaminants from datasets based on SRS and LRS NGS methods.

While there is great potential in our pipelines to identify internal and external contaminants in plant genomic data, great care has to be taken in the interpretation of the results because of uncertainties: success is limited to the completeness and taxonomic accuracy of source sequence databases. Sufficiently stringent cut-off points are needed in order to limit candidate names; we used 95% sequence identity and >250 bp alignment length. Even then, scaffolds with conserved gene regions may provide several candidate names. However, their common/shared biology may allow their categorisation after sampling into external or internal contaminants, pathogens or secondary contaminants. If the aim is to determine contaminants in glasshouses for quarantine or plant health purposes, a suitably planned comprehensive sampling is also required.

DOI 10.24823/Sibbaldia.2026.2130

In our pilot study it was possible to identify some contaminants with a great level of certainty whereas others would require some follow-up work to achieve positive identifications. For example, the presence of the citrus mealybug was indicated both in visual observations and in genomic data and could represent a successful ID, while the presence of others, such as *Spiroplasma phoeniceum*, the periwinkle yellowing pathogen, would require further study for a positive ID.

Our study indicated a great potential for contaminant detection of our pipelines, and further work should be focused on a standardisation of identification and protocols to allow studies to be comparable. We feel that this technique is useful and relevant in the horticultural setting of a botanic garden and that it provides a novel method using contemporary technology to aid biosecurity and quarantine efforts.

Acknowledgements

We would like to thank RBGE for access to the laboratory facilities and glasshouse collections, and particularly Andrew Ensoll (horticulture) for plant cultivation. We are also grateful to Milo Phillips (RBGE), for identifying the glasshouse 'critters'. Scientific and Technical Services and Digital and Technology Services staff at RBGE provided essential support for carrying out research. Pete Hollingsworth (RBGE) is thanked for general support and facilitating research at RBGE, as is Akitoshi Iwamoto (Kanagawa University) for KN's research at Kanagawa University. Computation was largely performed on the Crop Diversity server, James Hutton Institute, Dundee, UK and we thank Iain Milne for facilitating these analyses. The analyses were partly carried out on the NIG supercomputer at ROIS National Institute of Genetics, Shizuoka, Japan. The authors further thank the NGS community on the web for supporting the many analysis programs running on the servers. RBGE is supported by the Scottish Government's Rural and Environmental Science and Analytical Services Division (RESAS).

References

ABD-RABOU, S., SHALABY, H., GERMAIN, J.F., RIS, N., KREITER, P. & MALAUSA, T. (2012). Identification of mealybug pest species (Hemiptera: Pseudococcidae) in Egypt and France, using a DNA barcoding approach. *Bulletin of Entomological Research*, 102: 515–523. doi: https://doi.org/10.1017/S0007485312000041

ANDREWS, S. (2024). FastQC: A quality control tool for high throughput sequence data. Available online: www.bioinformatics.babraham.ac.uk/projects/fastqc (accessed July 2024).

BARD, N.W., DAVIES, T.J. & CRONK, Q.C.B. (2025). Teknonaturalist: A Snakemake pipeline for assessing fungal diversity from plant genome bycatch. *Molecular Ecology Resources*, 25: e14056. doi: https://doi.org/10.1111/1755-0998.14056

BEHJATI, S. & TARPEY, P.S. (2013). What is next generation sequencing? *Archives of Disease in Childhood–Education and Practice*, 98: 236–238. doi: https://doi.org/10.1136/archdischild-2013-304340

BGCI (2022). Guide to plant biosecurity in botanic gardens and arboreta. Available online: www.bgci.org/wp/wp-content/uploads/2022/08/BGCI-Guide-to-Biosecurity-in-Botanic-Gardens-and-Arboreta.pdf (accessed July 2024).

BLACKMAN, R.L. (2010). *Aphids — Aphidinae (Macrosiphini)*. Handbooks for the Identification of British Insects, Vol. 2, Part 7. Royal Entomological Society, St Albans.

BOLGER, A.M., LOHSE, M. & USADEL, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30: 2114–2120. doi: https://doi.org/10.1093/bioinformatics/btu170

BUCKLER IV, E.S. & THORNSBERRY, J.M. (2002). Plant molecular diversity and applications to genomics. *Current Opinion in Plant Biology*,

5: 107–111. doi: https:/doi.org/10.1016/s1369-5266(02)00238-8

CARROLL, G.C. (1988). Fungal endophytes in stems and leaves from latent pathogen to mutual symbiont. *Ecology*, 69: 2–9. doi: https://doi.org/10.2307/1943154

CD GENOMICS (2025). CD Genomics Blog: Error rate of PacBio vs Nanopore: How accurate are long-read sequencing technologies. Available online: www.cd-genomics.com/blog/pacbio-nanopore-error-rate-correction-strategies (accessed September 2025).

