Sequence-ID and Genotype-ID: Novel open source tools for genotype and species identification

**Javier F. Tabima**

Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR

**Sydney E. Everhart**

Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR

**Meredith M. Larsen**

Horticultural Crops Research Laboratory, USDA ARS, Corvallis, OR

**Zhian N. Kamvar**

Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR

**Niklaus J. Grünwald**

Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR

Horticultural Crops Research Laboratory, USDA ARS, Corvallis, OR

Corresponding Author:

**Niklaus J. Grünwald**

Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR

Horticultural Crops Research Laboratory, USDA ARS, Corvallis, OR

**Abstract**

Identification of *Phytophthora* strains is critical yet challenging, especially when morphological traits are limited and only molecular data can identify certain species. We created Phytophthora-ID 2.0, a website that allows rapid identification of species and strains using previously described genetic markers. For identification, Phytophthora-ID 2.0 has two modules, Sequence-ID and Genotype-ID. Sequence-ID allows users to identify *Phytophthora* species using the carefully curated database of *Phytophthora* species identified with two commonly used barcodes, the internal transcribed spacer (ITS) and the cytochrome oxidase (COX) spacer regions. The Sequence-ID component has been improved from version 1 of the website to be faster and improve uptime. Genotype-ID is new in version 2 and implements the ability to place individuals into clonal lineages based on simple sequence repeat (SSR) identification for the two species *P. infestans* and *P. ramorum*. Genotype-ID uses a curated reference dataset, R functions from the R package ‘poppr’ running in shinyR server to display placement of query genotypes into currently known lineages using dendrograms or minimum spanning networks based on Bruvo’s genetic distance. All code is open source and available via github and CRAN providing the ability for other groups to develop web-based, interactive genotyping capabilities with minimal technical expertise on any server running shinyR and R.

**Introduction**

Researchers rely on a suite of tools, from morphological to molecular, to identify species, strains, and lineages within a species. Molecular markers can be used for looking at variations at different taxonomic levels. Regions as 16s have been used to classify at the level of domains, other LSU and SSU have been used to clarify relationships from the rank of phylum to the rank of genus. Other regions such as mtDNA and the ITS region were proposed as genes that were able to delineate ranks from classes to species-subspecies (Coleman, 2003, Coleman, 2002) to the level that these markers are the most widely used markers in plants (Coleman, 2003, Coleman, 2009), fungi (Turenne, 1999; Blaalidet al, 2013), corals (Grajales et at, 2007) and oomycetes (Cooke et al. 2000, Landis and Gargas 2007, Robideau et al. 2011, Grunwald et al, 2011). Recently, the search for molecular markers that are able to identify species and behave as taxonomic barcodes has flooded the literature, as the discovery of a region that is able to identify any species is most valuable in situations of unknown species, sampling using environmental samples and identification of causal agents of diseases.

Rapidly evolving markers can detect variation within species or even between populations. Microsatellite markers (simple sequence repeat; SSR) are one such marker. These are codominant markers that prime repeating sequences of 2 to 6 base pairs of DNA. Microsatellites have been widely used in population genetic studies, as they have advantages such as that they are locus-specific, codominant, highly polymorphic, and easily obtainable by PCR-based methodologies. With large-scale efforts to sequence genomes, such as the 1000 fungal genomes project, the ability to develop diagnostic markers is easier and many times have already been developed. In addition, PCR-based approaches and technology are accessible by scientists worldwide. Thus, the major barrier in identifying species, strains, and lineages, is access to a curated database that provides a standard for comparison.

Online curated databases exist for commonly used molecular markers, such as the ITS2 database. There are also general online databases for several organisms, such as FungiDB (Stajich et al, 2012) for fungal and fungal-like species, EuPathDB (Aurrecoechea C, et al., 2007) for eukaryotic organisms, and Phytophthora-DB (?) for species in the genus *Phytophthora*. For markers such as microsatellites, there are few online databases that are specific for model organisms or economically important organisms, such as the Human SSRD database (Subramanian et al, 2003), the cotton marker database (Blenda et al, 2006) and the SSR database for cereals and legumes (Jayashree et al, 2006). Unfortunately, these databases do not provide customizable elements beyond the identification of the strain/cultivar/population.

