4-28-2016

Microbe-ID: An open source toolbox for microbial genotyping and species identification

Javier F. Tabima
Oregon State University

Sydney E. Everhart
University of Nebraska-Lincoln, everhart@unl.edu

Meredith M. Larsen
USDA-ARS, Corvallis, OR

Alexaandra J. Weisberg
Oregon State University

Zhian N. Kamvar
Oregon State University, zkamvar@unl.edu

See next page for additional authors

Microbe-ID: An open source toolbox for microbial genotyping and species identification

Javier F. Tabima[1], Sydney E. Everhart[1,#], Meredith M. Larsen[2], Alexandra J. Weisberg[1], Zhian N. Kamvar[1], Mathew A. Tancos[3], Christine D. Smart[3], Jeff H. Chang[1,4], Niklaus J. Grünwald[1,2,3,4]

[1] Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR, USA

[2] Horticultural Crops Research Laboratory, USDA-ARS, Corvallis, OR, USA

[3] Plant Pathology and Plant-Microbe Biology Section, School of Integrative Plant Science, Cornell University, Geneva, NY, USA

[4] Molecular and Cellular Biology Graduate Program and Center for Genome Biology and Biocomputing, Oregon State University, Corvallis, OR, USA

[#] Current address: Department of Plant Pathology, University of Nebraska, Lincoln, NE, USA

Corresponding Author:

Niklaus J. Grünwald[1,2,3,4]

3420 NW Orchard Ave., Corvallis, OR 97330, United States

E-mail address: nik.grunwald@ars.usda.gov

Abstract

Development of tools to identify species, genotypes, or novel strains of invasive organisms is critical for monitoring emergence and implementing rapid response measures. Molecular markers, although critical to identifying species or genotypes, require bioinformatic tools for analysis. However, user-friendly analytical tools for fast identification are not readily available. To address this need, we created a web-based set of applications called Microbe-ID that allow for customizing a toolbox for rapid species identification and strain genotyping using any genetic markers of choice. Two components of Microbe-ID, named Sequence-ID and Genotype-ID, implement species and genotype identification, respectively. Sequence-ID allows identification of species by using BLAST to query sequences for any locus of interest against a custom reference sequence database. Genotype-ID allows placement of an unknown multilocus marker in either a minimum spanning network or dendrogram with bootstrap support from a user-created reference database. Microbe-ID can be used for identification of any organism based on nucleotide sequences or any molecular marker type and several examples are provided. We created a public website for demonstration purposes called Microbe-ID (www.microbe-id.org) and provided a working implementation for the genus *Phytophthora* (www.phytophthora-id.org). In *Phytophthora*-ID, the Sequence-ID application allows identification based on ITS or *cox* spacer sequences. Genotype-ID groups individuals into clonal lineages based on simple sequence repeat (SSR) markers for the two invasive plant pathogen species *P. infestans* and *P. ramorum*. All code is open source and available on github and CRAN. Instructions for installation and use are provided at https://github.com/grunwaldlab/Microbe-ID.

Background

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2005v1 | CC-BY 4.0 Open Access | rec: 28 Apr 2016, publ: 28 Apr 2016

Development of tools for identification of species, genotypes or strains is critical for monitoring emergence of invasive organisms such as *Phytophthora ramorum* causing sudden oak death (Grünwald et al. 2008), *Hymenoscyphus fraxineus* causing ash dieback (Gross et al. 2014), *Aphanomyces astaci* causing crayfish plague (Holdich et al. 2009), *Cryptococcus gattii* causing cryptococcosis and meningitis (Byrnes et al. 2010), or Methicillin-resistant *Staphylococcus aureus* causing invasive MRSA disease (Klevens et al. 2007). Molecular markers provide a rapid means for identification, but require various bioinformatics tools for identification of species and/or novel genotypes. In eukaryotes, sequences from the rRNA internal transcribed spacer (ITS) region and various mitochondrial DNA regions are used to delineate species (Coleman, 2003, 2007). ITS and mtDNA markers are now the most widely used markers in plants (Coleman, 2007, 2009), fungi (James et al., 2006), corals (Grajales et al., 2007), and oomycetes (Cooke et al., 2012; Robideau et al., 2011) and have been coined "DNA barcodes" because of their broad ability to distinguish species (Schoch et al., 2012). Classification of individuals using various molecular markers has recently increased. Multi-locus sequence types (MLST) are being widely used by researchers working with bacterial taxa to reveal the identity of samples by classification relative to known reference strains (Maiden et al., 2013). Other molecular markers or methods used to distinguish genotypes might include microsatellites (or simple sequence repeats) to identify strains and clonal lineages (Cooke et al., 2012; Ivors et al., 2006), DNA sequences for specific genic regions (Maiden et al., 2013), single nucleotide polymorphism (SNP) genotyping using reduced representation approaches (Grünwald et al. 2016) such as RAD-seq (Etter et al., 2011) or genotyping by sequencing (GBS, Elshire et al., 2011), or genome wide SNP genotyping (Huang et al., 2009).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2005v1 | CC-BY 4.0 Open Access | rec: 28 Apr 2016, publ: 28 Apr 2016

In addition to the molecular methods developed, different types of online databases have been implemented to identify species using these molecular methods within groups of organisms. Examples of these databases are FungiDB for fungi and fungal-like organisms (Stajich et al., 2011), EuPathDB for eukaryotic organisms (Aurrecoechea et al., 2013), and the *Phytophthora* database which allows entries by experts in the *Phytophthora* community from different labs or countries for different species of the genus (Park et al., 2008). We previously reported on our development of a database for *Phytophthora* species and genotype identification using web tools to identify species using common barcodes, enabling the conjunction of modern laboratory techniques with highly curated databases for species identification (Grünwald et al. 2011).

