

Article

ASVmaker: A New Tool to Improve Taxonomic Identifications for Amplicon Sequencing Data

Clément Plessis ^{1,2} , Thomas Jeanne ^{1,2,*} , Antoine Dionne ³, Julien Vivancos ³, Arnaud Droit ² 
and Richard Hogue ¹ 

¹ Institut de Recherche et de Développement en Agroenvironnement, Québec, QC G1P 3W8, Canada

² Computational Biology Laboratory, CHU de Québec—Université Laval Research Center, Québec City, QC G1V 4G2, Canada

³ Laboratoire d'Expertise et de Diagnostic en Phytoprotection, Ministère de l'Agriculture, des Pêcheries et de l'Alimentation du Québec (MAPAQ), Québec City, QC G1P 3W6, Canada

* Correspondence: thomas.jeanne@irda.qc.ca

Abstract: The taxonomic assignment of sequences obtained by high throughput amplicon sequencing poses a limitation for various applications in the biomedical, environmental, and agricultural fields. Identifications are constrained by the length of the obtained sequences and the computational processes employed to efficiently assign taxonomy. Arriving at a consensus is often preferable to uncertain identification for ecological purposes. To address this issue, a new tool called “ASVmaker” has been developed to facilitate the creation of custom databases, thereby enhancing the precision of specific identifications. ASVmaker is specifically designed to generate reference databases for allocating amplicon sequencing data. It uses publicly available reference data and generates specific sequences derived from the primers used to create amplicon sequencing libraries. This versatile tool can complete taxonomic assignments performed with pre-trained classifiers from the SILVA and UNITE databases. Moreover, it enables the generation of comprehensive reference databases for specific genes in cases where no directly applicable database exists for taxonomic classification tools.

Keywords: taxonomic assignment; QIIME2; ASV-specific database; pre-trained classifiers; public reference databases



Citation: Plessis, C.; Jeanne, T.; Dionne, A.; Vivancos, J.; Droit, A.; Hogue, R. ASVmaker: A New Tool to Improve Taxonomic Identifications for Amplicon Sequencing Data. *Plants* **2023**, *12*, 3678. <https://doi.org/10.3390/plants12213678>

Academic Editor: Assunta Bertaccini

Received: 22 August 2023

Revised: 20 October 2023

Accepted: 22 October 2023

Published: 25 October 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

1.1. Amplicon Sequencing

High-throughput sequencing approaches and, more specifically, amplicon sequencing allow the generation of a large diversity of genetic variants. They represent the relative composition of a microbial group in an environmental DNA (eDNA) sample. This molecular approach is dependent on the specific primers used [1,2], and several systems allow us to analyze the diversity of bacteria [3], fungi [4,5], and other microbial groups [6,7] detected in eDNA samples.

Computer processing of high-throughput sequencing data is essential to obtain reliable and high-quality results. To reduce the influence of sequencing errors, the first strategy has been to generate similarity clusters by defining operational taxonomic units (OTU) at a similarity threshold of 97%. This approach was suitable for early sequencing technologies (e.g., 454). Over the past decade, significant advancements have been made in the tools and methods used to process this type of data. These advancements aim to reduce sequencing errors' impact and enhance downstream analyses' accuracy. New tools such as DADA2 [8] allow the generation of genetic variants with high accuracy [9]. The algorithms use machine learning approaches to optimize sequencing error handling. With this type of processing, amplicon sequence variants (ASV) can be obtained.

1.2. Public Reference Database and Taxonomic Limitations

Taxonomic classification is an important step in sequence processing. Representative ASVs can be compared with reference sequences. Historically, there are important reference databases such as Genbank [10], DDBJ [11], and the European Nucleotide Archive (ENA) that accumulate all deposited sequencing data. These reference data are not necessarily well-balanced between species, and there are many problematic annotations. Several research groups have developed curated and aligned reference databases to facilitate the processing of high-throughput sequencing (HTS) data to facilitate taxonomic assignment. The SILVA [12], Greengenes [13], RDP [14], WarcupRDS [15], and UNITE [16] databases are widely used in microbial ecology for consensus sequence identification, but several problems remain. Each taxonomic assignment is linked to an accuracy that depends on the amount of data available in the reference databases. Additionally, a significant ratio of microbial diversity remains unknown due to the inability to cultivate several microorganisms [17]. As a result, some taxonomic assignments remain imprecise. Moreover, in cases where species exhibit high genetic similarity within the amplified region, classifiers can only reach a consensus at the genus level. Consequently, the application of HTS approaches for species identification related to some microbial genera is limited.

1.3. Available Tools to Assign Taxonomy

Bioinformatics tools for taxonomic identification are becoming more and more powerful. Some classifiers use machine learning approaches like “SKlearn” to improve classification and speed up data processing [18], while other more conventional approaches allow more parameter settings, for example, Vsearch [19] or DADA2 [8]. To make the HTS data usable and to facilitate result presentation, a consensus assignment is provided for each previously identified ASV. An accuracy calculation is possible using a pre-trained classifier, but the taxonomic assignment decision is conservative. This procedure is generally suitable for most applications in microbial ecology. However, there are limitations when it comes to identifying non-cultivated species and genera having exact similarities within the targeted gene. Finally, more specific tools propose treatments to improve the accuracy of taxonomic assignment [20]. These tools are also very dependent on the available reference databases generated by taxonomist research groups. There is a lack of tools to easily generate and use more specific reference databases for less studied genes (e.g., EF1-alpha, Beta tubulin, cytochrome oxidase II).

Here, a new tool that allows the creation of specific and usable ASV-specific reference databases for HTS data purposes is presented. This provides information on all possible identifications for each ASV and contributes to a better taxonomic assignment.

