A Minimal bookdown document

Alex Ringeri

Contents

1 Introduction

2 Objectives and Research Question

3 Methods

3.1 Preparation of reads from fungal isolates for synthetic dataset

A mock dataset has been created to simulate scenarios where multiple fungal species are present in the same sample. Using reads from the sequencing of

65 known fungal isolates, we were able to control the proportions of species in each simulated scenario. These synthetically generated datasets aim provide insight into the limitations and sensitivity of this pipeline. Supplemental Table 1 provides the list of fungal species or strains and the number of raw reads available for the mock dataset. The steps to generate and prepare reads for this synthetic dataset are outlined below.

3.1.1 DNA extraction, amplification and sequencing

The genomic DNA of 65 fungal strains was extracted with the Qiagen DNeasy Plant mini kit. The DNA was extracted from mycelia or spore material of each fungal isolate. Partial SSU, full ITS region and partial LSU regions of each sample were amplified with primers NS5 (forward) and LR6 (reverse). The amplicons were prepared with the Oxford Nanopore's Native Barcoding Kit 96 V14 (code SQK-NBD114.96, ONT). Fungal samples plus a negative control were multiplexed and sequenced using the MinION R10.4 flow cell.

3.1.2 Bioinformatics - Basecalling, trimming and filtering

The raw ONT data was basecalled and demultiplexed using Guppy v6.4.2 with the super-high accuracy model (dna_r10.4.1_e8.2_400bps_sup). The mean read length of amplicons after basecalling was ~2.2 KB as seen in Figure 4.

- [alt] Dorado: dna_r10.4.1_e8.2_400bps_sup@v4.1.0

Sequencing adapters were trimmed from raw basecalled reads using Dorado v0.6.1 with the '`--no-trim-primers`' option to avoid removing primer sequences. Cutadapt v4.6 (Martin, 2011) has been used to select and trim primers from amplicons that contain both forward (NS5) and reverse primer (LR6) sequences. Amplicons where both of these primer sequences could not be detected were excluded from the analysis. Cutadapt was also used to re-orient reads that have been sequenced by the reverse strand ('`--revcomp`' option), making it easier to process by downstream tools (Figure 1)

The full ITS region of these reads was extracted using ITSxpress v2.0.1 (Rivers *et al.*, 2018) with default settings other than '`--single_end`' and '`--taxa Fungi`' options. Chopper v0.7.0 (De Coster and Rademakers, 2023) was used to select reads of the full ITS region having a length between 300-6000bp and mean Phred quality score above Q20. Chimeric reads were detected using VSEARCH's (Rognes *et al.*, 2016) denovo and reference based methods. The database used for reference based chimera detection of full ITS sequence was the UNITE general release v9.0 (sh_general_release_dynamic_s_all_25.07.2023).

3.1.3 Sample selection and taxonomic naming validation

To ensure that taxonomic names were used consistently between the fungal isolate samples and the reference database, manual validation of each sample name was performed. A search was conducted for the recorded species name of

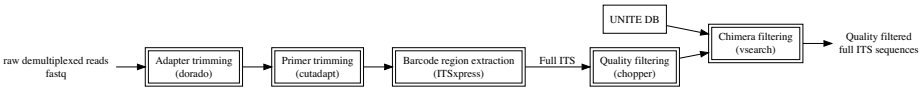


Figure 1: The steps taken to prepare raw reads for mock scenarios. Boxes inticate trimming and filtering stages of the pipeline with and the software used is in paretheses.

each sample in the 2024 UNITE reference database (Vu, 2024). Most species names that could not be found were either misspelled or used an older taxonomic synonym. For example, *Nakaseomyces glabratus* has been recorded by its former name *Candida glabrata* which the latter could not be found in the reference database. Names of samples were manually updated using current names found in both Index Fungorum (www.indexfungorum.org) and the 2024 UNITE reference database. For samples that were not classified at the species level (e.g. *Entoleuca sp* CCL052), the genus label was confirmed to exist in the reference database.

Seven samples that had less than a total of 2500 reads after basecalling, trimming and filtering were removed from the analysis leaving 58 samples (a total of 55 species). The excluded taxa included: *Aspergillus niger*, *Naganishia albida* (*Cryptococcus albidus*), *Geotrichum candidum* (*Galactomyces geotrichum*), *Meyerozyma guilliermondii*, Yarrowia lipolytica, *Fusarium proliferatum* and *Puccinia recondita* (*Puccinia triticina*). Bias in chimera detection leading to a loss 99.0% and 68.8% of reads in *Puccinia triticina* and *Fusarium proliferatum* respectively.

3.1.4 Synthetic dataset - Scenario 1 - Equal abundance

Four scenarios have been designed to test the pipeline under varying community structures. Each mock scenario consists of full ITS sequences that pass the quality filtering steps and are combined in different proportions to control the abundance of each sample.

Scenario 1 considers an even community structure where every fungal isolate has equal abundance (in terms of number of reads). Libraries of differing sizes were generated by subsampling the same number of reads from each sample. For each of the 58 samples: 20, 50, 167, 1000, 2000 and 2500 reads were randomly selected to produce libraries with sizes of 1160, 2900, 9686, 58000, 116000 and 145000 reads. Seqtk v1.4 (Lh3/seqtk: Toolkit for processing sequences in FASTA/Q formats) was used to perform the subsampling (with the '`sample`' command). The subsampling was repeated five times for each library size producing a total of 30 libraries (6 libraries $\times$ 5 repetitions). The seed values given to seqtk's 'sample' command for each repetition were generated deterministically so that each run of the pipeline is reproducible.

3.1.5 Synthetic dataset - Scenario 2 - Uneven (5 low abundance)

- 5 samples with very low, rest high

3.1.6 Synthetic dataset - Scenario 3 - Uneven (5 high abundance)

- all very low, 5 high

3.1.7 Synthetic dataset - Scenario 4 - Uneven (range of abundance)

- sampling in stepwise abundance increments

3.2 Bioinformatics - Clustering

For each scenario full ITS sequences were grouped into approximate species-level clusters. Two de-novo clustering approaches were explored and can be seen in Figure 2.

The first approach uses centroid-based clustering on sequence similarity as implemented in VSEARCH (Rognes *et al.*, 2016). The sequences were first de-replicated using VSEARCH's '–fastx_uniques' option to merge identical sequences into a single record, then clustered with the '–cluster_size' option. A 97% identity was used as the pairwise sequence similarity threshold. This clustering approach resulted in many low-abundance operational taxonomic units (OTUs), largely overestimating the number of species when unfiltered. A minimum OTU size threshold of 0.1% of the library size was used to filter out low-abundance OTUs and more closely estimate the number of species (see Section 4.2 for the determination of this threshold).

The second approach clusters similar sequences through transformations of k-mer signatures as used in the NanoCLUST pipeline (Rodríguez-Pérez *et al.*, 2021; Langsiri *et al.*, 2023). Each sequence was transformed into a k-mer frequency vector and stored in a tabular format. In our pipeline, 6-mer frequencies were computed (as opposed to 5-mers in NanoCLUST). The multidimensional tabular structure was then projected into two-dimensions using Uniform Manifold Approximation and Projection (UMAP) (McInnes *et al.*, 2020). Sequences (represented as points in the two-dimensional space) were then clustered using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (McInnes *et al.*, 2017). The 'cluster_selection_epsilon' parameter used in HDBSCAN was set to 0.5 as used in (Langsiri *et al.*, 2023), while the 'min_cluster_size' parameter was set to 0.5% of the library size. See Section 4.2 for determination of the 'min_cluster_size' parameter.

3.3 Bioinformatics - Taxonomic assignments

A representative sequence from each cluster was selected to give taxonomic assignments to the cluster. Our initial approach used the most abundant sequence of each cluster as the representative sequence. The VSEARCH clustering

Figure 2: Clustering stages of the pipeline with the tools used in paretheses.

approach provides these sequences with the '`--centroids`' option. To determine the most abundant sequence from clusters in the UMAP + HDBSCAN approach, the VSEARCH '`--fastx_uniques`' command in combination with '`--topn 1`' option was used.

The representative sequence from each cluster was then given a taxonomic assignment with dnabarcoder v1.0.6 (Vu *et al.*, 2022). In this case full ITS sequences were classified against the UNITE 2024 reference database (Vu, 2024) using precomputed similarity cutoffs provided by the dnabarcoder project (unite2024ITS.unique.cutoffs.best.json).

- TODO: assign based on polished sequence of each cluster (NanoCLUST method)
  - fastANI - compute draft sequence from each cluster (based on highest Average Nucleotide Identity)
  - minimap2 - map reads to draft
  - racon - polishing
  - medaka - polishing step
  - unable to do read correction with canu due to full ITS region being too short. Canu requires a '`genomeSize`' of at least 1000, where mean full ITS from our samples were ~400.