Labs routinely develop organism-specific databases and analyses that are shared with other researchers. However, simply sharing the database puts the burden of data analysis on researchers that may not have the proper background to perform the analyses or lack computational resources to perform these analyses. Our goal in this paper was to use openly available online resources to create a customized, organism-specific website database that provides a standardized approach to species and strain identification. We illustrate the capabilities of this approach to online database development with two notorious plant pathogens in the genus *Phytophthora.*

The genus *Phytophthora* is comprised of more than 110 species of oomycetes that are plant pathogens (Kroon et al, 2012). Formerly classified as fungi, oomycetes were moved to Kingdom Stramenopile. Commonly known as watermolds,

the genus comprises several plant pathogens that have caused devastating economical and social disasters such as the Irish potato famine in 1845. One reason to use the genus *Phytophthora* to exemplify our suite of tools is that, to identify species in the genus *Phytophthora*, sequence identification has been used as the most common mechanisms to identify and unknown strain, using barcodes such as the ITS region and COX spacer loci to identify the most similar species by a comparison with a curated dataset (Grunwald et al. 2011). Another reason to use the genus *Phytopthora* is the existence of clonal lineages in several of the species. A clonal lineage can be defined as a population derived from clonal reproduction (Lamour, 2013). Clonal lineages were discovered by use of AFLP (Flier et al, 2007) but have been recently studied using microsatellite data (Goss et al, 2006, Cooke et al, 2006). Still, no databases of well-curated genotypes are available, and the existence of such tools can facilitate the study of clonal lineages or any other populational differences.

**Methods**

**1. Development of tools for species and genotype identification: Phytophthora-ID 2.0**

The website for *Phytophthora* species and clonal lineage identification (Phytophthora-ID) is comprised by two modules: Sequence-ID and Genotype-ID. Sequence-ID is a module that permits species identification using sequence data while Genotype-ID determines genotype identification by using R packages that are specifically designed for population genetics, such as adegenet, pegas and poppr; using web-frameworks in html5 based on a web-interphase package (Shiny).

**1.1 Sequence-ID (Species identification using sequence similarity)**

For the Sequence-ID module, we modified a previous version of a blast search based on *Phytophthora* species (Grunwald et al. 2011) by adding new PERL functionality such as the communication between the user interphase (designed in html5) and the BLAST table of results for easy web visualization. Sequence-ID uses a PERL CGI script that permits the communication between the blast 2.2.7+ suite () and the user. In order to maintain updated data, the module permits the update of the database using a multifasta file into the *makeblastdb* program to generate the database and a simple web form to run the *blastn* to compare between the sequences available in the database.

1.2. **Genotype-ID (Genotype identification using molecular marker data using R).**

For Genotype\_ID, we developed a web application of R statistical language and specific population genetic packages. To develop the web app we used the package Shiny R (Rstudio Inc, 2013) which permits the communication of the web pages with the R structure. The Shiny web framework relies on reactive programming, which allows dynamic deployment of traditional R scripts in response to data input through a website console, with results generated and deployed to the web. Thus, ordinary R packages can be deployed with a website interface. Genotype-ID was built using this framework and mainly interacts with the R package ‘poppr’, which is a package for population genetics (Kamvar et al.). Web forms for query submission (sample genotypes) have a handling mechanism only accepts copy/paste inputs from Microsoft Excel so the researcher can use familiar tools and avoid reformatting matrices. Using the queries as input, poppr converts the input into R readable format and adds them into the curated dataset. The R package poppr then calculates Bruvo’s distance and recreates the relationships between the queries and the samples by constructing a dendogram using UPGMA or neighbor joining (NJ) with statistical support calculated by bootstrapping, and creates minimum spanning networks. These methods reconstruct relationships between queries and dataset samples in the web interface, thus allowing the user to visualize the relationship of their query submission in relation to the curated database.

2. **Testing the tools: The case of the genus Phytophthora.**

**2.1. Sequence-ID:**

We developed scripts that improve BLAST searches against two regularly updated sequence databases. The two databases include the ITS region sequences and the *cox1* and -*2* spacer region. Both data sets presented follow the requirements established in Phytophthora-ID 1.0 version, which are Sequences that are associated with a published *Phytophthora* species description, any of the *Phytophthora* phylogenetics references by Cooke et al. , Martin and Tooley , and Blair et al. 2008.