CHALLIS, R., RICHARDS, E., RAJAN, J., COCHRANE, G. & BLAXTER, M. (2020). BlobToolKit–interactive quality assessment of genome assemblies. *G3: Genes, Genomes, Genetics*, 10: 1361–1374. doi: https://doi.org/10.1534/g3.119.400908

DE COSTER, W., D'HERT, S., SCHULTZ, D.T., CRUTS, M. & VAN BROECKHOVEN, C. (2018). NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, 34: 2666–2669. doi: https://doi.org/10.1093/bioinformatics/bty149

DE COSTER, W. & RADEMAKERS, R. (2023). NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics*, 39: btad311. doi: https://doi.org/10.1093/bioinformatics/btad311

DENCHEV, C.M. & DENCHEV, T.T. (2021). Validation of the generic names *Meira* and *Acaromyces* and nineteen species names of basidiomycetous yeasts. *Mycobiota*, 11: 1–10. doi: https://doi.org/10.12664/mycobiota.2021.11.01

DOHM, J.C., PETERS, P., STRALIS-PAVESE, N. & HIMMELBAUER, H. (2020). Benchmarking of long-read correction methods. *NAR Genomics and Bioinformatics*, 2: lqaa037. doi: https://doi.org/10.1093/nargab/lqaa037

FRANIĆ, I., ALLAN, E., PROSPERO, S., ADAMSON, K., ATTORRE, F., AUGER-ROZENBERG, M.-A., AUGUSTIN, S., AVTZIS, D., BAERT, W., BARTA, M., BAUTERS, K., BELLAHIRECH, A. *ET AL.* (2023). Climate, host and geography shape insect and fungal communities of trees. *Scientific Reports*, 13: 11570. doi: https://doi.org/10.1038/s41598-023-36795-w

GALANTI, D., JUNG, J.H., MÜLLER, C. & BOSSDORF, O. (2024). Discarded sequencing reads uncover natural variation in pest resistance in *Thlaspi arvense*. *eLife*, 13: RP95510. doi: https://doi.org/10.7554/eLife.95510.3

GATHERCOLE, L.A.P., NOCCHI, G., BROWN, N., COKER, T.L.R., PLUMB, W.J., STOCKS, J.J., NICHOLS, R.A., DENMAN, S. & BUGGS, R.J.A. (2021). Evidence for the widespread occurrence of bacteria implicated in Acute Oak Decline from incidental genetic sampling. *Forests*, 12: 1683. doi: https://doi.org/10.3390/f12121683

GEYER, J.K., GRUNBERG, R.L., WANG, J. & MITCHELL, C.E. (2024). Leaf age structures phyllosphere microbial communities in the field and greenhouse. *Frontiers in Microbiology*, 15. doi: https://doi.org/10.3389/fmicb.2024.1429166

GUREVICH, A., SAVELIEV, V., VYAHHI, N. & TESLER, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29: 1072–1075. doi: https://doi.org/10.1093/bioinformatics/btt086

HERNÁNDEZ-TASCO, A.J., TRONCHINI, R.A., APAZA-CASTILLO, G.A., HOSAKA, G.K., QUIÑONES, N.R., GOULART, M.C., FANTINATTI-GARBOGGINI, F. & SALVADOR, M.J. (2023). Diversity of bacterial and fungal endophytic communities presents in the leaf blades of *Sinningia magnifica*, *Sinningia schiffneri* and *Sinningia speciosa* from different cladus of Gesneriaceae family: A comparative analysis in three consecutive years. *Microbiological Research*, 271: 127365. doi: https://doi.org/10.1016/j.micres.2023.127365

HU, T., CHITNIS, N., MONOS, D. & DINH, A. (2021). Next-generation sequencing technologies: An overview. *Human Immunology*, 82: 801–811. doi: https://doi.org/10.1016/j.humimm.2021.02.012

JAIN, M., FIDDES, I.T., MIGA, K.H., OLSEN, H.E., PATEN, B. & AKESON, M. (2015). Improved data analysis for the MinION nanopore sequencer. *Nature Methods*, 12: 351–356. doi: https://doi.org/10.1038/nmeth.3290

JO, Y., BACK, C.G., KIM, K.H., CHU, H., LEE, J.H., MOH, S.H. & CHO, W.K. (2021). Comparative study of metagenomics and metatranscriptomics to reveal microbiomes in overwintering pepper fruits. *International Journal of Molecular Sciences*, 22: 6202. doi: https://doi.org/10.3390/ijms22126202