4 Results

4.1 Quality control

Tracking read counts at each stage of the pipeline has shown that large losses of reads occurred in the primer trimming and quality filtering stages of the pipeline (Figure 3). 32.17% of reads were lost after applying cutadapt to select and trim amplicons that contain both forward (ITS1F) and reverse primer (LR3) sequences. The extraction of full ITS regions led to a loss of 11.83% trimmed reads. 71.35% of full ITS sequences were lost after selecting reads between 300-6000bp in length and having a mean quality above Q20 (Phred scale). Such large loss of reads can be attributed to many of the reads from the dataset being below the minimum mean quality threshold of Q20 as in Figure 4.

It should be noted that no reads were removed after trimming ONT adapter sequences at the ends of reads with Dorado. While the overall read loss for the chimera filtering step was relatively low (0.75%), we observed sample bias in reference-based chimera detection step for samples which were excluded

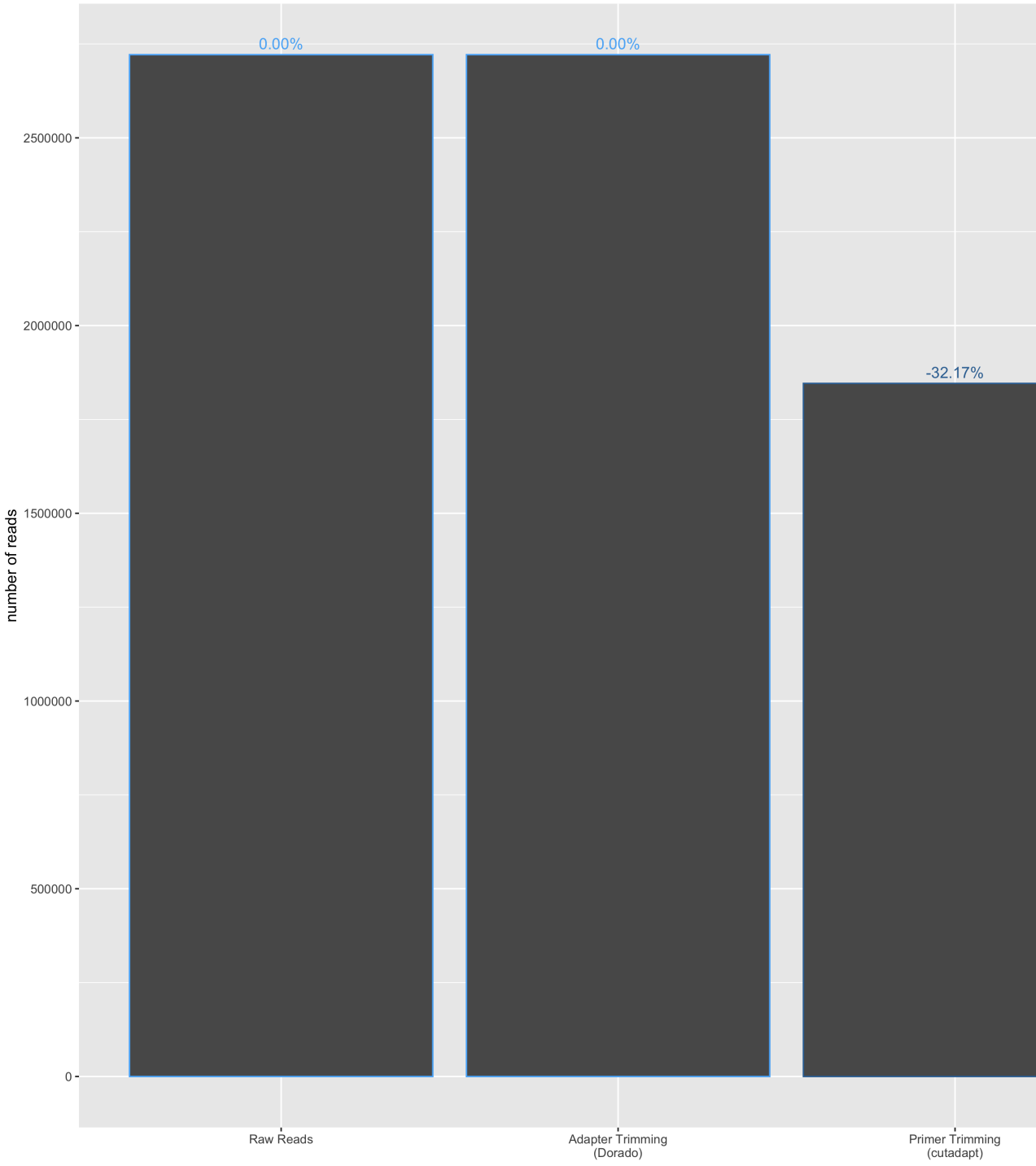


Figure 3: Read loss at each filtering and trimming stage of the pipeline. The percentages on each bar indicate the proportion of reads lost from the previous step. Each step is executed in order from left to right.

from the analysis (99.0% loss in Puccinia triticina and 68.8% loss in Fusarium proliferatum).

4.2 Minimum cluster size threshold can recover expected number of species from even abundance scenario

The clustering approach that was adopted aimed to group full ITS sequences from the same species together. The total number of clusters was used as a measure for successfully estimating number of species in the community. In this case we are considering the even abundance mock community where all samples are equally represented (Scenario 1).

Clustering using VSEARCH at 97% sequence identity consistently over-estimated the number of species in mixed read scenarios. Large numbers of OTUs were encountered with many of them having low abundance. The method of removing OTUs with abundance levels below a specified threshold was explored in Figure 5. The minimum OTU size threshold was selected as a proportion of the total library size. When clustering with VSEARCH at 97% sequence identity, the threshold that gave consistent recovery of the number of species was 0.15% of total library size. This threshold was robust across different library sizes (1000-150000 total reads) generated by subsampling the full dataset.

A k-mer based clustering method (used by NanoCLUST) was explored to test whether we could improve the accuracy of species level clustering by potentially being more tolerant of noise in the ONT reads compared to VSEARCH. The minimum cluster size parameter used by the HDBSCAN greatly affects the number of clusters by determining the smallest size grouping that is considered a cluster. We tested multiple library sizes to determine an optimal value for this minimum cluster size parameter. When the minimum cluster size parameter was set to its minimum value (2), the NanoCLUST method overestimated the number of species for library sizes 10000 and above. The number of clusters increased significantly with library size for small values of the minimum cluster parameter (Figure 6). For library sizes of 10000 reads and above, a minimum cluster size of 0.65% (of the total library size) approximately recovered the expected number of species.

A direct comparison between the clustering methods was performed by comparing the read loss when specifying a minimum clustering threshold. Read loss in this instance refers to the proportion of reads that were removed from the analysis due to being placed in a cluster that was smaller than the given threshold. We found that for low library sizes (20-167 reads per sample), the NanoCLUST method produced closer estimates of the actual number of species with lower read loss than VSEARCH for all threshold values (Figure 7). For larger library sizes (1000-2500 reads per sample), the NanoCLUST method's minimum cluster size of 0.65% produced consistently accurate estimates of the actual number of species while VSEARCH's minimum cluster threshold of 0.15% performed similarly. When considering read loss at these thresholds, the NanoCLUST

Figure 4: The mean read quality of unprocessed reads before trimming or filtering steps. Quality scores are in the Phred (Q) scale. Read lengths are shown in thousand basepairs. Histograms on X and Y axes indicate the density of reads at respective quality scores and read lengths. (Plot generated by NanoPlot [@DeCoster2023])