For we gathered a total of 211 sequences representing 108 species of the genus, while for we created a database of 150 sequences representing 106 specieshttp://phytophthora-id.org/files/Phytophthora-ID%20sequencing%20protocols.pdf.

The complete list of ITS sequences, cox spacer sequences, GenBank accession numbers, and *Phytophthora* spp. included can be downloaded at www.phytophthora-id.org or as supplements (include supplements).

**2.2. Genotype-ID**

We decided to use two of the most commonly knows species of *Phytophthora* worldwide: *P. infestans,* which is the pathogen that causes the potato leaf blight disease, and *P. ramorum*, which causes sudden oak death*. P. infestans* has more than 18 reported clonal lineages, while *P. ramorum* has 4 reported clonal lineages. For this case, samples of *P. infestans* and *P. ramorum* were obtained from around the world in order to have a wide representation of the clonal lineages in each of the species.

In the case of *P. ramorum*, 48 isolates representing the 4 reported clonal linages (NA1, NA2, EU1, EU2) were genotyped for 9 SSR loci reported for this species (Goss et al. 2009, Grunwald et al. 2009, Ivors et al. 2006, Prospero et al. 2007 and Vercauteren et al. 2010). For *P. infestans*, 11 clonal lineages (US11,US12, US8, US20, US21, US23, US24, EU4, EU5, EU8, EU13) represented in 96 isolates were genotyped using Lees et al. (2006) and Li et al. (2013) protocols for 11 SSR microsatellites.

Using the SSR data from *P. infestans* and *P. ramorum.* we reconstructed the databases using poppr, and recreated the dendrograms and minimum spanning networks based on Bruvo’s distance with queries similar to two of the clonal lineage of each species (for *P. infestans*, queries identical to US8 and US23, for *P. ramorum* queries identical to NA1 and NA2). For dendrograms, we used both neighbor joining and UPGMA, and added the support values after 1000 boostrap calculations. Minimum spanning networks were also reconstructed for both species.

**Results**

References:

**[FungiDB: an integrated functional genomics database for fungi](http://www.ncbi.nlm.nih.gov/pubmed/22064857)**. Stajich JE, Harris T, Brunk BP, Brestelli J, Fischer S, Harb OS, Kissinger JC, Li W, Nayak V, Pinney DF, Stoeckert CJ Jr, Roos DS. Nucleic Acids Res. 2012 Jan 1;40(D1):D675-D681.

[ApiDB: integrated resources for the apicomplexan bioinformatics resource center.](http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed&cmd=search&term=17098930) 2007 Jan;35(NAR Database issue):D427-30. Aurrecoechea C, et al. (to be updated soon)

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2448451/

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2448451/

<http://www.bioinfo.de/isb/2006060054/>

Flier, W. G., Kroon, L. P. N. M., Hermansen, A., van Raaij, H. M. G., Speiser, B., Tamm, L., Fuchs, J. G., Lambion, J., Razzaghian, J., Andrivon, D., Wilcockson, S., and Leifert, C. 2007. Genetic structure and pathogenicity of populations of Phytophthora infestans from organic potato crops in France, Norway, Switzerland and the United Kingdom. Plant Pathol. 56:562-572.

**Figure 1.** Scheme representing the key objects in the deployement of Genotype-ID. Rstudio shiny uses a User interphase file (can be using R code or html code) to genereate the user interphase scheme such as forms, views, lists and input windows. Shiny also uses a Server specification file. This file has all of the information for R to run on the background and use the user interphase file as inputs, function modifiers and menus. Inside of the server file, there must be a deployement of the database file (SSR tab delimited text) and the package to be used (in this case, poppr), as R shiny must process these basic objects before deploying the package. When all of the elements are taken by Shiny R and deployed, we have a functional version of Genotype-ID working. Now the user has to copy/paste the queries from a template excel spreadsheet (included on the webpage) and select the analysis to visualize, either the dendrogram reconstruction or the minimum spanning network.