Our objective here was to report the development of a toolbox for microbe identification (Microbe-ID) that can readily be customized for sequence based species identification (Sequence-ID) or molecular marker-based identification of genotypes (Genotype-ID) for any group of organisms. Our objectives were two-fold: 1) to implement Microbe-ID as a demonstration site that is customizable for any group of organisms and 2) to demonstrate a working implementation at *Phytophthora*-ID.org version 2.0 with significant updates from version 1.0 (Grünwald et al. 2011). Microbe-ID includes two modules, Sequence-ID and Microbe-ID. Sequence analysis is implemented based on use of a well characterized barcode region for the genus *Phytophthora*, but can be implemented to use any barcode sequence of interest. Genotype analysis can be implemented to use a variety of marker data types. To demonstrate the breadth of the developed tools and applicability to the diversity of microorganisms, the following three examples were included: codominant microsatellite data (SSR/Microsatellite) for the oomycete *P. ramorum* (Grünwald et al. 2009), concatenated Multi Locus Sequence Type (MLST) or individual locus sequences for the bacterium *Clavibacter michiganensis* subsp. *michiganensis* (Tancos et al. 2015, Supplementary Figure 1), and dominant Amplified Fragment Length Polymorphism (Binary

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2005v1 | CC-BY 4.0 Open Access | rec: 28 Apr 2016, publ: 28 Apr 2016

(AFLP) data, Supplementary Figure 2) for the oomycete *Aphanomyces euteiches* (Grünwald and Hoheisel, 2006). Moreover, Genotype-ID can be expanded to include other marker systems including gene sequences for resistance to antibiotics or fungicides as well as presence/absence polymorphisms for effector genes or other adaptive loci. Two sequence databases were developed that help us demonstrate the utility of Microbe-ID. These databases are for the genus *Phytophthora* (*Phytophthora*-ID) with two sequence databases containing over 110 species that are mostly plant pathogens (Kroon et al., 2012), and two genotyping databases for populations of the potato late blight pathogen, *P. infestans*, and the sudden oak death pathogen, *P. ramorum* (Genotype-ID) (Grünwald et al. 2008; Kamoun et al. 2015). All of these tools are readily customizable and open source (http://www.github.com/grunwaldlab/microbe-ID), provided as a demonstration site (http://microbe-id.org/), and a working implementation of Sequence-ID and Genotype-ID that we use in our own work for the genus *Phytophthora* (www.phytophthora-id.org). Finally, a companion paper describes application of Microbe-ID for the implementation of a new website, Gall-ID, with novel tools for identification of gall-forming bacteria (Davis et al. 2016).

Microbe-ID: a toolkit for web-based genotype and species identification

The Microbe-ID toolbox. We developed a template website named Microbe-ID (http://microbe-id.org/) with two separate modules, Sequence-ID and Genotype-ID, for sequence-based species identification and genotyping, respectively. The website is written using bootstrap (http://getbootstrap.com/), a HTML, CSS, and JS framework for developing responsive, mobile first projects on the web. Specific instructions, code, and resources necessary for implementation of Microbe-ID are provided on the github repository (https://github.com/grunwaldlab/Microbe-

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2005v1 | CC-BY 4.0 Open Access | rec: 28 Apr 2016, publ: 28 Apr 2016

ID). The server currently hosting Microbe-ID is running Centos Linux release 6.6, NCBI BLAST 2.2.28+ (Altschul et al., 1990), MAFFT version 7.221 (Katoh et al. 2002), and R version 3.1.2. Below we describe specific tools required for customization of each module.

Navigating and using Microbe-ID. Navigating the demonstration site of Microbe-ID (http://Microbe-ID.org) the user will encounter a menu bar with links to Sequence-ID, Genotype-ID, and an about page. The home page has a general description of the functionality and components implemented in Microbe-ID, as well as links to the github site. Since input from the user is entered into forms as a query, each form implemented in Microbe-ID is encoded to check that the format of the data supplied by the user is supported. If the format is incorrect, the page will prompt the user with an error message, making the use of Microbe-ID more user-friendly.

Sequence-ID. We created a module called Sequence-ID that uses BLAST analysis of common sequence loci for species identification (Figure 1). Sequence-ID includes a PERL_CGI script that permits the communication between a user interface form (implemented in HTML5) and a BLAST database for the desired sequence data. The form recognizes the input in FASTA format and uses BLASTN to search for the most similar DNA sequences in the marker database. The PERL-CGI script receives the information of the HTML tabulated output from BLAST and displays a table of hits to the end user. Sequence-ID is customizable as the FASTA data file can readily be updated using the makeblastdb program of the BLAST suite. It can also similarly be implemented for BLASTP analysis of amino acid sequences.

In the Sequence-ID webpage, the user will find two main tabs: A “Blast” tab in which the web-app is contained, and a “Help” tab which contains a link to a “Site Help” prompt. This “Site Help” prompt shows an example of the FASTA sequence format recognized by Sequence-ID. The user can copy the FASTA sequence and paste it in the “Blast” tab text form to perform a BLAST search on the example query, or can download the database in FASTA format. After the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2005v1 | CC-BY 4.0 Open Access | rec: 28 Apr 2016, publ: 28 Apr 2016

BLAST is performed, the web-page will provide a table including the BLAST search results for the query of interest using the BLAST alignment format.