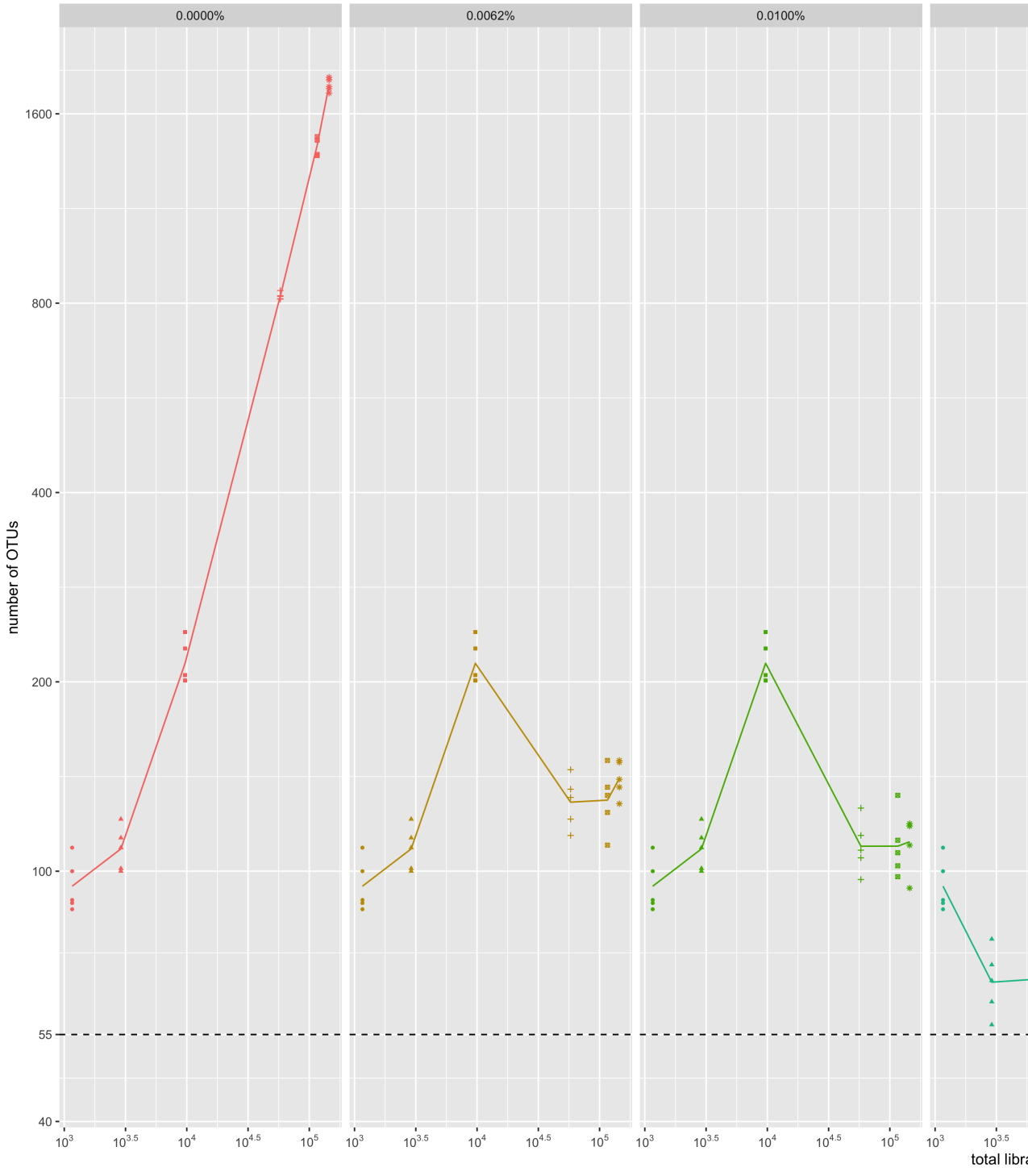


Figure 5: The effect of minimum OTU size threshold and total library size on the number of OTUs when clustering with VSEARCH. Colours indicate the minimum OTU size threshold as a proportion of the total library size. The dashed line indicates the actual number of species in the synthetic dataset. Five repetitions were performed for each library size. X and Y axes are on a logarithmic scale.

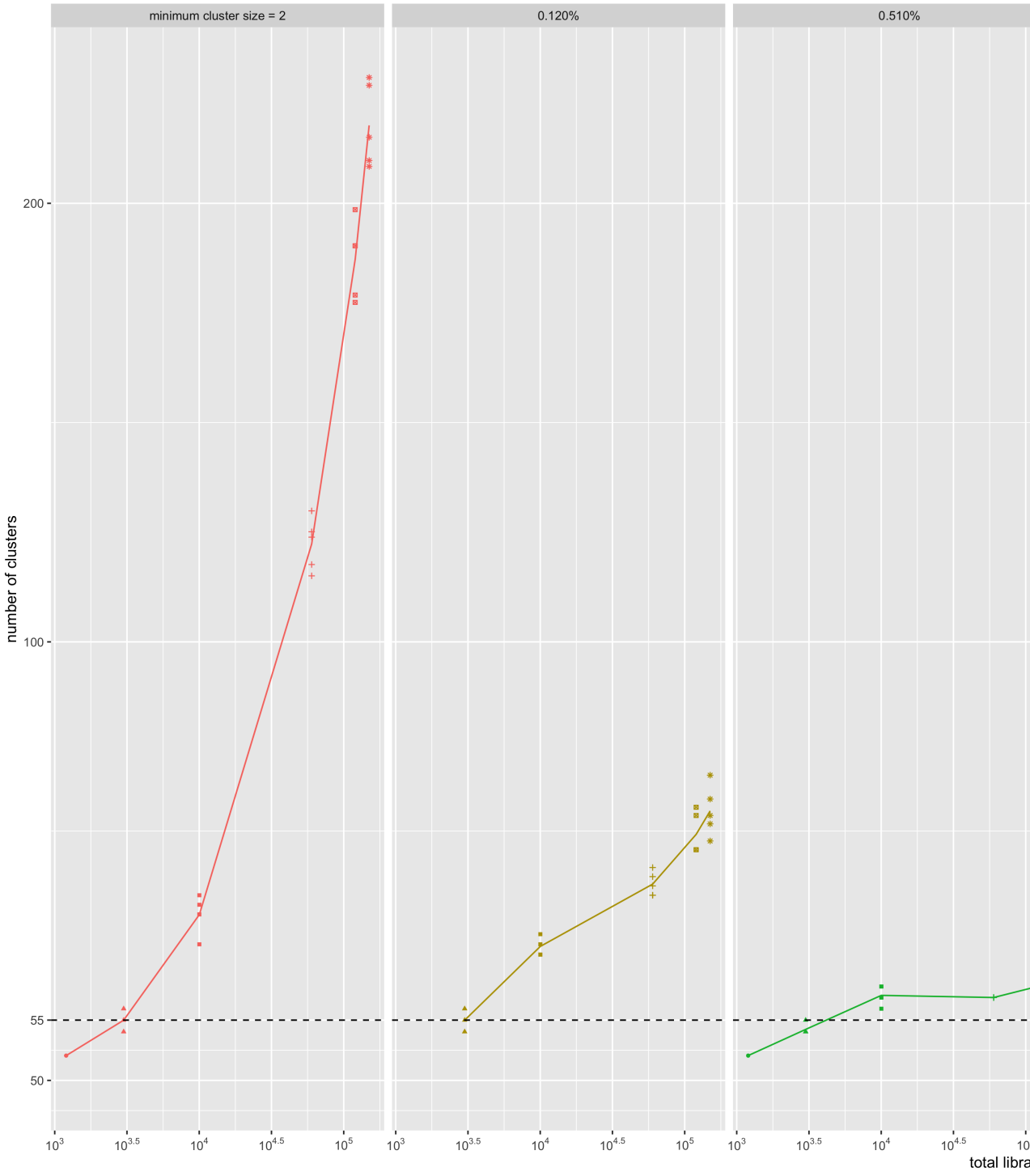


Figure 6: The effect of minimum cluster size parameter and total library size on the number of clusters when following the UMAP and HDBSCAN clustering approach. Colours indicate the minimum cluster size parameter as a proportion of the total library size. The dashed line indicates the actual number of species in the mock dataset. Five repetitions were performed for each library size. X and Y axes are on a logarithmic scale.

method performed better at 1000 reads per sample, similarly at 2000 reads per sample and worse at 2500 reads per sample compared to VSEARCH.

4.2.1 Splitting of species into clusters (Scenario 1)

To investigate the effect on clustering of closely related species we looked at a single execution of the even abundance scenario (Scenario 1), sampling at 2000 reads per sample for a total library size of 116K reads. Reads were clustered using the UMAP and HDBSCAN method (NanoCLUST) using a min cluster size of 580 (0.5% of the library size). The most abundant sequence of each cluster was assigned taxonomy with dnabarcoder using UNITE 2024 reference database.

The samples from the order *Pucciniales* are shown in Figure 8. We can see that the majority of reads from the *Puccinia striiformis (var tritici)* sample (BC 25) have clustered together into a single cluster (OTU 43), which has been given the expected species-level classification (*Puccinia striiformis*). The remaining reads from the *P. striiformis (var tritici)* sample have clustered into two groups (5 reads each) which do not correspond to the expected sample taxonomy. The taxonomic classification given to these clusters corresponds to other samples present in the library (*Zymoseptoria tritici* and the *Trichomonascus* genus) and may be indicative of index-switching, where a sequencing error has occurred in the barcode region of the read.