2. Materials and Methods

2.1. Environment

ASVmaker is an open-source tool available at <https://github.com/cplessis/ASVmaker>. This is a Python-based tool that is completely interoperable. It can be deployed using the Python Package Index (PyPi) Python-based tool. We recommend using it by command line. The installation and use procedure is described in the tool’s GitHub repository, available at the following address: <https://github.com/cplessis/ASVmaker> (accessed on 1 August 2023).

2.2. ASVmaker Functionalities

2.2.1. Structure

ASVmaker is designed to be used by modules (Figure 1). (1) The first step involves downloading a FASTA file for a specific genus of interest from a general database: Silva, Unite, RNACentral, ENA, NCBI, or DDBJ. This file contains the genomic data necessary for subsequent analysis. (2) Next, ASVmaker enables the creation of a genus-specific database using the downloaded FASTA file. Each sequence lineage is verified by accession number through the European Nucleotide Archive API if possible and through the NCBI Entrez API if the ENA one does not match. Users must specify primers to be used during

the simulation of the amplification process, allowing for precise targeting of the desired genomic regions and ASV creation. (3) To enhance the quality and specificity of the analysis, ASVmaker provides the functionality to filter out redundant amplicons and exclude unwanted taxonomy. Redundant amplicons are ASVs sharing the same taxonomy. Unwanted taxonomy or species that are not of interest (e.g., “sp.” or “aff.”) can also be filtered out, ensuring a more focused analysis of the target genus. (4) ASVmaker creates shared amplicon (SA) groups, which involve clustering identical ASVs with different taxonomies. This grouping allows for a comprehensive understanding of the taxonomic diversity within the selected ASV, providing valuable precisions into the composition and dynamics of microbial communities. (5) Moreover, ASVmaker offers the option to merge ASV-specific databases from different general databases, providing flexibility to combine data from various sources. When two reference databases for the same genus are built from two different FASTA files, it is possible to merge them. This step creates new SA groups if necessary and eliminates duplicates. This merging process allows for a more comprehensive dataset, enabling comparative analysis and broader insights into the studied genus.

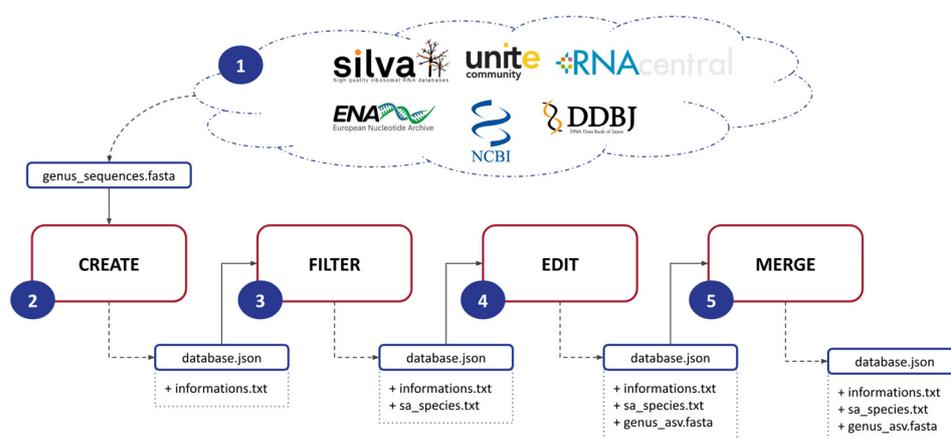


Figure 1. Schematic of the five steps of ASVmaker’s process. (1) Download the FASTA file for one genus from a general database, (2) produce the database for a specific genus and primers, (3) filter redundant amplicons or unwanted taxonomy, (4) produce shared amplicon (SA) groups, and (5) prepare facultative merging of specific genus ASV databases from different general databases.

2.2.2. Taxonomy

A taxon is defined as the most precise taxonomic description that can be obtained for a sequence variant. With the currently available tools, a variant with a different possible taxonomy is, by default, assigned to the consensus taxonomy at a truncated level (e.g., “GenusName_spp.”). This results in the loss of crucial information. A solution to this problem is to assign a group of species sharing the same amplified sequence as a taxon. In this case, the amplicon is defined as “Shared Amplicon” (SA). The taxon of an ASV related to a single species will, therefore, be “GenusName_SpeciesName.” The taxon of an ASV related to several species will be “GenusName_SAn,” where “n” is the sequential number of the SA in the database. The choice to group these sequences under the name of the SA is an important step to avoid losing information on genetic variability. Thus, the identification by HTS will return a maximum of answers to the user without passing by a consensus attribution. Hence, it is possible to attribute a taxonomic identification by grouping very similar sequences. The SA groups give the same information as Blast at 100% identity on same-length sequences for multiple species but stored in the sequence taxonomy. ASVmaker does not rely on any specific algorithm, unlike other classifiers.

2.2.3. Amplicon

To create an ASV-specific database, a simulation of amplification must be performed on all the sequences to select the amplifiable fragments. The original amplification system was based on the PCR function of the Python package Pydna [21]. However, a custom module was created because this package does not offer customizable parameters for primer mismatch tolerance. This module uses local primer alignment scores on a given sequence. To favor the positions where the primer can attach, the calculation of the scores favors match and mismatches rather than gaps: match +1, mismatch 0, open-gap −1, extend-gap −0.5. The “sense” leader is directly aligned to the sequence, and the position with the highest score is saved. Then, the complementary strand of the reverse primer is synthesized before it is also aligned to the sequence. If the alignment score of the two primers passes the threshold set by the user, then the amplicon is generated on the primer positions with or without end primers (as desired). An amplicon is created only if the last three bases at the 3′ ends do not contain any mismatches.