Genotype-ID. Genotype-ID is a web application designed to be user-friendly, and to facilitate the interaction with R using the set of tools listed in Table 1 (Figure 2). To develop the web application, we used the R package 'shiny' (Rstudio, 2013). Shiny facilitates the interactions between user, server, and R (Figure 2). 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 whereby results generated in R are subsequently pushed to the end user. Thus, ordinary R packages, which otherwise require familiarity with the language, can be deployed behind a user-friendly interface. Genotype-ID interacts principally with the R packages 'poppr' (Kamvar et al., 2014), `adegenet` (Jombart, 2008), `ape` (Paradis et al, 2004), and `pegas` (Paradis, 2010) (Table 1).

Genotype-ID has three different modules, which are shown as tabs on the website, each specific to different molecular markers: SSR/microsatellite data (SSR/Microsatellite data), multilocus genotype sequence types (MLST data), and AFLP, RFLP, SNP and other binary datasets (Binary (AFLP) data). The user supplies a query via the web framework that is read into R and queried against a curated dataset. Genetic distances (Table 2) are calculated between the query and genotypes in the reference database, with relationship of the query to the database presented either as a UPGMA or neighbor-joining dendrogram with bootstrap support or a minimum spanning network. These methods reconstruct relationships of the query relative to the genotypes found in the curated database. Each module has its own customization scheme to analyze each particular molecular marker type. The SSR data tool calculates Bruvo's distance (Bruvo et al. 2004) to reconstruct the UPGMA or NJ dendrogram and minimum spanning networks. The MLST data tool permits the comparison of user-submitted gene sequences in

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2005v1 | CC-BY 4.0 Open Access | rec: 28 Apr 2016, publ: 28 Apr 2016

FASTA format. The MLST data tool uses MAFFT (Katoh et al. 2002) to align each query sequence to the corresponding curated gene database. It then concatenates all separate alignments, calculates the genetic distance for the alignment, and reconstructs distance dendrograms and minimum spanning networks. Lastly, the binary tool uses molecular markers such as AFLPs or RFLPs to reconstruct relationships. This tool uses binary data (coded as 1 and 0) and different genetic distances (Table 2) to reconstruct the UPGMA or NJ dendrogram and the minimum spanning network.

The Microbe-ID implementation. Microbe-ID contains implementations of each of the modules for different custom databases: the SSR implementation uses nine diploid SSR loci for the oomycete species *P. ramorum* (Figure 3); the MLST implementation uses eight multilocus sequence types for the bacterial species *Clavibacter michiganensis* subsp. *michiganensis*; and a binary implementation that uses 56 loci for AFLP data for the oomycete *Aphanomyces euteiches*. Each of the implementations of Genotype-ID contains collapsible instructions on how to format the query and a link to download a tabulated file with example queries formatted for use in Genotype-ID. The user can download the spreadsheet in order to edit, copy, and paste their custom queries into the "Data input" form. To make analyses that use a random seed repeatable, the user has the option to specify a seed number. In MLST-ID and Binary-ID, the user can select the genetic distance to be used in the analysis. After the user inputs the query and runs the web application, the web page will proceed to the *Analysis* section, where the user can choose between two different visualizations, a distance tree with bootstrap support values or a minimum spanning network. If the distance tree is selected, the user can change the tree algorithm (either neighbor joining or UPGMA) and number of bootstrap replicates. The user can also download results as a PDF or in NEWICK format. For the minimum spanning network, the user can adjust the grey

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2005v1 | CC-BY 4.0 Open Access | rec: 28 Apr 2016, publ: 28 Apr 2016

scale for edge distances and download results as a PDF file. Implementations for other genetic data and different visualizations can be added.

***Phytophthora*-ID implementation.** Tools provided in Microbe-ID were implemented in http://Phytophthora-ID.org, a functional website for identifying samples from the genus *Phytophthora*, a group of economically important plant pathogens in the stramenopile branch of the tree of life (Kamoun et al. 2015). Phytophthora-ID version 1.0 (Grünwald et al. 2011) had a BLAST script to identify samples using the ITS barcode. *Phytophthora*-ID version 2.0 was substantially revised and upgraded from the first iteration and now includes a faster BLAST search implemented in Sequence-ID and a new genotype identification system implemented in Genotype-ID.

The current version of Phytophthora-ID contains a Sequence-ID module customized for two molecular barcodes used for *Phytophthora* species identification (ITS and *cox* spacer). This particular version of the sequence identification tool was created in PERL-CGI, and permits the search of a FASTA sequence query against a curated database of *Phytophthora* species. The PERL-CGI for *Phytophthora*-ID version 2.0 was also redesigned to run directly on the server to make the web application more stable and faster. In contrast, in *Phytophthora*-ID 1.0, the web application was designed as a communication wrapper to an external cluster, thus rendering functionality dependent on the external server.