KANWAR, N., BLANCO, C., CHEN, I.A. & SEELIG, B. (2021). PacBio sequencing output increased through uniform and directional fivefold

concatenation. *Scientific Reports*, 11: 18065. doi: https://doi.org/10.1038/s41598-021-96829-z

KENNEDY, C. & SOUTHWOOD, T. (1984). The number of species of insects associated with British trees: a re-analysis. *The Journal of Animal Ecology*, 53: 455–478. doi: https://doi.org/10.2307/4528

LAETSCH, D.R. & BLAXTER, M.L. (2017). BlobTools: Interrogation of genome assemblies. *F1000Research*, 6: 1287. doi: https://doi.org/10.12688/f1000research.12232.1

LI, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34: 3094–3100. doi: https:/doi.org/10.1093/bioinformatics/bty191

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. & 1000 GENOME PROJECT DATA PROCESSING SUBGROUP (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25: 2078–2079. doi: https://doi.org/10.1093/bioinformatics/btp352

LI, Y. (2024). What lives within and on a plant: our understanding from genome NGS data. Unpublished MSc thesis, University of Edinburgh.

LIN, X., TANG, W., AHMAD, S., LU, J., COLBY, C.C., ZHU, J. & YU, Q. (2012). Applications of targeted gene capture and next-generation sequencing technologies in studies of human deafness and other genetic disabilities. *Hearing Research*, 288: 67–76. doi: https://doi.org/10.1016/j.heares.2012.01.004

LIU, B.W., LI, S.Y., ZHU, H. & LIU, G.X. (2023). Phyllosphere eukaryotic microalgal communities in rainforests: Drivers and diversity. *Plant Diversity*, 45: 45–53. doi https://doi.org/10.1016/j.pld.2022.08.006

LU, J. & SALZBERG, S.L. (2018). Removing contaminants from databases of draft genomes. *PLoS Computational Biology*, 14: e1006277. doi: https://doi.org/10.1371/journal.pcbi.1006277

LUO, R., LIU, B., XIE, Y., LI, Z., HUANG, W., YUAN, J., HE, G., CHEN, Y., PAN, Q. & LIU, Y. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, 1: 2047-2217X-2041-2018. doi: https://doi.org/10.1186/2047-217X-1-18

LUPO, V., VAN VLIERBERGHE, M., VANDERSCHUREN, H., KERFF, F., BAURAIN, D. & CORNET, L. (2021). Contamination in reference sequence databases: time for divide-and-rule tactics. *Frontiers in Microbiology*, 12: 755101. doi: https://doi.org/10.3389/fmicb.2021.755101

MARSBERG, A., KEMLER, M., JAMI, F., NAGEL, J.H., POSTMA-SMIDT, A., NAIDOO, S., WINGFIELD, M.J., CROUS, P.W., SPATAFORA, J.W. & HESSE, C.N. (2017). *Botryosphaeria dothidea*: a latent pathogen of global importance to woody plant health. *Molecular Plant Pathology*, 18: 477–488. doi: https://doi.org/10.1111/mpp.12495

MEHL, J., WINGFIELD, M.J., ROUX, J. & SLIPPERS, B. (2017). Invasive everywhere? Phylogeographic analysis of the globally distributed tree pathogen *Lasiodiplodia theobromae*. *Forests*, 8: 145. doi: https://doi.org/10.3390/f8050145

MICCOLI, C., PALMIERI, D., DE CURTIS, F., LIMA, G., HEITMAN, J., CASTORIA, R. & IANIRI, G. (2020). The necessity for molecular classification of basidiomycetous biocontrol yeasts. *BioControl*, 65: 489–500. doi: https://doi.org/10.1007/s10526-020-10008-z

MIN, S.H. & ZHOU, J. (2021). smplot: an R package for easy and elegant data visualization. *Frontiers in Genetics*, 12: 802894. doi: https://doi.org/10.3389/fgene.2021.802894

MÖLLER, M. & CRONK, Q.C.B. (2001). Phylogenetic studies in *Streptocarpus* (Gesneriaceae): reconstruction of biogeographic history and distribution patterns. *Systematics and Geography of Plants*, 71(2): 545–555. doi: https://doi.org/10.2307/3668699

NISHII, K., HART, M., KELSO, N., BARBER, S., CHEN, Y.Y., THOMSON, M., TRIVEDI, U., TWYFORD, A.D. & MÖLLER, M. (2022). The first genome for the Cape Primrose *Streptocarpus rexii* (Gesneriaceae), a model plant for studying meristem-driven shoot diversity. *Plant Direct*, 6: e388. doi: https://doi.org/10.1002/pld3.388