2.2.4. Usage

When seeking to identify an ASV from the amplification of a large microbial group (e.g., bacteria or fungi), an ASV-specific database generated by ASVmaker can be used on ASVs that have generated an initial identification at the genus level. This constitutes a case of double identification, firstly by a general database such as Silva or UNITE and secondly with the specific one from ASVmaker. For other applications, when dealing with ASVs generated by the amplification of a specific genus (e.g., *Fusarium* for the EF1-alpha gene), the specific database generated by ASVmaker can be used directly. In all cases, the taxonomic assignment with the specific reference database must be used with 100% alignment and 100% coverage.

2.3. Creation of a New Database

To evaluate the performance of ASVmaker, ASV-specific reference databases have been generated. It was chosen as a microbial genus that may include plant pathogens. This application in plant pathology is not the only one, but it was chosen because we are involved in a project to evaluate the potential of HTS for the identification of several plant pathogenic organisms. For bacteria, the targeted genera have been *Erwinia*, *Streptomyces*, *Pseudomonas*, and *Xanthomonas*. The fungal genera have been *Colletotrichum*, *Septoria*, *Ustilagi*, and *Verticillium*. These ASV-specific databases can be combined with the taxonomic assignment with the pre-trained classifiers (SILVA version 138 or UNITE version 8.3) to improve the species-level identification.

An additional ASV-specific database has been generated to present an example of a direct and specific amplification targeting a non-ribosomal gene. The *Fusarium* elongation factor alpha gene was targeted. In the 3 targeted examples, we used ASVmaker with the primers described in Table 1 and sequences from queries from UNITE for fungal genera, SILVA for bacterial ones, and RNACentral for both. Since there is no sequence of EF1α in the UNITE database, the sequences available from the ENA database were downloaded.

Table 1. List of primers used to produce the ASV-specific database and for the amplifications performed on the environmental DNAs.

Microbial Group	Target	Region	Forward Primer	Reverse Primer	Reference
Bacteria	16S	V4V5	515FB: GTGYCAGCMGCCGCGTAA	926R: CCGYCAATTYMTTTRAGTTT	[22,23]
Fungi	ITS	ITS1	BITS: ACCTGCGGARGGATCA	B58S3 GAGATCCRTTGYTRAAAGTT	[24]
Fusarium	EF1α	EF1α	Fa-150: CCGGTCACTTGATC- TACCAG	Ra-2: ATGACGGTGACATAGTAGCG	[25]

2.4. Application on Environmental Samples

To provide examples of applications, plant samples from a large study focused on the potential for identifying plant pathogenic organisms using HTS were used. These examples compared the identification process using public reference databases (SILVA, UNITE) to the dual identification method based on the reference database generated with ASVMaker.

2.4.1. Sample and DNA Extraction

Plant tissues were collected by the Ministère de l'Agriculture, des Pêcheries et de l'Alimentation du Québec (MAPAQ) plant pathologists based on specific disease symptoms. The fresh tissues were homogenized, and 0.2 g were used for DNA extraction. DNA extractions were performed with the DNeasy Plant Mini Kit (Qiagen, Mississauga, ON, Canada) according to the manufacturer's instructions. Each DNA pellet was suspended in 100 µL of sterile molecular-grade deionized water. The quality and quantity of the DNA extracts were evaluated by spectrophotometry using a Biophotometer (Eppendorf, Mississauga, ON, Canada) with readings at 260, 280, 230, and 320 nm.

2.4.2. Amplicon Sequencing

Prokaryote and fungal diversity were assessed by HTS as described [26], using 515FB and 926R primers and BITS-ITS1 and B58S3 primers, respectively, for bacteria and fungi. Specific *Fusarium* spp. amplification was performed using the primers Fa-150 and Ra-2, targeting the elongation factor 1-alpha gene (Table 1). Briefly, a two-step dual-indexed PCR approach was specifically designed for Illumina instruments by the Plateforme d'analyses génomiques (IBIS, Université Laval, Quebec City, QC, Canada) was performed. Indexed PCR products were purified, checked for quality on a DNA7500 Bioanalyzer chip (Agilent, Santa Clara, CA, USA), and then quantified spectrophotometrically using the Biophotometer with a G1.0 µCuvette. Barcoded amplicons were pooled in equimolar concentrations for sequencing on the Illumina MiSeq platform using a 2 × 300 bp sequencing kit.

2.4.3. Bioinformatic Analysis

Raw MiSeq sequences (FASTQ) were filtered under the QIIME2 platform [27] using the DADA2 plugin [8] filtration approach for determining amplicon sequence variants (ASV). For fungi sequences of the ITS1 region, primers were previously removed with the Cutadapt tool [28].

Taxonomic assignments were carried out using a classification approach with the sklearn function in the q2-feature-classifier plugin [18] and pre-trained classifiers from the SILVA (version 138) and UNITE (version 8.3) databases for bacteria and fungi, respectively. The secondary assignment was generated with 100% similarity identification using the ASV-specific database obtained with ASVMaker. For the specific EF1α gene, the ASV-specific database generated from the EF1α sequences was used directly.

3. Results

3.1. ASV Specific Database for 16S rRNA, ITS and EF1α Gene

Three ASV-specific databases were created to showcase ASVMaker use cases (all code available in the "data and code availability" section). Two of these were designed to complete the analysis with the Silva and UNITE pre-trained classifiers. Four bacterial genera and four fungal genera were chosen to present a simple and complex case study for each microbial group targeted in the phytopathological application. The raw sequences were then retrieved from the Silva database for bacterial genus and primers targeting the 16S region, from the Unite database for fungal genus and primers targeting the ITS region, and from the RNACentral database for both. The tool can concatenate specific ASV bases from different generalist bases (Figure 2). For all the genera studied, ASVMaker made it possible to increase the number of variants by concatenating the two generalist bases. The developed tool enables us to better characterize identical variants with different taxonomies

(SA). These variants represent, on average, 10% of the ASVs of the four bacterial genera and 11% of the fungal ASVs.