We implemented the Sequence-ID module to use sequences from two genetic regions: the nuclear internal transcribed spacer (ITS) and mitochondrial *cox* spacer region spanning the *cox1* and *cox2* loci. Databases were created using sequences from published *Phytophthora* species descriptions (Blair et al., 2008; Martin and Tooley, 2003) or those that belong to classical *Phytophthora* species described at least 25 years ago and are still recognized as valid (Erwin and Ribeiro, 1996).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2005v1 | CC-BY 4.0 Open Access | rec: 28 Apr 2016, publ: 28 Apr 2016

For the ITS region, we gathered a total of 211 sequences representing 108 species. For the *cox* spacer region, we created a database of 150 sequences representing 106 species. Laboratory protocols for preparing samples for sequence analysis are available on the website and in a previous publication (Grünwald et al., 2011). Additionally, a file with the complete set of ITS sequences, *cox* spacer sequences, GenBank accession numbers, and *Phytophthora* spp. can be downloaded from the website. Documentation on updates to the databases is provided on the website.

We implemented two modules for genotype identification of two *Phytophthora* species: *P. infestans* that causes potato late blight, and *P. ramorum* that causes sudden oak death. *P. infestans* has more than 18 reported clonal lineages (Hu et al., 2012), while *P. ramorum* has 4 reported clonal lineages (Grünwald et al., 2009; Van Poucke et al., 2012). To establish a database with a wide representation of clonal lineages in each species, we obtained and prepared DNA samples from *P. infestans* and *P. ramorum* that were collected from various regions of the world. A total of 48 *P. ramorum* isolates representing the 4 reported clonal linages (NA1, NA2, EU1, EU2) were genotyped at nine SSR loci (Cooke et al., 2012; Prospero et al., 2007; Vercauteren et al., 2010; Grünwald et al., 2009; Ivors et al., 2006). Similarly, 11 *P. infestans* clonal lineages that are dominant in the US (including US11, US12, US8, US20, US21, US23, US24, EU4, EU5, EU8, EU13) represented by a total population of 96 isolates were genotyped at 11 SSR loci, using protocols from Lees et al. (2006) and Li et al. (2013). We constructed SSR reference databases compatible with the poppr/adegenet R packages that provide dendrograms and minimum spanning networks for queries. Bootstrap support can be calculated for dendrograms. To test the datasets, we used two queries per dataset, one sample of NA1 and NA2 clonal lineages each for *P. ramorum* and one sample of clonal lineages US-23 and US-8 clonal lineages for *P. infestans*.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2005v1 | CC-BY 4.0 Open Access | rec: 28 Apr 2016, publ: 28 Apr 2016

All samples grouped with the corresponding reference lineages with high support values (Figures 4 and 5), demonstrating functionality of the Microbe-ID tool to identify samples correctly.

Conclusions

We constructed a web framework that can use a wide array of molecular markers to rapidly determine identities of species and genotypes. This web framework is provided as open source code on github (https://github.com/grunwaldlab/Microbe-ID). We implemented a demonstration available at http://Microbe-ID.org. New implementations for any organisms require reference databases, scripts, and web pages including choices from the set of computational tools shown in Table 1. The Microbe-ID framework is currently implemented for two fully functional sites, *Phytophthora*-ID (http://Phytophthora-ID.org/) and Gall-ID (http://Gall-ID.cgrb.oregonstate.edu/; Davis et al. 2016).

Modularity of Microbe-ID permits implementation of a range of markers using curated databases of known species or genotypes to determine identity of specimens. Use of a custom BLAST database permits identification of any sample of a species of interest by using any sequence-based molecular marker. Note that the amount of time required for BLAST execution is dependent on size of the databases. For the Sequence-ID ITS region implementation for *Phytophthora*, each run was completed and loaded to the web-app in less than one second, with results shown as a BLAST output table.

Most innovative in this web framework is the Genotype-ID module, which permits the researcher to use, in a simple, web-based interface, any type of molecular marker with a corresponding and curated reference dataset. Use of a shiny server for strain identification is completely novel and modularity permits use of virtually any molecular marker suitable for

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2005v1 | CC-BY 4.0 Open Access | rec: 28 Apr 2016, publ: 28 Apr 2016

existing R packages. For genotyping, population genetic markers demonstrated include SSR, MLST, and AFLP/RFLP data, but can also be expanded to presence/absence of genes or alleles or any other genetic features of interest. Custom implementations must include a reference dataset and custom R scripts. Implementations can be customized for virtually any organism and any molecular marker that can be analyzed in R.

Development of tools for species identification has increased in recent years; but lack of modularity and complicated methodological procedures make the process of species identification tedious and challenging for researchers without sufficient skills in computational biology. We developed an easy-to-set-up web framework, which permits quick and flexible deployment of species identification and genotyping tools using sequence, microsatellite/SSR, AFLP/RFLP, or MLST data for any organism. Given use of the bootstrap html framework, webpages are light in terms of computer resources required and are thus usable on any device including smartphones and tablets, and function with any browser. These tools have been deployed and used successfully for the genus *Phytophthora* and gall-forming bacteria, providing examples of two curated websites (Grünwald et al. 2011; Davis et al. 2016).

Funding

This research is supported in part by US Department of Agriculture (USDA) Agricultural Research Service Grant 5358-22000-039-00D (NJG), USDA National Institute of Food and Agriculture (NIFA) Grant 2011-68004-30154 (NJG), the USDA ARS Floriculture Nursery Research Initiative (NJG), USDA NIFA Grant 2014-51181-22384 (JHC and NJG), and USDA NIFA 2012-67012-19844 (SEE).