In the *Puccinia graminis* sample (BC 27), the majority of the reads have been split into two clusters, both of which have been classified as the expected *P. graminis* species. This likely indicates biological variation in the *P. graminis* sample that is causing the single sample to be split into separate clusters. The majority of reads for both *Austropuccinia psidii* samples (BC 28 and 36) have been split into the same two clusters (OTU 38 and 21) which have been given the classification *Puccinia psidii*. A similar plot for all samples in this scenario can be seen in Figure 12

- difficulty of clustering fungal ITS regions at species level due to variation
- classification of Austropuccinia psidii -> Puccinia psidii significant?
- limitation of using single representative sequence

Splitting of cryptococcus:

4.2.2 Effect of newer base calling model on clustering

4.2.3 Classification proportion and precision in even abundance scenario (Scenario 1)

In order to explore the accuracy of the taxonomic classifications in the pipeline, two metrics were computed for the even abundance scenario (Scenario 1). The genera classification proportion metric has been defined as $\frac{\text{\# of reads classified at the genera level}}{\text{total \# of reads in cluster}}$. The genera precision metric has been defined as $\frac{\text{\# of reads classified correctly at the genera level}}{\text{\# of reads classified at the genera level}}$. These metrics were calculated for

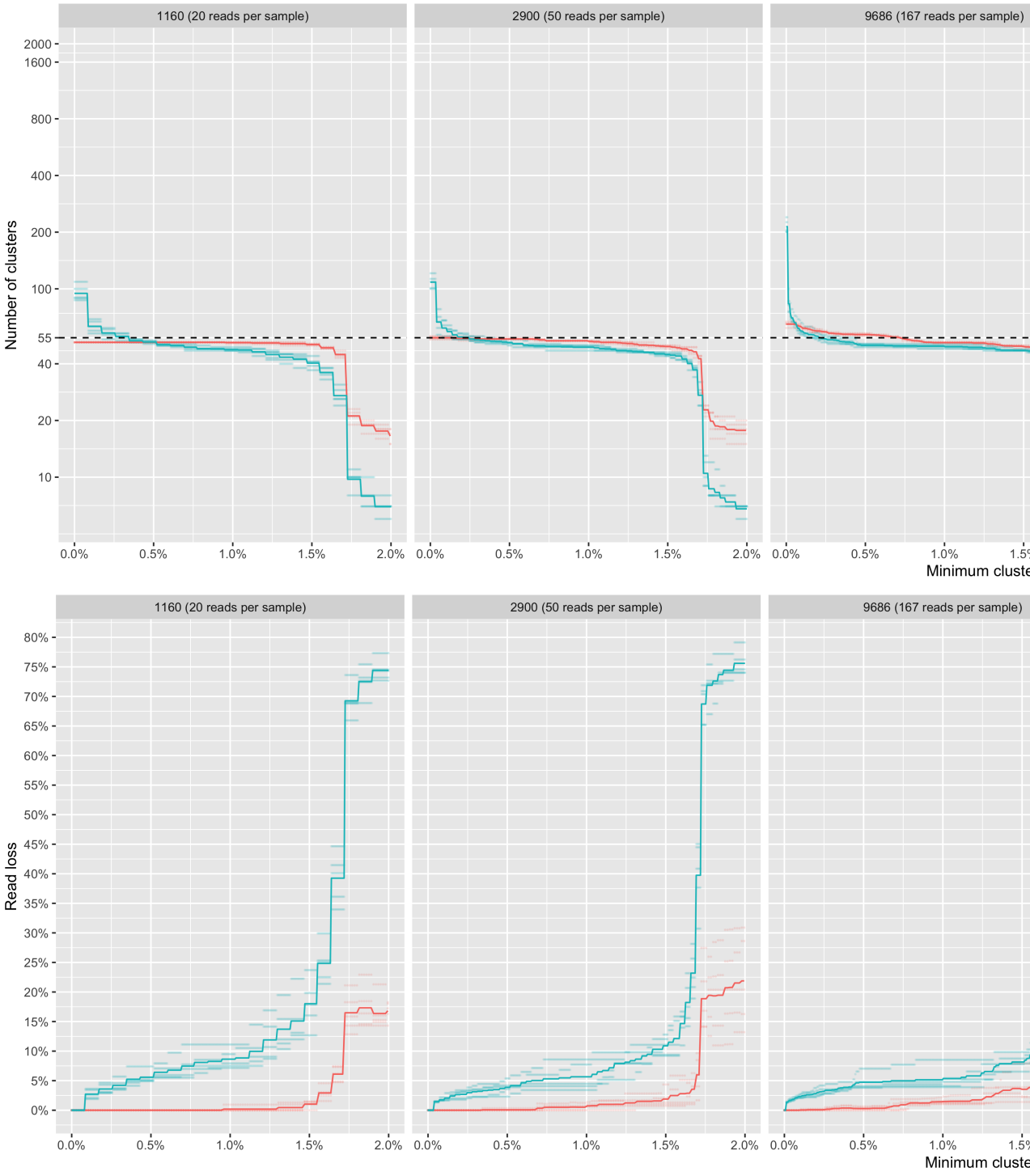


Figure 7: Comparing the impact of a minimum cluster size threshold on the number of clusters and read loss between NanoCLUST and VSEARCH clustering methods. Plots have been organised in columns by increasing library size. The top row of plots show the number of clusters after applying a minimum cluster size threshold. The bottow row of plots show the proportion of reads that are lost after applying the minimum cluster size threshold. The mean values of each have been plotted after five random resamplings at each library size.

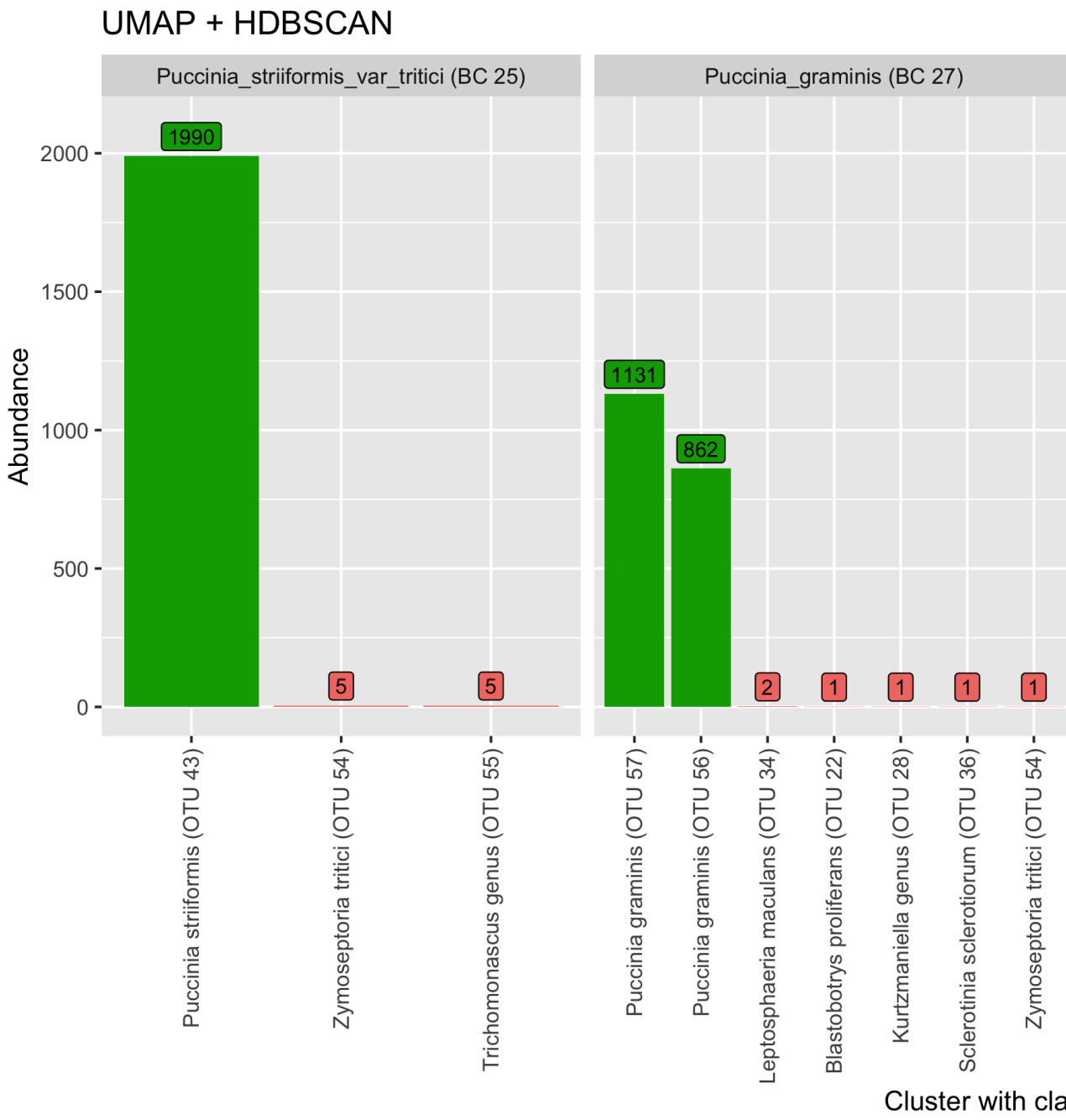


Figure 8: Plot indicating the splitting of Pucciniales samples into clusters using the UMAP + HDBSCAN method. Each plot shows how reads from each sample are distributed into clusters. Bars indicate the abundance of a cluster (number of reads). Taxonomic classification of each cluster is shown in the X axis labels. Green indicates that a cluster matches the expected species-level classification.