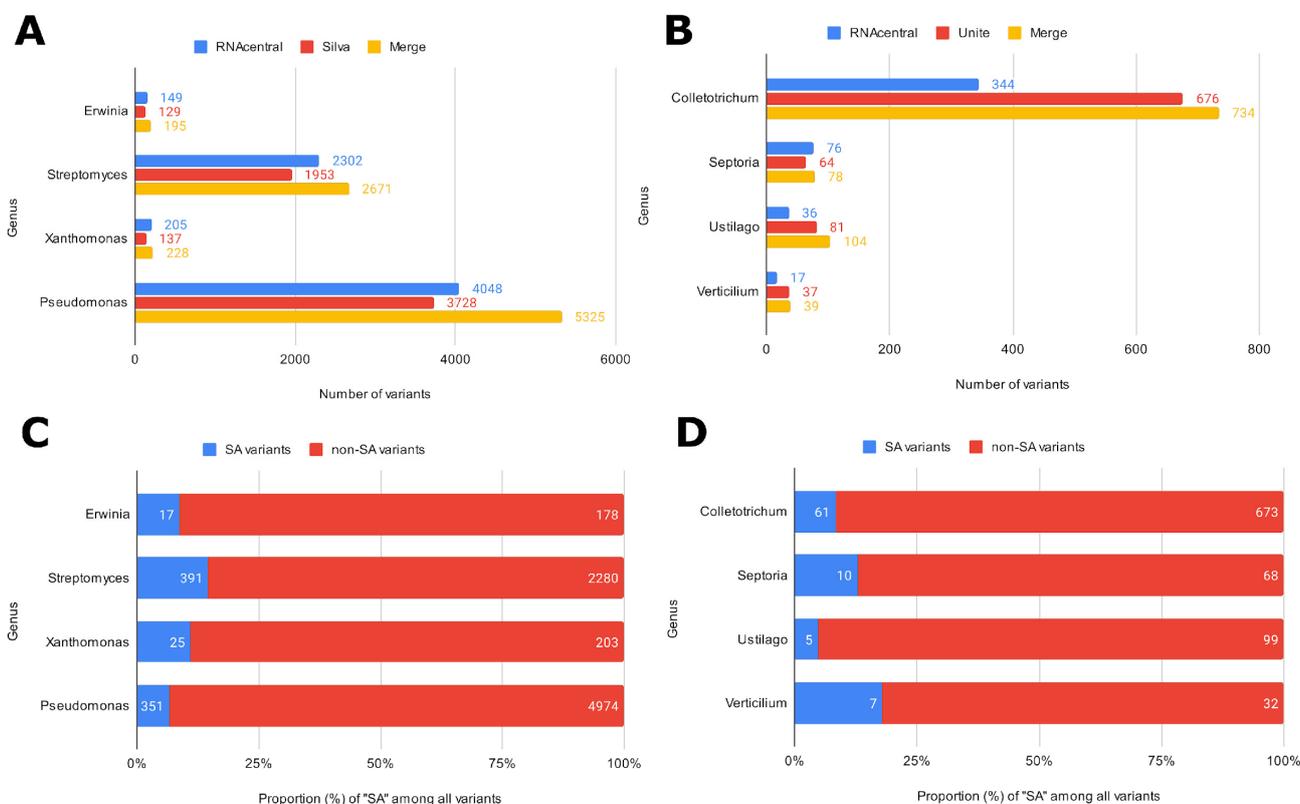


Figure 2. Number of variants retained according to the data source and ASV combination for bacterial genus (A) and targeted fungal genus (B). Number of unique variants (non-SA) or variants with at least one different taxonomic identification (SA) for bacterial genus (C) and fungal genus (D).

For the third example, we targeted the gene *EF1 α* to evaluate the *Fusarium* species diversity. This ASV-specific database was created from sequences present in a non-specialized generalist database to reach a better diversity of sequences for less studied genes, the ENA (European Nucleotide Archive). However, these databases may have taxonomic assignment errors on their sequences, unlike databases such as Silva and UNITE, which are more accurate. A total of 43,509 raw sequences were retrieved from the ENA website. After processing with ASVmaker, 3353 unique variants were identified, including 126 SA variants and 2784 species complex variants (Figure 3A). A total of 77 unique species taxa (including species complex) and 126 SA taxa were isolated in the *Fusarium* *EF1 α* ASV specific database for a total of 203 possible taxonomic attributions. Most of the variants of the created specific database targeting the gene *EF1 α* are species complex taxa or SA taxa (Figure 3B).

3.2. Environmental Samples Application

One possible application of ASVmaker is to provide an additional level of information aiming at plant pathogen identification. The use of high-throughput sequencing could complement or enhance phytopathologists' ability to detect plant diseases. As part of a large-scale study in collaboration with the MAPAQ's phytopathologists, several hundred diseased plants were tested, and plant pathogens identification was obtained using conventional methods (Microscopic, qPCR) and with HTS w compared. To illustrate the benefits of using the databases generated with ASVmaker, samples that could be used in the five following situations (code C1 to C5) were identified:

- C1: Confirmation of the identification obtained with pre-trained classifiers (from the Silva/UNITE databases) with the ASV-specific database;

- C2: Precision increase to the species level with the ASV-specific database;
- C3: Change of species identification with the ASV-specific database;
- C4: Precision obtained with the ASV-specific database with a few species possibilities (simple case);
- C5: Precisions obtained with the ASV-specific database with several species possibilities (complex case).