References

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2005v1 | CC-BY 4.0 Open Access | rec: 28 Apr 2016, publ: 28 Apr 2016

Aurrecoechea, C., Barreto, A., Brestelli, J., Brunk, B.P., Cade, S., Doherty, R., Fischer, S., Gajria, B., Gao, X., Gingle, A. & Grant, G. (2012). EuPathDB: The eukaryotic pathogen database. Nucleic Acids Research 41, D684-D691.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). "Basic local alignment search tool." Journal of Molecular Biology 215,403-410.

Blair, J.E., Coffey, M.D., Park, S.-Y., Geiser, D.M., & Kang, S. (2008). A multi-locus phylogeny for *Phytophthora* utilizing markers derived from complete genome sequences. Fungal Genetics and Biology 45, 266-277.

Bruvo, R., Michiels, N.K., D'Souza, T.G., & Schulenburg, H. (2004). A simple method for the calculation of microsatellite genotype distances irrespective of ploidy level. Molecular Ecology, 13, 2101-2106.

Byrnes III, E.J., Li, W., Lewit, Y., Ma, H., Voelz, K., Ren, P., Carter, D.A., Chaturvedi, V., Bildfell, R.J., May, R.C. and Heitman, J. (2010). Emergence and pathogenicity of highly virulent *Cryptococcus gattii* genotypes in the northwest United States. PLoS Pathogens *6*(4), p.e1000850.

Coleman, A.W. (2003). ITS2 is a double-edged tool for eukaryote evolutionary comparisons. Trends in Genetics 19, 370-355.

Coleman, A.W. (2007). Pan-eukaryote ITS2 homologies revealed by RNA secondary structure. Nucleic Acids Research 35, 3322-3329.

Coleman, A.W. (2009). Is there a molecular key to the level of "biological species" in eukaryotes? A DNA guide. Molecular Phylogenetics and Evolution 50, 197-203.

Cooke, D.E., Cano, L.M., Raffaele, S., Bain, R.A., Cooke, L.R., Etherington, G.J., Deahl, K.L., Farrer, R.A., Gilroy, E.M., Goss, E.M. & Grünwald, N.J. (2012). Genome analyses of an

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2005v1 | CC-BY 4.0 Open Access | rec: 28 Apr 2016, publ: 28 Apr 2016

304 aggressive and invasive lineage of the Irish potato famine pathogen. PLoS Pathogens 8,
305 e1002940.
306 Davis, E.W., Weisberg, A.J., Tabima, J.F., Grünwald, N.J., & Chang, J.H. (2016). Gall-ID: Tools
307 for genotyping gall-causing phytopathogenic bacteria. PeerJ Preprints 4, e1998v1. DOI:
308 10.7287/peerj.preprints.1998v1.
309 Edwards, A.W.F. (1971). Distances between populations on the basis of gene frequencies.
310 Biometrics 27, 873-881.
311 Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., & Mitchell, S.E.
312 (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity
313 species. PLoS ONE 6, e19379.
314 Erwin, D.C. and Ribeiro, O.K., (1996). Phytophthora diseases worldwide. American
315 Phytopathological Society, St. Paul, MN (USA).
316 Etter, P.D., Bassham, S., Hohenlohe, P.A., Johnson, E.A. and Cresko, W.A. (2010). SNP
317 discovery and genotyping for evolutionary genetics using RAD sequencing. In Methods in
318 Molecular Biology (Clifton, NJ), 772, 157-178.
319 Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood
320 approach. Journal of Molecular Evolution 17, 368-376.
321 Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5, 164-
322 166.
323 Grajales, A., Aguilar, C., & Sánchez, J.A. (2007). Phylogenetic reconstruction using secondary
324 structures of Internal Transcribed Spacer 2 (ITS2, rDNA): Finding the molecular and
325 morphological gap in Caribbean gorgonian corals. BMC Evolutionary Biology 7, 90.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2005v1 | CC-BY 4.0 Open Access | rec: 28 Apr 2016, publ: 28 Apr 2016

326 Grigoriev, I.V., Nikitin, R., Haridas, S., Kuo, A., Ohm, R., Otillar, R., Riley, R., Salamov, A.,
327 Zhao, X., Korzeniewski, F. & Smirnova, T. (2013). MycoCosm portal: Gearing up for
328 1000 fungal genomes. Nucleic Acids Research 42, D699-D704.
329 Gross A, Holdenrieder O, Pautasso M, Queloz V, & Sieber TN. (2014). *Hymenoscyphus*
330 *pseudoalbidus*, the causal agent of European ash dieback. Molecular Plant Pathology 1, 5-
331 21.
332 Grünwald, N.J., Goss, E.M., Ivors, K., Garbelotto, M., Martin, F.N., Prospero, S., Hansen, E.,
333 Bonants, P.J., Hamelin, R.C., Chastagner, G. & Werres, S. (2009). Standardizing the
334 nomenclature for clonal lineages of the sudden oak death pathogen, *Phytophthora*
335 *ramorum*. Phytopathology 99, 792-795.
336 Grünwald, N.J., Martin, F.N., Larsen, M.M., Sullivan, C.M., Press, C.M., Coffey, M.D., Hansen,
337 E.M., & Parke, J.L. (2011). *Phytophthora*-ID.org: A sequence based *Phytophthora*
338 identification tool. Plant Disease 95, 337-342.
339 Grünwald, N. J., McDonald, B. M., & Milgroom, M. G. (2016). Population genomics of fungal
340 and oomycete pathogens. Annual Review of Phytopathology, in press.
341 Grünwald, N.J., Goss, E.M. & Press, C.M. (2008). *Phytophthora ramorum*: A pathogen with a
342 remarkably wide host-range causing sudden oak death on oaks and ramorum blight on
343 woody ornamentals. Molecular Plant Pathology 9, 729-740.
344 Grünwald, N.J., & Hoheisel, G.-A. (2006). Hierarchical analysis of diversity, selfing and genetic
345 differentiation in populations of the oomycete *Aphanomyces euteiches*. Phytopathology
346 96, 1134-1141.
347 Holdich, D.M., Reynolds, J.D., Souty-Grosset, C. & Sibley, P.J. (2009). A review of the ever
348 increasing threat to European crayfish from non-indigenous crayfish species. Knowledge
349 and Management of Aquatic Ecosystems 11, 394-395.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2005v1 | CC-BY 4.0 Open Access | rec: 28 Apr 2016, publ: 28 Apr 2016