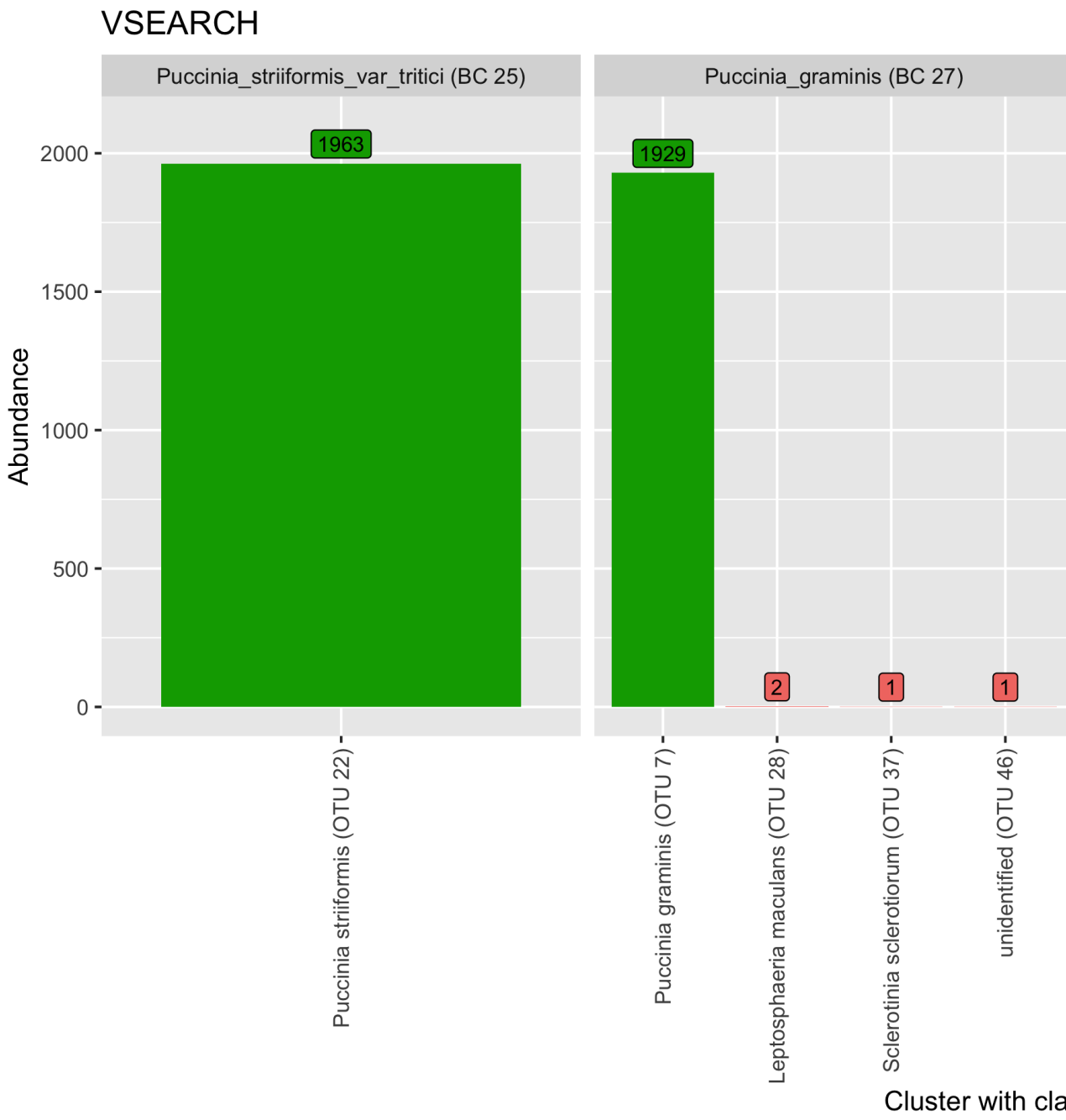


Figure 9: Plot indicating the splitting of Pucciniales samples into clusters using the VSEARCH method.

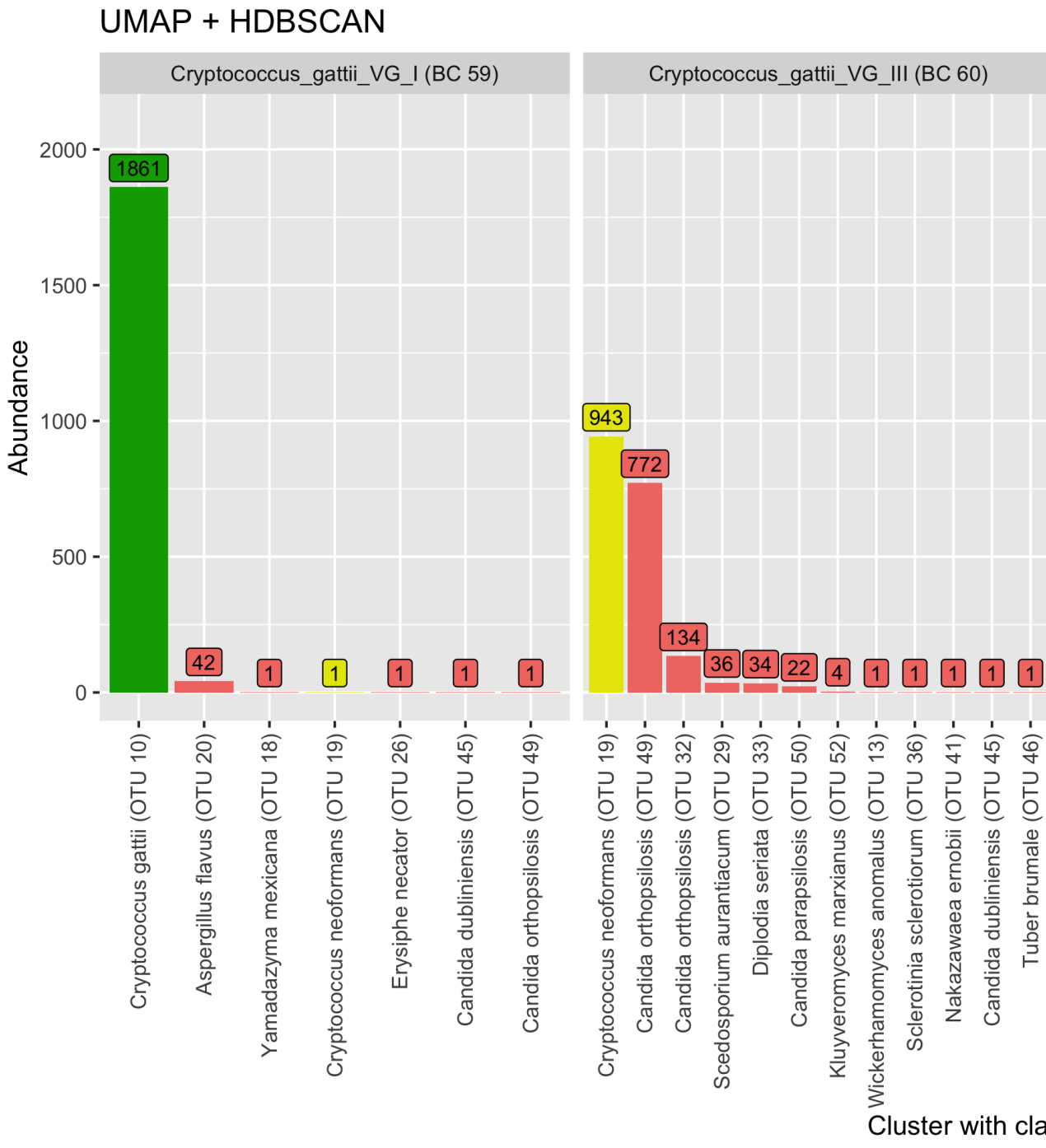


Figure 10: Plot indicating the splitting of Cryptococcus samples into clusters using the UMAP + HDBSCAN method. Each plot shows how reads from each sample are distributed into clusters. Bars indicate the abundance of a cluster (number of reads). Taxonomic classification of each cluster is shown in the X axis labels. Green indicates that a cluster matches the expected species-level classification.