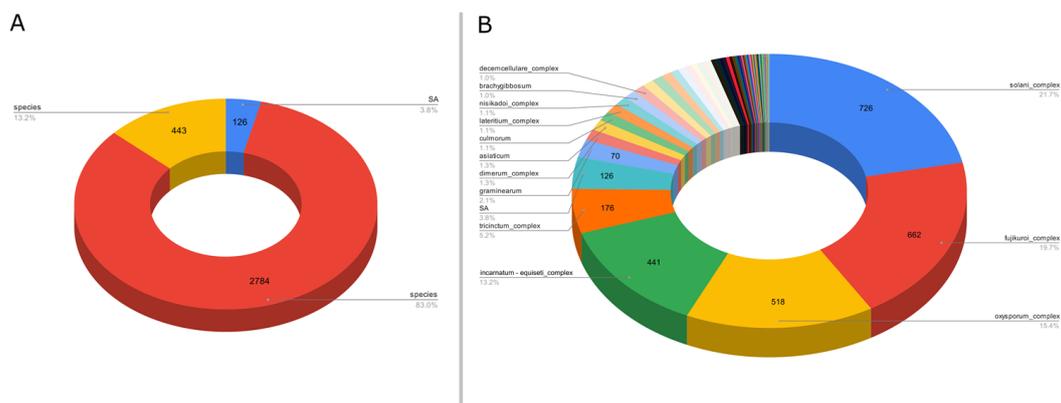


Figure 3. (A) Proportion of variants taxonomically assigned to species, species complex, or SA from 3353 variants *Fusarium* EF1 α database created from 43,509 sequences retrieved from the ENA. (B) Major taxa proportion among 3353 variants in the *Fusarium* EF1 α database created from 43,509 sequences retrieved from the ENA.

Table 2 shows the results obtained for the taxonomic identification of the selected cases and according to the overall diversity of bacteria, fungi, and fusarium-specific diversity determined by EF1 α gene diversity. A first interpretation illustrates that, whatever the microbial group, it can be easy or more complex to make a good taxonomic identification with HTS data. It is, therefore, not possible to generalize about identification problems. On the other hand, the cases selected for bacteria present more problems compared to fungi. Without being exhaustive, identifications are more problematic for *Pseudomonas*, *Xanthomonas*, and *Streptomyces*, and the number of possible species can vary widely (from a few species to 44). However, ASV-specific databases can improve taxonomic identifications, such as Cases 3 and 5 for *Streptomyces*, or enable identification at the species level, such as Case 3 for *Erwinia tracheiphila*.

On the other hand, the taxonomic identification improvement provided by ASV-specific databases can be used to discriminate variants potentially associated with a given species. In a case when *Pseudomonas syringae* is targeted, it is possible to discard some variants that do not present this species in the shared amplicon list.

In the case of fungi, identifications are generally more accurate. Examples in Table 2 illustrate these observations with the identifications of *Colletotrichum*, *Ustilago*, and *Verticillium*. For *Colletotrichum*, the secondary identification detailed a more problematic identification with three possible species against one with the pre-trained classifier (Case 7) or change the species identification (Cases 9 and 10). This example highlights the problem of dataset training size of the classifiers. The same observations are reported for *Verticillium* with two possible species identified with ASV-specific database (cases 7, 9, and 10) and for a more problematic case with *Septoria* (Case 9). However, for *Ustilago*, which was a simple case, the same identification was obtained with both databases. The same observations generally apply to other genus.

Table 2. Detailed results of the best-taxonomic identifications obtained with the pre-trained classifiers from the SILVA and UNITE databases and with the ASV-specific database created with ASVmaker for the selected samples. The table shows three sections for the amplification system targeting bacteria, fungi, and specifically *Fusarium* spp. using the *EF1 α* gene.