350 Hu, C.H., Perez, F.G., Donahoo, R., McLeod, A., Myers, K., Ivors, K., Secor, G., Roberts, P.,
351 Deahl, K.L., Fry, W.E. & Ristaino, J.B. (2012). Recent genotypes of *Phytophthora*
352 *infestans* in the eastern United States reveal clonal populations and reappearance of
353 mefenoxam sensitivity. Plant Disease 96, 1323-1330.
354 Huang, X., Feng, Q., Qian, Q., Zhao, Q., Wang, L., Wang, A., Guan, J., Fan, D., Weng, Q.,
355 Huang, T. and Dong, G. (2009). High-throughput genotyping by whole-genome
356 resequencing. Genome Research 19, 1068-1076.
357 Ivors, K., Garbelotto, M., Vries, I.D.E., Ruyter-Spira, C., Te Hekkert, B., Rosenzweig, N., &
358 Bonants, P. (2006). Microsatellite markers identify three lineages of *Phytophthora*
359 *ramorum* in US nurseries, yet single lineages in US forest and European nursery
360 populations. Molecular Ecology 15, 1493-1505.
361 James, T.Y., Kauff, F., Schoch, C.L., Matheny, P.B., Hofstetter, V., Cox, C.J., Celio, G.,
362 Gueidan, C., Fraker, E., Miadlikowska, J. & Lumbsch, H.T. (2006). Reconstructing the
363 early evolution of fungi using a six-gene phylogeny. Nature 443, 818-822.
364 Jombart, T. (2008). Adegenet: a R package for the multivariate analysis of genetic markers.
365 Bioinformatics 24, 1403-1405.
366 Jukes, T.H. and Cantor, C.R. (1969). Evolution of protein molecules: Mammalian Protein
367 Metabolism 3, 132.
368 Kamoun, S., Furzer, O., Jones, J.D.G., Judelson, H.S., Shad Ali, G., Dalio, R.J.D., Guha Roy, S.,
369 Schena, L., Zampounis, A., Panabières, F., Cahill, D., Ruocco, M., Figueiredo, A., Chen,
370 X.-R., Hulvey, J., Stam, R., Lamour, K., Gijzen, M., Tyler, B.N., Grünwald, N. J., Tor,
371 M., Mukhtar, S.M., Tome, D., van den Ackerveken, G., McDowell, J., Daayf, F., Fry,
372 W.E., Lindqvist-Kreuze, H., Meijer, H.J.G., Petre, B., Ristaino, J., Yoshida, K., Birch, P.,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2005v1 | CC-BY 4.0 Open Access | rec: 28 Apr 2016, publ: 28 Apr 2016

& Govers, F. (2015). The top 10 oomycete pathogens in molecular plant pathology. Molecular Plant Pathology 16, 413-434.

Kamvar, Z.N., Tabima, J.F., & Grünwald, N.J. (2014). Poppr: An R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ 2, e281.

Kamvar, Z.N., Brooks, J.C., & Grünwald, N.J. (2015). Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. Frontiers in Genetics 6:208.

Katoh, K., Misawa, K., Kuma, K. I., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30, 3059-3066.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution. 16, 111-120.

Klevens RM, Morrison MA, Nadle J, Petit S, Gershman K, Ray S, Harrison LH, Lynfield R, Dumyati G, Townes JM, & Craig AS. (2007). Invasive methicillin-resistant *Staphylococcus aureus* infections in the United States. Jama 298, 1763-71.

Kroon, L.P., Brouwer, H., de Cock, A.W. & Govers, F. (2012). The genus *Phytophthora* anno 2012. Phytopathology 102, 348-364.

Lees, A.K., Wattier, R., Shaw, D.S., Sullivan, L., Williams, N.A., & Cooke, D.E.L. (2006). Novel microsatellite markers for the analysis of *Phytophthora infestans* populations. Plant Pathology 55, 311-319.