PALMIERI, D., BARONE, G., CIGLIANO, R.A., DE CURTIS, F., LIMA, G., CASTORIA, R. & IANIRI, G. (2021). Complete genome sequence of the biocontrol yeast *Papiliotrema terrestris* strain LS28. *G3: Genes, Genomes, Genetics*, 11: jkab332. doi: https://doi.org/10.1093/g3journal/jkab332

PAPPALARDO, P., HEMMI, J.M., MACHIDA, R.J., LERAY, M., COLLINS, A.G. & OSBORN, K.J. (2025). Taxon-specific BLAST percent identity thresholds for identification of unknown sequences using metabarcoding. *Methods in Ecology*

and Evolution, 16: 2380–2394. doi: https://doi.org/10.1111/2041-210X.70147

PATON, J. (2023). What's in an NOR? Nuclear ribosomal DNA variation in *Aeschynanthus* (Gesneriaceae) – what can be revealed with genome skimming? Unpublished MSc thesis, University of Edinburgh.

PÉREZ-COBAS, A.E., GOMEZ-VALERO, L. & BUCHRIESER, C. (2020). Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses. *Microbial Genomics*, 6: e000409. doi: https://doi.org/10.1099/mgen.0.000409

PIĄTEK, M., LUTZ, M. & WELTON, P. (2012). *Exobasidium darwinii*, a new Hawaiian species infecting endemic *Vaccinium reticulatum* in Haleakala National Park. *Mycological Progress*, 11: 361–371. doi: https://doi.org/10.1007/s11557-011-0751-4

PIOMBO, E., ABDELFATTAH, A., DROBY, S., WISNIEWSKI, M., SPADARO, D. & SCHENA, L. (2021). Metagenomics approaches for the detection and surveillance of emerging and recurrent plant pathogens. *Microorganisms*, 9: 188. doi: https://doi.org/10.3390/microorganisms9010188

PUCKER, B., IRISARRI, I., DE VRIES, J. & XU, B. (2022). Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions. *Quantitative Plant Biology*, 3: e5. doi: https://doi.org/10.1017/qpb.2021.18

QIN, A., DING, Y., JIAN, Z., MA, F., WORTH, J.R., PEI, S., XU, G., GUO, Q. & SHI, Z. (2021). Low genetic diversity and population differentiation in *Thuja sutchuenensis* Franch., an extremely endangered rediscovered conifer species in southwestern China. *Global Ecology and Conservation*, 25: e01430. doi: https://doi.org/10.1016/j.gecco.2020.e01430

RAMAKRISHNAN, D.K., JAUERNEGGER, F., HOEFLE, D., BERG, C., BERG, G. & ABDELFATTAH, A. (2024). Unravelling the microbiome of wild flowering plants: a comparative study of leaves and flowers in alpine ecosystems. *BMC Microbiology*, 24: 417. doi: https://doi.org/10.1186/s12866-024-03574-0

R CORE TEAM (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Available online: www.R-project.org (accessed August 2025).

ROMAN-REYNA, V., PINILI, D., BORJA, F.N., QUIBOD, I.L., GROEN, S.C., ALEXANDROV, N., MAULEON, R. & OLIVA, R. (2020). Characterization of the leaf microbiome from whole-genome sequencing data of the 3000 Rice Genomes Project. *Rice*, 13: 72. doi: https://doi.org/10.1186/s12284-020-00432-1

RRWICK (2018). Adapter trimmer for Oxford Nanopore reads. Github. Available online: https://github.com/rrwick/Porechop (accessed June 2024).

RUAN, J. & LI, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nature Methods*, 17: 155–158. doi: https://doi.org/10.1038/s41592-019-0669-3

RUNGJINDAMAI, N. & JONES, E.B.G. (2024). Why are there so few Basidiomycota and basal fungi as endophytes? A review. *Journal of Fungi*, 10: 67. doi: https://doi.org/10.3390/jof10010067

SAIKKONEN, K., FAETH, S., HELANDER, M. & SULLIVAN, T. (1998). Fungal endophytes: a continuum of interactions with host plants. *Annual Review of Ecology and Systematics*, 29: 319–343. doi: https://doi.org/10.1146/annurev.ecolsys.29.1.319