Sample Code	Crop	Diagnostic (Conventional)	Code	Best Taxonomic Identification by Pre-Trained Classifiers (SILVA/UNITE)	Conf.	Complementary Identification with ASV-Specific Database	Shared Amplicon (SA)	Relative Abund. (%)
Bacterial amplification—16S rADN								
Case1	Squash	<i>Pseudomonas_syringae</i>	C5	<i>Pseudomonas</i>	1	<i>Pseudomonas_SA46</i>	35 species	8.55
			C5	<i>Pseudomonas</i>	1	<i>Pseudomonas_SA63</i>	28 species	30.89
Case2	Cabbage	<i>Xanthomonas campestris</i>	C5	<i>Xanthomonas</i>	0.997	<i>Xanthomonas_SA1</i>	27 species	57.35
Case3	Squash	<i>Erwinia tracheiphila</i>	C1	<i>Erwinia tracheiphila</i>	1	<i>Erwinia tracheiphila</i>	44 species	31.03
			C5	<i>Pseudomonas</i>	1	<i>Pseudomonas_SA22</i>		
			C4	<i>Streptomyces</i>	1	<i>Streptomyces_SA167</i>	<i>S.roseirectus</i> <i>S.niveiscabiei</i> <i>S.acidiscabies</i>	0.06
Case4	Cabbage	<i>Xanthomonas campestris</i>	C5	<i>Xanthomonas</i>	0.997	<i>Xanthomonas_SA1</i>	27 species	63.11
Case5	Wheat	<i>Xanthomonas campestris</i>	C5	<i>Pseudomonas</i>	1	<i>Pseudomonas_SA22</i>	44 species	0.26
			C5	<i>Xanthomonas</i>	0.997	<i>Xanthomonas_SA1</i>	27 species	3.97
Case5	Potato	<i>Streptomyces scabies</i>	C5	<i>Xanthomonas</i>	1	<i>Xanthomonas_SA3</i>	15 species	49.50
			C5	<i>Pseudomonas</i>	1	<i>Pseudomonas_SA39</i>	30 species	0.18
			C5	<i>Pseudomonas</i>	1	<i>Pseudomonas_SA46</i>	35 species	0.92
			C5	<i>Streptomyces</i>	0.999	<i>Streptomyces_SA63</i>	25 species	7.19
			C2	<i>Streptomyces</i>	1	<i>Streptomyces scabrisporus</i>		0.36
Fungal amplification—ITS1								
Case7	Potato	<i>Colletotrichum, Dickeya sp., Fusarium, Pythium, Verticillium</i>	C4	<i>Colletotrichum coccodes</i>	0.998	<i>Colletotrichum_SA61</i>	<i>C.nigrum</i> <i>C.coccodes</i>	24.46
			C4	<i>Verticillium nubilum</i>	0.998	<i>Verticillium_SA1</i>	<i>C.gloeosporioides_complex</i> <i>V.longisporum</i> <i>V.dahliae</i>	
Case8	Corn	<i>Ustilago maydis</i>	C1	<i>Ustilago maydis</i>	1	<i>Ustilago maydis</i>		13.32
Case9	Melon	<i>Verticillium dahliae</i>	C3	<i>Colletotrichum fuscum</i>	1	<i>Colletotrichum destructivum</i> complex		0.10
			C5	<i>Septoria epilobii</i>	0.93	<i>Septoria_SA3</i>	38 species	0.01
			C4	<i>Verticillium nubilum</i>	0.998	<i>Verticillium_SA1</i>	<i>V.longisporum</i> <i>V.dahliae</i>	12.51

Table 2. Cont.

Sample Code	Crop	Diagnostic (Conventional)	Code	Best Taxonomic Identification by Pre-Trained Classifiers (SILVA/UNITE)	Conf.	Complementary Identification with ASV-Specific Database	Shared Amplicon (SA)	Relative Abound. (%)
Case10	Melon	<i>Verticillium_dahliae</i>	C3	<i>Colletotrichum fuscum</i>	0.998	<i>Colletotrichum destructivum</i> complex		0.01
			C4	<i>Verticillium nubilum</i>	0.998	<i>Verticillium_SA1</i>	<i>V.longisporum</i> <i>V.dahliae</i>	28.52
Fusarium-specific amplification—EF1A (only with ASV-specific database)								
Case11	Corn	<i>Fusarium graminearum</i> , <i>Fusarium avenaceum</i>	C2			<i>Fusarium tricinctum</i> complex		10.01
			C2			<i>Fusarium tricinctum</i> complex		23.89
Case12	Corn	<i>Fusarium sporotrichoides</i> <i>Fusarium graminearum</i> <i>Fusarium equiseti</i>	C2			<i>Fusarium fujikuroi</i> complex		28.56
			C2			<i>Fusarium incarnatum-equiseti</i> complex		67.20
Case13	Corn	<i>Kebatiellose</i> <i>Fusarium</i>	C2			<i>Fusarium incarnatum-equiseti</i> complex		1.11
			C4			<i>Fusarium_SA89</i>	<i>F.incarnatum equiseti</i> complex <i>F.sporotrichioides</i> <i>F.asiaticum</i> <i>F.armeniicum</i> <i>F.boothii</i>	46.91
			C4			<i>Fusarium_SA93</i>	<i>F.graminearum</i> <i>F.meridionale</i>	3.19
			C2			<i>Fusarium_serpentinum</i>		0.97
			C2			<i>Fusarium_sporotrichioides</i>		0.79
			C2			<i>Fusarium_sporotrichioides</i>		0.53
			C2			<i>Fusarium_tricinctum</i> complex		7.37

Samples analyzed for the *Fusarium*-specific gene (*EF1 α*) generally showed a very good level of identification. Unlike the application for bacteria and fungi, the results for the *EF1 α* gene allow direct identification. The identifications obtained by HTS and ASV-specific databases can be corroborated with microbial isolations on selective media. In all samples where *Fusarium* spp. was identified by isolation, it was possible to obtain identification by HTS. On the other hand, species identifications may be different or expressed by different names or species complexes. Identifications coupled with relative abundance enable the identification of variants detected in the same sample and to assess their respective representation. Except for the *Fusarium_SA89* and *Fusarium_SA93* variants, which have 2 and 6 possible identifications, respectively, all other variants are identified as species or species complex.

4. Discussion

ASVmaker is a specialized tool that addresses various application gaps using amplicon sequencing data. It offers additional taxonomic information to confirm species identification or improve identification challenges encountered with conventional classifiers.

While many existing tools aim to enhance taxonomic attributions through database generation, either by refining existing databases or employing more powerful algorithms [29–31], ASVmaker is more specifically designed to target a particular genus or a list of genera, adapting accordingly to the primers used in sequencing library preparation.

ASVmaker can also be used to improve a specific already-generated ASV database. The merge function allows the addition and integration of additional sequences into a newly documented structure. However, it is important to note that ASVmaker is not able to treat multiple genera simultaneously. In this study, it was tested on 10 bacterial genera and 38 fungal genera. As the tool does not address inter-genus issues, employing it as a subsequent step following taxonomic assignment with a pre-trained classifier is crucial.

Additionally, ASVmaker can be used to generate a genus-specific ASV reference database for non-ribosomal genes. The results with the *EF1 α* gene showed that ASVmaker can improve taxonomic assignment directly compared to other studies using conventional classifiers [32]. Identifying species with conventional classifiers can be difficult due to conflicts with multiple taxonomies for a single variant. However, ASVmaker can isolate and retain this information in the taxonomic assignment. It is feasible to prepare similar reference databases for other genes of interest in microbial ecology, such as beta-tubulin or cytochrome oxidase II.