Li, Y., Cooke, D.E.L., Jacobsen, E., & van der Lee, T. (2013). Efficient multiplex simple sequence repeat genotyping of the oomycete plant pathogen *Phytophthora infestans*. Journal of Microbiological Methods 92, 316-322.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2005v1 | CC-BY 4.0 Open Access | rec: 28 Apr 2016, publ: 28 Apr 2016

396 Maiden, M.C.J., van Rensburg, M.J.J., Bray, J.E., Earle, S.G., Ford, S.A., Jolley, K.A., &
397 McCarthy, N.D. (2013). MLST revisited: the gene-by-gene approach to bacterial
398 genomics. Nature Reviews Microbiology 11, 728-736.
399 Martin, F.N., & Tooley, P.W. (2003). Phylogenetic relationships among *Phytophthora* species
400 inferred from sequence analysis of mitochondrially encoded cytochrome oxidase I and II
401 genes. Mycologia 95, 269-284.
402 Nei, M. (1972). Genetic distance between populations. The American Naturalist 106, 283-292.
403 Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in
404 R language. Bioinformatics 20, 289-290.
405 Paradis E. (2010). Pegas: An R package for population genetics with an integrated-modular
406 approach. Bioinformatics 26, 419-420
407 Park, J., Park, B., Veeraraghavan, N., Jung, K., Lee, Y.-H., Blair, J. E., Geiser, D. M., Isard, S.,
408 Mansfield, M. A., Nikolaeva, E., Park, S.-Y., Russo, J., Kim, S. H., Greene, M., Ivors, K.
409 L., Balci, Y., Peiman, M., Erwin, D. C., Coffey, M. D., Rossman, A., Farr, D., Cline, E.,
410 Grünwald, N. J., Luster, D. G., Schrandt, J., Martin, F., Ribeiro, O. K., Makalowska, I.,
411 and Kang, S. (2008). Phytophthora Database: A forensic database supporting the
412 identification and monitoring of *Phytophthora*. Plant Disease 92, 966-972.
413 Prevosti, A., Ocaña, J., & Alonso, G. (1975). Distances between populations of *Drosophila*
414 *subobscura*, based on chromosome arrangement frequencies. Theoretical and Applied
415 Genetics 45, 231-241.
416 Prospero, S., Hansen, E.M., Grünwald, N.J., & Winton, L.M. (2007). Population dynamics of the
417 sudden oak death pathogen *Phytophthora ramorum* in Oregon from 2001 to 2004.
418 Molecular Ecology 16, 2958-2973.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2005v1 | CC-BY 4.0 Open Access | rec: 28 Apr 2016, publ: 28 Apr 2016

419 Reynolds, J., Weir, B.S., & Cockerham, C.C. (1983). Estimation of the coancestry coefficient:
420 basis for a short-term genetic distance. Genetics 105, 767-779.
421 Robideau, G.P., de Cock, A.W., Coffey, M.D., Voglmayr, H., Brouwer, H., Bala, K., Chitty,
422 D.W., Désaulniers, N., Eggertson, Q.A. & Gachon, C.M.M. (2011). DNA barcoding of
423 oomycetes with cytochrome c oxidase subunit I and internal transcribed spacer. Molecular
424 Ecology Resources 11, 1002-1011.
425 Rogers, J.S. (1972). Measures of genetic similarity and genetic distance. Studies in Genetics 7,
426 145-153.
427 Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., & Chen, W.
428 (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA
429 barcode marker for Fungi. Proceedings of the National Academy of Sciences of the
430 United States of America 109, 6241-6246.
431 Stajich, J.E., Harris, T., Brunk, B.P., Brestelli, J., Fischer, S., Harb, O.S., Kissinger, J.C., Li, W.,
432 Nayak, V., Pinney, D.F. & Stoeckert, C.J., (2011). FungiDB: An integrated functional
433 genomics database for fungi. Nucleic Acids Research 40, D675-81.
434 Tamura, K., & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control
435 region of mitochondrial DNA in humans and chimpanzees. Molecular Biology and
436 Evolution 10, 512-526.
437 Tancos, M.A., Lange, H.W., & Smart, C.D. (2015). Characterizing the genetic diversity of the
438 *Clavibacter michiganensis* subsp. *michiganensis* population in New York.
439 Phytopathology 105, 169-179.
440 Van Poucke, K., Franceschini, S., Webber, J.F., Vercauteren, A., Turner, J.A., McCracken, A.R.,
441 Heungens, K., & Brasier, C.M. (2012). Discovery of a fourth evolutionary lineage of
442 *Phytophthora ramorum*: EU2. Fungal Biology 116, 1178-1191.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2005v1 | CC-BY 4.0 Open Access | rec: 28 Apr 2016, publ: 28 Apr 2016

443 Vercauteren, A., De Dobbelaere, I., Grünwald, N.J., Bonants, P., Van Bockstaele, E., Maes, M.,
444 & Heungens, K. (2010). Clonal expansion of the Belgian *Phytophthora ramorum*
445 populations based on new microsatellite markers. Molecular Ecology 19, 92-107.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2005v1 | CC-BY 4.0 Open Access | rec: 28 Apr 2016, publ: 28 Apr 2016

446 **Tables**

447

448 **Table 1.** Open source computational tools required to install and deploy Microbe-ID on a server.

449 - = reference not available; see website provided for information.

Tool	Description	Source	Reference
Ape	R package for phylogenetic and evolutionary analysis	http://ape-package.ird.fr/	Paradis et al., 2004
BLAST	Basic Local Alignment and Search Tool implemented as an algorithm for comparing DNA, RNA or protein query sequences against a reference database	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/	Altschul et al., 1990
Adegenet	R package for multivariate analysis of genetic data	https://github.com/thibautjombart/adegenet/	Jombart, 2008
Bootstrap	A framework for developing responsive, mobile-first projects on the web	http://www.getbootstrap.com/	-
Microbe-ID	Set of web-apps for identification of species, genotypes, and strains of any organism	https://github.com/grunwaldlab/Microbe-ID	This paper.
MAFFT	Multiple sequence alignment algorithm to find homology between sequences using Fourier algorithms	http://mafft.cbrc.jp/alignment/software/	Katoh et al., 2002
Pegas	R package for analysis of population genetic data	https://github.com/grunwaldlab/poppr	Kamvar et al., 2014; Kamvar et al., 2015
Poppr	R package for genetic analysis of populations with mixed reproduction	https://github.com/grunwaldlab/poppr	Kamvar et al., 2014
Shiny	Interactive web application framework for R	http://shiny.rstudio.com/	-