SAILLARD, C., VIGNAULT, J., BOVÉ, J., RAIE, A., TULLY, J., WILLIAMSON, D., FOS, A., GARNIER, M., GADEAU, A. & CARLE, P. (1987). *Spiroplasma phoeniceum* sp. nov., a new plant-pathogenic species from Syria. *International Journal of Systematic and Evolutionary Microbiology*, 37: 106–115. doi: https://doi.org/10.1099/00207713-37-2-106

SAINI, M.K., GAURAV, H., KUMAR, J. & SANU, K. (2023). DNA sequencing techniques: Sanger to next generation sequencing. *The Science World*, 3: 2378–2393. doi: https://doi.org/10.5281/zenodo.8376905

SANGIOVANNI, M., GRANATA, I., THIND, A.S. & GUARRACINO, M.R. (2019). From trash to treasure: detecting unexpected contamination in unmapped NGS data. *BMC Bioinformatics*, 20: 168. doi: https://doi.org/10.1186/s12859-019-2684-x

SCHÖNROGGE, K., GIBBS, M., OLIVER, A., CAVERS, S., GWEON, H.S., ENNOS, R.A., COTTRELL, J., IASON, G.R. & TAYLOR, J. (2022). Environmental factors and host genetic variation

shape the fungal endophyte communities within needles of Scots pine (*Pinus sylvestris*). *Fungal Ecology*, 57–58: 101162. doi: https://doi.org/10.1016/j.funeco.2022.101162

SHEN, W., LE, S., LI, Y. & HU, F. (2016). SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PloS One*, 11: e0163962. doi: https://doi.org/10.1371/journal.pone.0163962

SOHRABI, R., PAASCH, B.C., LIBER, J.A. & HE, S.Y. (2023). Phyllosphere microbiome. *Annual Review of Plant Biology*, 74: 539–568. doi: https://doi.org/10.1146/annurev-arplant-102820-032704

THITLA, T., KUMLA, J., KHUNA, S., LUMYONG, S. & SUWANNARACH, N. (2022). Species diversity, distribution, and phylogeny of *Exophiala* with the addition of four new species from Thailand. *Journal of Fungi*, 8: 766. doi: https://doi.org/10.3390/jof8080766

THOMAS, G., KAY, W.T. & FONES, H.N. (2024). Life on a leaf: the epiphyte to pathogen continuum and interplay in the phyllosphere. *BMC Biology*, 22: 168. doi: https://doi.org/10.1186/s12915-024-01967-1

VALENCIA, C.A., PERVAIZ, M.A., HUSAMI, A., QIAN, Y. & ZHANG, K. (2013). *Next Generation Sequencing Technologies in Medical Genetics*. Springer: SpringerBriefs in Genetics, New York.

VASSE, M., VOGLMAYR, H., MAYER, V., GUEIDAN, C., NEPEL, M., MORENO, L., DE HOOG, S., SELOSSE, M.-A., MCKEY, D. & BLATRIX, R. (2017). A phylogenetic perspective on the association between ants (Hymenoptera: Formicidae) and black yeasts (Ascomycota: Chaetothyriales). *Proceedings of the Royal Society B: Biological Sciences*, 284: 20162519. doi: https://doi.org/10.1098/rspb.2016.2519

VENBRUX, M., CRAUWELS, S. & REDIERS, H. (2023). Current and emerging trends in techniques for plant pathogen detection. *Frontiers in Plant Science*, 14: 1120968. doi: https://doi.org/10.3389/fpls.2023.1120968

VORHOLT, J.A. (2012). Microbial life in the phyllosphere. *Nature Reviews Microbiology*, 10: 828–840. doi: https://doi.org/10.3929/ethz-b-000059727

WALLER, J.M., LENNÉ, J.M. & WALLER, S.J. (2001). *Plant Pathologist's Pocketbook*. CABI Publishing, Wallingford.

WICKHAM, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer, Cham.

YANG, F., PU, X., MATTHEW, C., NAN, Z. & LI, X. (2024). Exploring phyllosphere fungal communities of 29 alpine meadow plant species: composition, structure, function, and implications for plant fungal diseases. *Frontiers in Microbiology*, 15: 1451531. doi: https://doi.org/10.3389/fmicb.2024.1451531

ZEINELDIN, M., HICKS, J., WARD, H.J., WÜNSCHMANN, A., CAMP, P., FARRELL, D., LEHMAN, K., THACKER, T.C. & CUTHBERT, E. (2023). Complete genome sequence of *Candidatus Mycobacterium wuenschmannii*, a nontuberculous mycobacterium isolated from a captive population of Amazon milk frogs. *Microbiology Resource Announcements*, 12: e00547-00523. doi: https://doi.org/10.1128/MRA.00547-23