Presently, ASVmaker is restricted to data generated from the Illumina platform, as it requires high-quality sequences for successful implementation. Therefore, using an ASV-specific database on sequences from sequencing approaches involving Oxford Nanopore Technology (ONT) is not feasible. Conversely, it may exhibit promising performance for approaches such as Pacbio or other techniques generating high-quality sequences.

5. Conclusions

By allowing users to easily prepare their own ASV-specific database and complete the taxonomic annotation from public pre-trained classifiers, ASVmaker will enable researchers in microbial ecology to improve taxonomic identifications for specific microbial genera. The use of ASV-specific databases does not guarantee precise microbial species identification but clarifies potential issues with pre-trained classifiers. ASVmaker also proves to be a powerful tool for constructing a genus-specific ASV reference database for non-ribosomal genes. It was tested on the *EF1 α* gene, and it achieved highly interesting performance, obtaining species-specific identifications in most cases. This tool has a wide range of applications, including plant pathology, studying the results of microbial inoculants and biostimulants, as well as applications in biomedical research.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/plants12213678/s1>.

Author Contributions: Conceptualization, C.P., T.J., A.D. (Arnaud Droit), and R.H.; methodology, C.P. and T.J.; software, C.P. and T.J.; diagnostic validation, A.D. (Antoine Dionne) and J.V.; resources, R.H., and A.D. (Arnaud Droit); data curation, C.P. and T.J.; writing—original draft preparation, C.P. and T.J.; writing—review and editing, C.P., T.J., A.D. (Antoine Dionne), J.V., A.D. (Arnaud Droit), and R.H.; project administration, R.H. and A.D. (Arnaud Droit); funding acquisition, R.H. and A.D. (Arnaud Droit). All authors have read and agreed to the published version of the manuscript.

Funding: This project was carried out under subcomponent 2.2 of the Prime-Vert 2018-2023 program and was funded by the Ministère de l’Agriculture, des Pêcheries et de l’Alimentation du Québec (MAPAQ). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability Statement: Supplementary Materials describe the Python notebooks used to generate ASV-specific databases and analyze environmental samples. All raw sequence files (FASTQ) have been deposited on NCBI. <https://www.ncbi.nlm.nih.gov/sra/PRJNA1023220> (accessed on 1 August 2023).

Acknowledgments: We want to extend a special thank you to Vanessa Villeneuve for her contribution to conducting the molecular analyses.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Mbareche, H.; Veillette, M.; Bilodeau, G.; Duchaine, C. Comparison of the performance of ITS1 and ITS2 as barcodes in amplicon-based sequencing of bioaerosols. *PeerJ* **2020**, *8*, e8523. [\[CrossRef\]](#)
2. Bukin, Y.S.; Galachyants, Y.P.; Morozov, I.V.; Bukin, S.V.; Zakharenko, A.S.; Zemskaya, T.I. The effect of 16S rRNA region choice on bacterial community metabarcoding results. *Sci. Data* **2019**, *6*, 190007. [\[CrossRef\]](#)
3. Abellan-Schneyder, I.; Machado, M.S.; Reitmeier, S.; Sommer, A.; Sewald, Z.; Baumbach, J.; List, M.; Neuhaus, K. Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing. *mSphere* **2021**, *6*, e01202-20. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Tedersoo, L.; Lindahl, B. Fungal identification biases in microbiome projects: Fungal identification biases in microbiome projects. *Environ. Microbiol. Rep.* **2016**, *8*, 774–779. [\[CrossRef\]](#)
5. Schoch, C.L.; Seifert, K.A.; Huhndorf, S.; Robert, V.; Spouge, J.L.; Levesque, C.A.; Chen, W.; Fungal Barcoding Consortium; Fungal Barcoding Consortium Author List; Bolchacova, E.; et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 6241–6246. [\[CrossRef\]](#)
6. Bahram, M.; Anslan, S.; Hildebrand, F.; Bork, P.; Tedersoo, L. Newly designed 16S rRNA metabarcoding primers amplify diverse and novel archaeal taxa from the environment. *Environ. Microbiol. Rep.* **2018**, *11*, 487–494. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Comeau, A.M.; Vincent, W.F.; Bernier, L.; Lovejoy, C. Novel chytrid lineages dominate fungal sequences in diverse marine and freshwater habitats. *Sci. Rep.* **2016**, *6*, 30120. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Callahan, B.J.; Mcmurdie, P.J.; Rosen, M.J.; Han, A.W.; Johnson, A.J.A.; Holmes, S.P. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **2016**, *13*, 581–583. [\[CrossRef\]](#)
9. Prodan, A.; Tremaroli, V.; Brodin, H.; Zwinderman, A.H.; Nieuwdorp, M.; Levin, E. Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PLoS ONE* **2020**, *15*, e0227434. [\[CrossRef\]](#)
10. Benson, D.A.; Cavanaugh, M.; Clark, K.; Karsch-Mizrachi, I.; Lipman, D.J.; Ostell, J.; Sayers, E.W. GenBank. *Nucleic Acids Res.* **2013**, *41*, D36–D42. [\[CrossRef\]](#)
11. Tateno, Y. DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic Acids Res.* **2002**, *30*, 27–30. [\[CrossRef\]](#)
12. Quast, C.; Pruesse, E.; Yilmaz, P.; Gerken, J.; Schweer, T.; Yarza, P.; Peplies, J.; Glöckner, F.O. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **2013**, *41*, D590–D596. [\[CrossRef\]](#)
13. DeSantis, T.Z.; Hugenholtz, P.; Larsen, N.; Rojas, M.; Brodie, E.L.; Keller, K.; Huber, T.; Dalevi, D.; Hu, P.; Andersen, G.L. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Appl. Environ. Microbiol.* **2006**, *72*, 5069–5072. [\[CrossRef\]](#)
14. Martin, D.; Rybicki, E. RDP: Detection of recombination amongst aligned sequences. *Bioinformatics* **2000**, *16*, 562–563. [\[CrossRef\]](#)
15. Deshpande, V.; Wang, Q.; Greenfield, P.; Charleston, M.; Porrás-Alfaro, A.; Kuske, C.R.; Cole, J.R.; Midgley, D.J.; Tran-Dinh, N. Fungal identification using a Bayesian classifier and the Warcup training set of internal transcribed spacer sequences. *Mycologia* **2016**, *108*, 1–5. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Nilsson, R.H.; Larsson, K.-H.; Taylor, A.F.S.; Bengtsson-Palme, J.; Jeppesen, T.S.; Schigel, D.; Kennedy, P.; Picard, K.; Glöckner, F.O.; Tedersoo, L.; et al. The UNITE database for molecular identification of fungi: Handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Res.* **2019**, *47*, D259–D264. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Pham, V.H.; Kim, J. Cultivation of unculturable soil bacteria. *Trends Biotechnol.* **2012**, *30*, 475–484. [\[CrossRef\]](#)