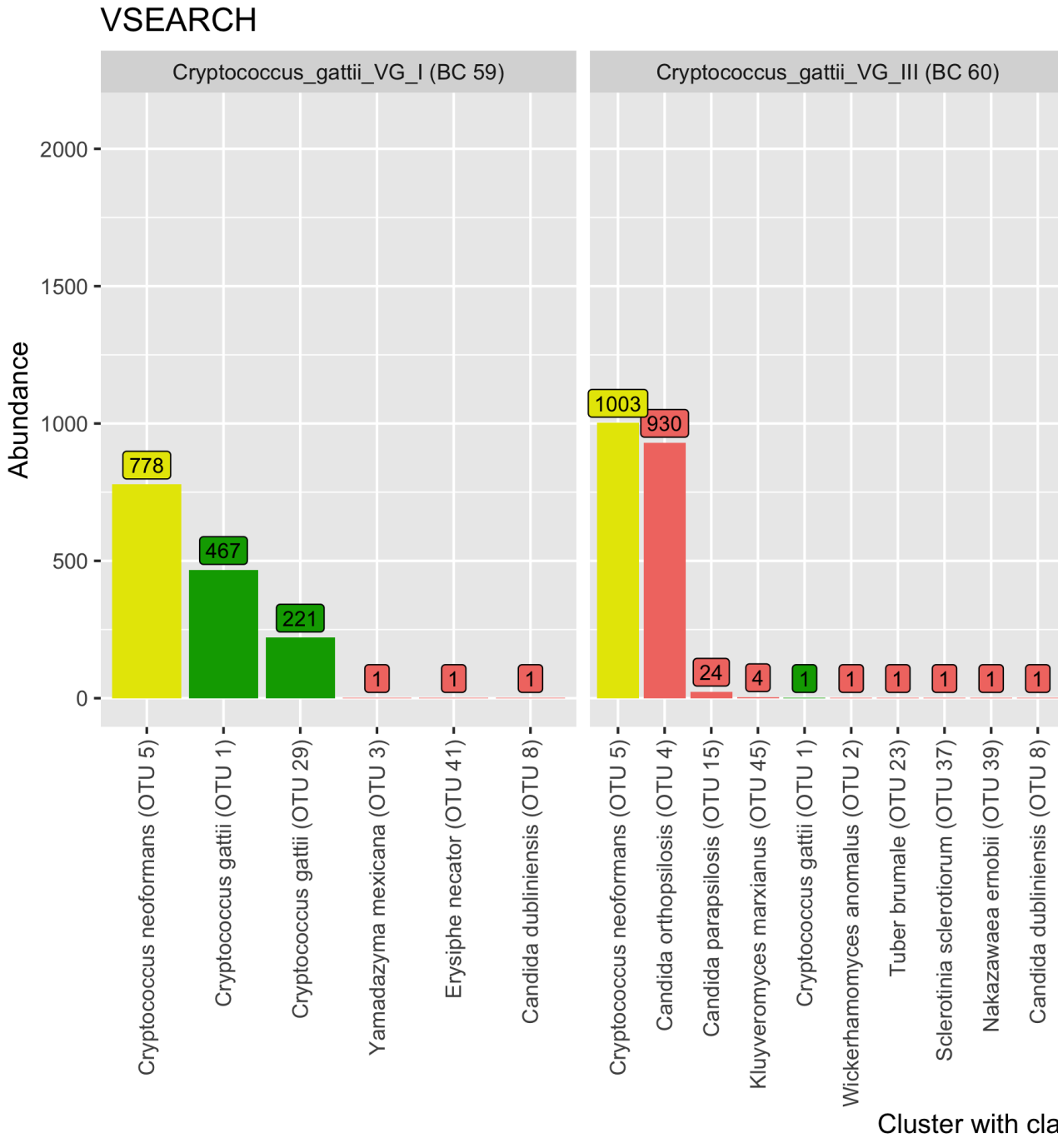


Figure 11: Plot indicating the splitting of Cryptococcus samples into clusters using the VSEARCH method.

each library size and compared by clustering methods in Figure **??**.

- NOTE: varying interpretation when classification proportion:
  - $\frac{\text{\# of reads classified at the genera level}}{\text{total \# of reads in cluster}}$ (better for nanoclust)
  - $\frac{\text{\# of reads classified at the genera level}}{\text{\# of reads per sample}}$ (similar outcomes)
- precision of nanoclust
  - using consensus

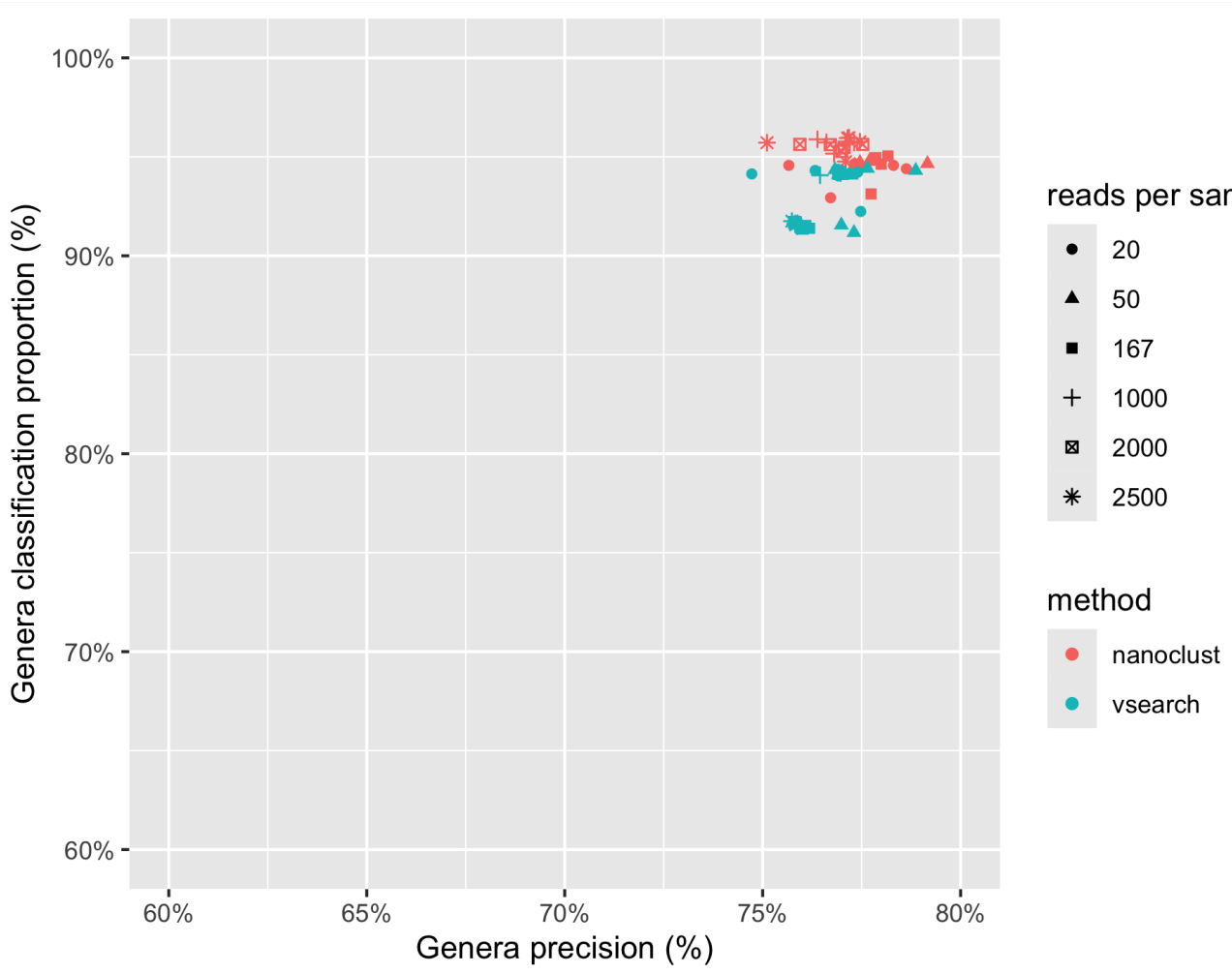

100%
90%
80%
70%
60%
Genera classification proportion (%)
60%
65%
70%
75%
80%
Genera precision (%)
reads per sar
20
50
167
1000
2000
2500
method
nanoclust
vsearch

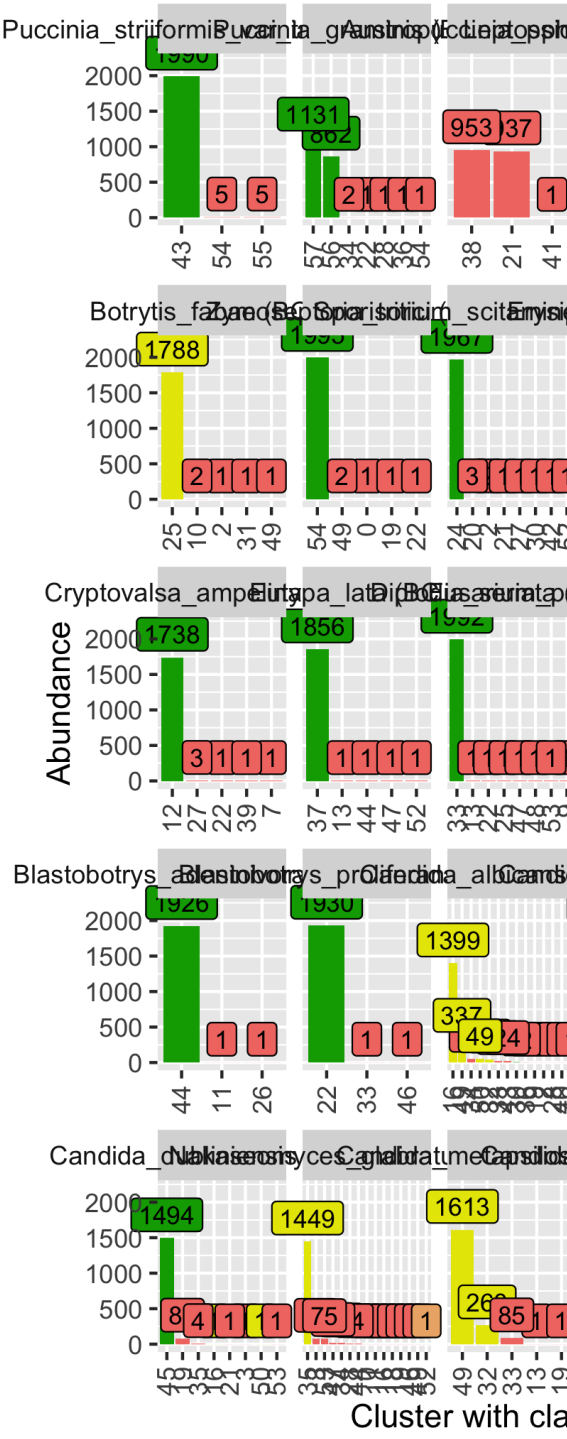
Table 1: Samplesheet showing the species, sequencing barcode and number of raw reads after basecalling.

Barcode	Sample	Number of raw reads	Number of rea
25	Puccinia striiformis var tritici	38761	
26	Puccinia triticina (Puccinia recondita)	21731	
27	Puccinia graminis	24352	
28	Austropuccinia psidii	28916	
29	Leptosphaeria maculans	40207	
30	Sclerotinia sclerotiorum	30922	
31	Botrytis cinerea	35580	
32	Botrytis fabae	28674	
33	Zymoseptoria tritici	31584	
34	Sporisorium scitamineum	29141	
35	Erysiphe necator	50836	
36	Austropuccinia psidii	29523	
37	Pyrenophora tritici-repentis	51246	
38	Cryptovalsa ampelina	37861	
39	Eutypa lata	25259	
40	Diplodia seriata	39280	
41	Fusarium pseudograminearum	37072	
42	Aspergillus flavum (Aspergillus flavus)	52132	
43	Aspergillus fumigatus	48170	
44	Aspergillus niger	106560	
45	Blastobotrys adeninivorans	56595	
46	Blastobotrys proliferans	45388	
47	Candida albicans	35572	
48	candida boletica (Candida boleticola)	45525	
49	Candida caryicola	42127	
50	Candida catenulata (Diutina catenulata)	70924	
51	Candida dublinsiensis (Candida dubliniensis)	62732	
52	Candida glabrata (Nakaseomyces glabratus)	48252	
53	Candida metapsilosis	48123	
54	Candida orthopsilosis	45771	
55	Candida parapsilosis	53561	
56	Candida tropicalis	48486	
57	Candid zeylanoides (Candida zeylanoides)	47125	
58	Cryptococcus albidus (Naganishia albida)	262	
59	Cryptococcus gattil VG I (Cryptococcus gattii VG I)	47998	
60	Cryptococcus gattii VG III	21127	
61	Cryptococcus neoformans VNI	19970	
62	Cryptococcus neoformans VN IIII	54596	
63	Fusarium proliferatum	79693	
64	Galactomyces geotrichum (Geotrichum candidum)	15962	
65	Geotrichum candidum	82477	
66	Kluyveromyces lactis	70007	
67	Kluyveromyces marxianus	70959	
68	Kodamaea ohmeri	65595	
69	Meyerozyma guillermondii (Meyerozyma guilliermondii)	29370	
70	Nakazawaea ernobil (Nakazawaea ernobii)	69285	
71	Penicillium chrysogenum	64025	
72	Pichia kudriavzevii	56683	
73	Pichia membranifacien (Pichia membranifaciens)	39163	
74	Rhodotorula mucilaginosa	53680	