18. Bokulich, N.A.; Kaehler, B.D.; Rideout, J.R.; Dillon, M.; Bolyen, E.; Knight, R.; Huttley, G.A.; Gregory Caporaso, J. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* **2018**, *6*, 90. [[CrossRef](#)] [[PubMed](#)]
19. Rognes, T.; Flouri, T.; Nichols, B.; Quince, C.; Mahé, F. VSEARCH: A versatile open source tool for metagenomics. *PeerJ* **2016**, *2016*, e2584. [[CrossRef](#)] [[PubMed](#)]
20. Zahariev, M.; Chen, W.; Visagie, C.M.; Lévesque, C.A. Cluster oligonucleotide signatures for rapid identification by sequencing. *BMC Bioinform.* **2018**, *19*, 395. [[CrossRef](#)] [[PubMed](#)]
21. Pereira, F.; Azevedo, F.; Carvalho, A.; Ribeiro, G.F.; Budde, M.W.; Johansson, B. Pydna: A simulation and documentation tool for DNA assembly strategies using python. *BMC Bioinform.* **2015**, *16*, 142. [[CrossRef](#)] [[PubMed](#)]
22. Parada, A.E.; Needham, D.M.; Fuhrman, J.A. Every base matters: Assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples: Primers for marine microbiome studies. *Environ. Microbiol.* **2016**, *18*, 1403–1414. [[CrossRef](#)] [[PubMed](#)]
23. Apprill, A.; McNally, S.; Parsons, R.; Weber, L. Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquat. Microb. Ecol.* **2015**, *75*, 129–137. [[CrossRef](#)]
24. Bokulich, N.A.; Mills, D.A. Improved Selection of Internal Transcribed Spacer-Specific Primers Enables Quantitative, Ultra-High-Throughput Profiling of Fungal Communities. *Appl. Environ. Microbiol.* **2013**, *79*, 2519–2526. [[CrossRef](#)]
25. Cobo-Díaz, J.F.; Baroncelli, R.; Le Floch, G.; Picot, A. A novel metabarcoding approach to investigate *Fusarium* species composition in soil and plant samples. *FEMS Microbiol. Ecol.* **2019**, *95*, fiz084. [[CrossRef](#)] [[PubMed](#)]
26. Jeanne, T.; D'astous-Pagé, J.; Hogue, R. Spatial, temporal and technical variability in the diversity of prokaryotes and fungi in agricultural soils. *Front. Soil Sci.* **2022**, *2*, 945888. [[CrossRef](#)]
27. Bolyen, E.; Rideout, J.R.; Dillon, M.R.; Bokulich, N.A.; Abnet, C.C.; Al-Ghalith, G.A.; Alexander, H.; Alm, E.J.; Arumugam, M.; Asnicar, F.; et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **2019**, *37*, 852–857. [[CrossRef](#)]
28. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **2011**, *17*, 10–12. [[CrossRef](#)]
29. Aurrecochea, C.; Barreto, A.; Brestelli, J.; Brunk, B.P.; Cade, S.; Doherty, R.; Fischer, S.; Gajria, B.; Gao, X.; Gingle, A.; et al. EuPathDB: The Eukaryotic Pathogen database. *Nucleic Acids Res.* **2012**, *41*, D684–D691. [[CrossRef](#)]
30. Chen, W.; Radford, D.R.; Hambleton, S. Towards Improved Detection and Identification of Rust Fungal Pathogens in Environmental Samples Using a Metabarcoding Approach. *Phytopathology* **2022**, *112*, 535–548. [[CrossRef](#)]
31. Grinevich, D.; Harden, L.; Grinevich, D.O.; Callahan, B.J. Serovar-level Identification of Bacterial Foodborne Pathogens from Full-length 16S rRNA Gene Sequencing. *Microbiology* **2023**, preprint. [[CrossRef](#)]
32. Boutigny, A.-L.; Gautier, A.; Basler, R.; Dauthieux, F.; Leite, S.; Valade, R.; Aguayo, J.; Ios, R.; Laval, V. Metabarcoding targeting the EF1 alpha region to assess *Fusarium* diversity on cereals. *PLoS ONE* **2019**, *14*, e0207988. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.