5 Supplemental Tables

6 Supplemental Figures

6.1 Nanoclust splitting (all samples)



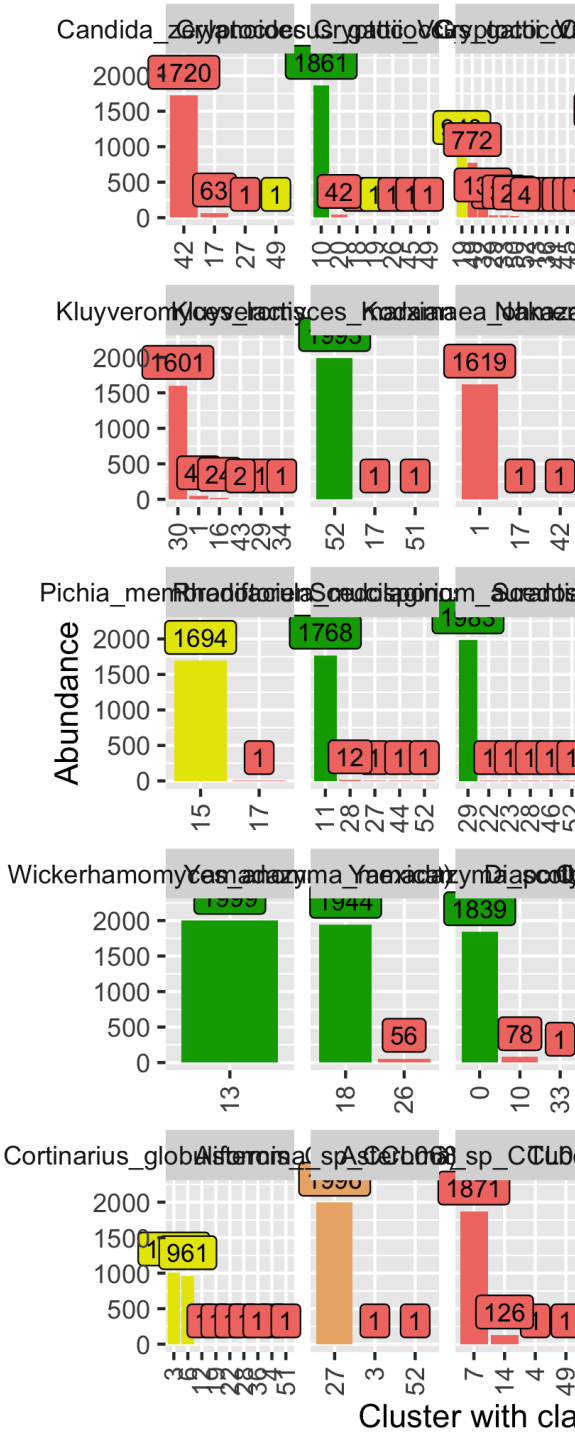
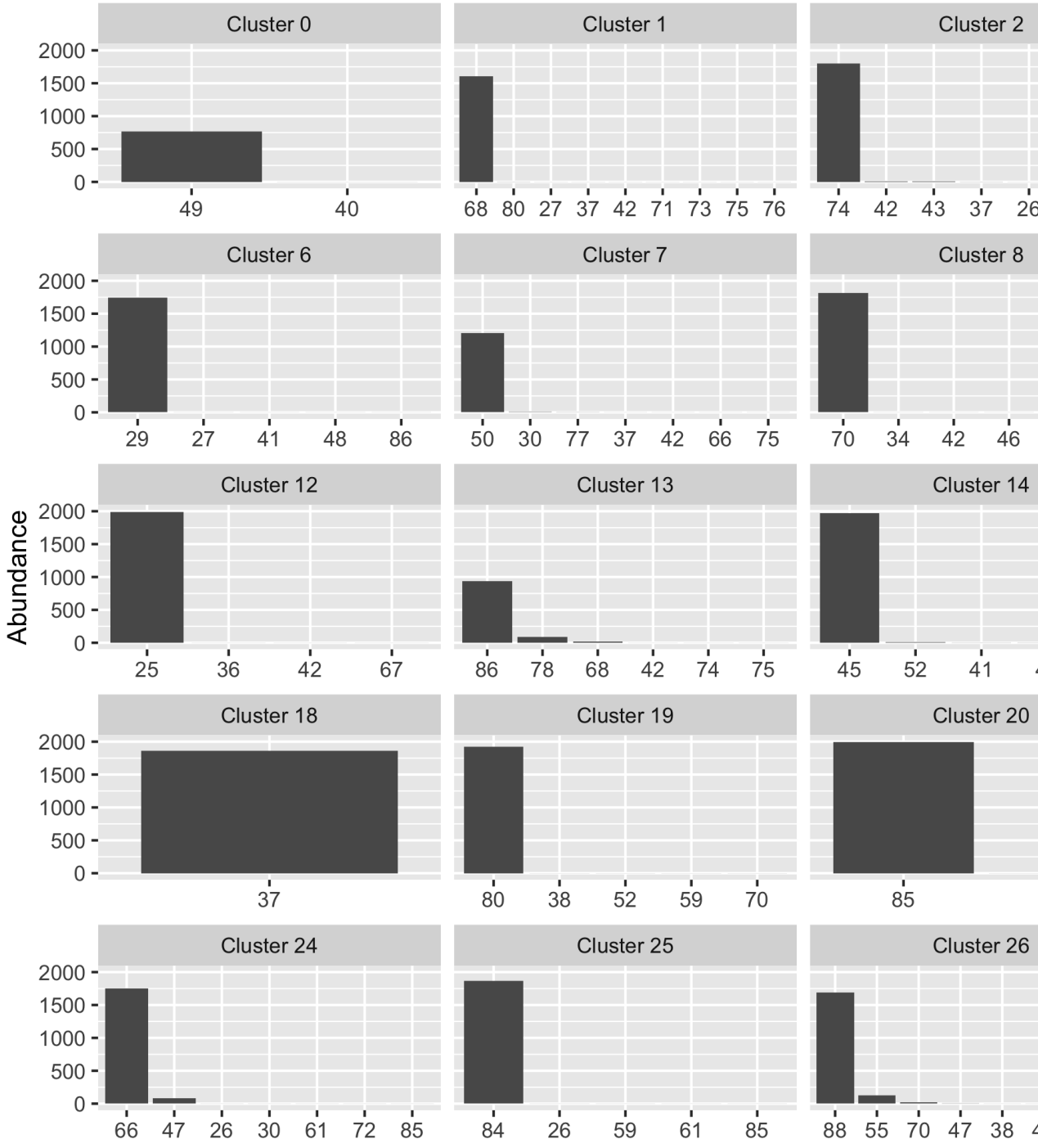
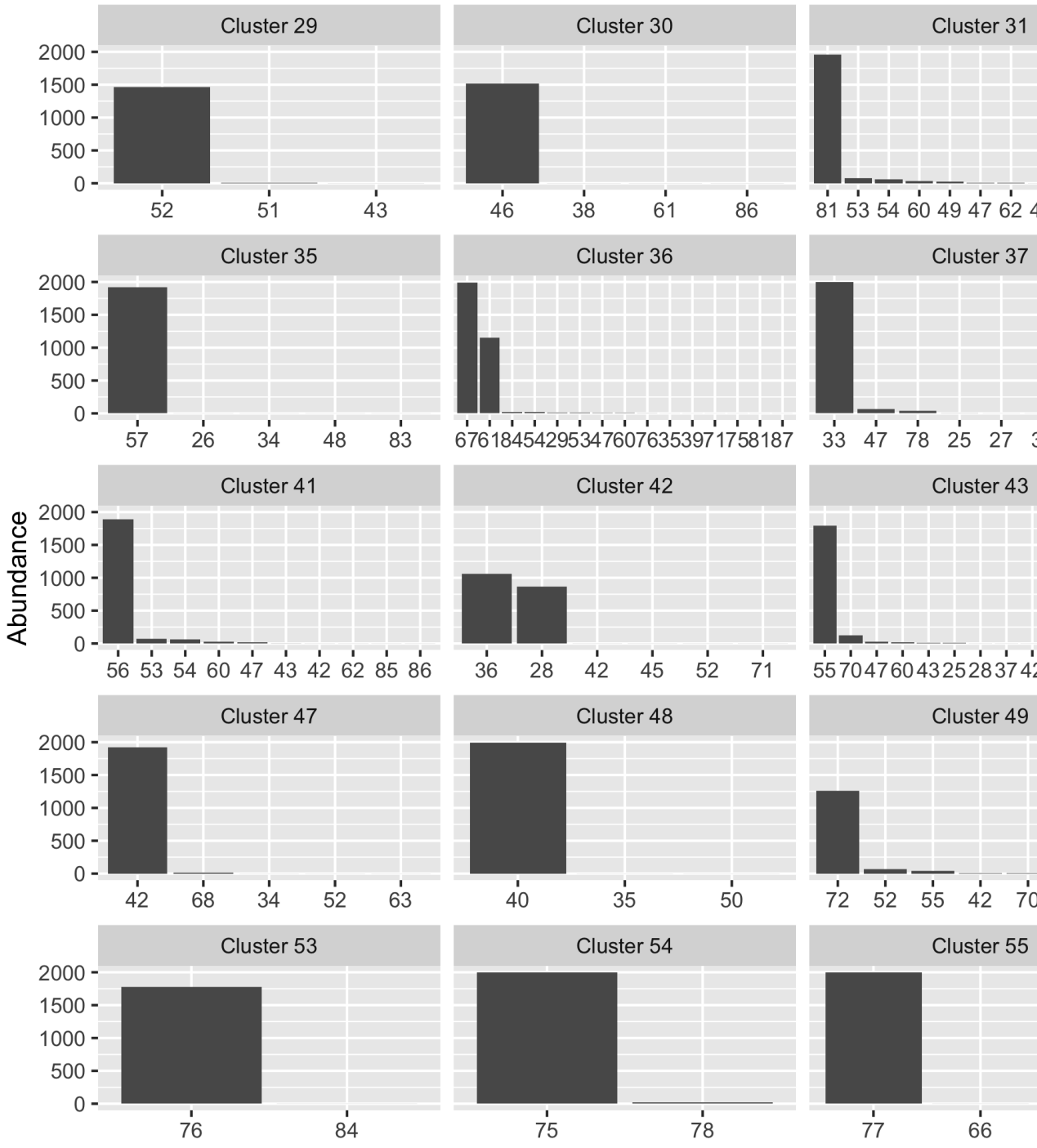


Figure 12: Plot indicating the splitting of all samples into clusters using the UMAP + HDBSCAN method. Each plot shows how reads from each sample are distributed into clusters. Bars indicate the abundance of a cluster (number of reads). Taxonomic classification of each cluster is shown in the X axis labels. Green indicates that a cluster matches the expected species-level classification.

6.2 Nanoclust Clumping (All OTUs)




Cluster 29
Cluster 30
Cluster 31
Cluster 35
Cluster 36
Cluster 37
Cluster 41
Cluster 42
Cluster 43
Cluster 47
Cluster 48
Cluster 49
Cluster 53
Cluster 54
Cluster 55
Abundance
2000
1500
1000
500
0
52 51 43
46 38 61 86
81 53 54 60 49 47 62
57 26 34 48 83
676184542953476076353971758187
33 47 78 25 27
56 53 54 60 47 43 42 62 85 86
36 28 42 45 52 71
55704760432528374
42 68 34 52 63
40 35 50
72 52 55 42
76 84
75 78
77 66

7 References

De Coster,W. and Rademakers,R. (2023) NanoPack2: Population-scale evaluation of long-read sequencing data. *Bioinformatics*, **39**, 0–2.

Langsiri,N. *et al.* (2023) Targeted sequencing analysis pipeline for species identification of human pathogenic fungi using long-read nanopore sequencing. *IMA Fungus*, **14**, 1–18.

Lh3/seqtk: Toolkit for processing sequences in FASTA/Q formats.

Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 10–12.

McInnes,L. *et al.* (2017) Hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, **2**, 205.

McInnes,L. *et al.* (2020) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

Rivers,A.R. *et al.* (2018) ITSxpress: Software to rapidly trim internally transcribed spacer sequences with quality scores for marker gene analysis [version 1; peer review: 2 approved]. *F1000Research*, **7**.

Rodríguez-Pérez,H. *et al.* (2021) NanoCLUST: A species-level analysis of 16S rRNA nanopore sequencing data. *Bioinformatics*, **37**, 1600–1601.

Rognes,T. *et al.* (2016) VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, **4**, e2584.

Vu,D. *et al.* (2022) Dnabarcoder: An open-source software package for analysing and predicting DNA sequence similarity cutoffs for fungal sequence identification. *Molecular Ecology Resources*, **22**, 2793–2809.

Vu,D. (2024) UNITE+INSD 2024 fungal ITS, ITS1, and ITS2 reference sequences.