450

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2005v1 | CC-BY 4.0 Open Access | rec: 28 Apr 2016, publ: 28 Apr 2016

451 **Table 2.** Genetic distances implemented in Microbe-ID.
452

Distance model	Module	R package	References
Felsenstein 81 (F81)	MLST-ID	ape	Felsenstein (1981)
Felsenstein 84 (F84)	MLST-ID	ape	Felsenstein (1989)
Indel	MLST-ID	ape	Paradis et al. (2004)
Jukes-Cantor (JC69)	MLST-ID	ape	Jukes and Cantor (1969)
Kimura 80 (K80)	MLST-ID	ape	Kimura (1980)
Kimura 81 (K81)	MLST-ID	ape	Kimura (1980)
Raw	MLST-ID	ape	Paradis et al. (2004)
Tamura and Nei 93 (TN93)	MLST-ID	ape	Tamura and Nei (1993)
Transitions (TS)	MLST-ID	ape	Paradis et al. (2004)
Transversions (TV)	MLST-ID	ape	Paradis et al. (2004)
Bruvo	SSR-ID	poppr	Bruvo et al. (2004)
Edwards	Binary-ID	poppr/adegenet	Edwards (1971)
Nei	Binary-ID	poppr/adegenet	Nei (1972)
Prevosti	Binary-ID	poppr/adegenet	Prevosti et al. (1974)
Reynolds	Binary-ID	poppr/adegenet	Reynolds et al. (1983)
Rogers	Binary-ID	poppr/adegenet	Rogers (1972)

453

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2005v1 | CC-BY 4.0 Open Access | rec: 28 Apr 2016, publ: 28 Apr 2016

454 **Figure Legends**

455

456 **Figure 1.** Screen capture of the Sequence-ID user interface of Microbe-ID. The interface
457 includes tabs for input of barcodes and a dropdown help link that contains laboratory protocols,
458 examples data, and logs of each query dataset. The input frame has a control scheme that only
459 permits data in FASTA format.

460

461 **Figure 2.** Diagram representing implementation of Genotype-ID, which is comprised of a user
462 interface file (index.html) and a server file (server.R). Each file communicates with the R
463 framework (via shiny) and user (via HTML5). On the user side (left side), user input is provided
464 by copy/paste of a query and selects/specifies the desired application modifiers (seed number,
465 genetic distance calculation). This information is subsequently received and processed by the
466 server file, prompting the application to run in R. On the server side (right side) a database file
467 (Marker DB), R packages, and functions are retrieved and executed. When the run is complete,
468 the server file provides output to the user interface file and displayed on the app output.

469

470 **Figure 3.** Screen capture of the Genotype-ID user interface of Microbe-ID. Shown is the SSR-ID
471 module for *P. ramorum*. Genotype-ID includes an *Intstructions* link with an example Excel file
472 that the user can modify, copy, and paste into the data input form. There is no limit to the number
473 of queries submitted. In the *Analysis* section, tabs are provided to select either between a distance
474 tree or a minimum spanning network rendering.

475

476 **Figure 4.** Results of SSR-ID for NA1 and NA2 queries of *P. ramorum* provided in the example
477 data file. Each color represents a clonal lineage pre-assigned to each reference sample (NA1,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2005v1 | CC-BY 4.0 Open Access | rec: 28 Apr 2016, publ: 28 Apr 2016

NA2, EU1, EU2) with queries colored in red. **A.** UPGMA tree with 1,000 bootstrap replicates and support values above branches. Queries are represented in red and all are correctly placed with reference samples of the presumptive clonal lineage while also representing the relationship between clonal lineages in the reference dataset. **B.** Minimum spanning network reconstruction. Edge shade and width are inversely proportional to Bruvo's distance as shown in the horizontal scale bar. Queries are represented in red and placed in nodes with the most similar reference sample in the dataset, indicating the NA1 query is most similar to the PR-12-044 reference sample and the NA2 query is more closely related to the PR-05-156 and PR-12-103 samples, which also belong to the NA2 clonal lineage.

Figure 5. Results of SSR-ID queries for strains placed into the US8 and US23 clonal lineages of the potato late blight pathogen, *P. infestans*. Colors correspond to clonal lineages assigned to each reference sample (B, C, EU-13, EU-14, etc.) except for the queries which are colored in red. **A.** UPGMA tree with 1,000 bootstrap replicates with support values above branches. Queries are represented in red and all are correctly placed with samples of the presumptive clonal lineage while also representing relationships between clonal lineages in the reference dataset. **B.** Minimum spanning network reconstruction. Edge shade and width are proportional to Bruvo's distance shown in the horizontal scale bar. Queries are represented in red nodes and appear in legend as '???'. Queries placed in nodes with the most similar reference sample, indicating that the US8 query is most similar to the PI-12-016 reference sample (US-8 clonal lineage) and the US23 query is most closely related to the PI-12-023 sample, part of the US-23 lineage.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2005v1 | CC-BY 4.0 Open Access | rec: 28 Apr 2016, publ: 28 Apr 2016

Supplementary Figures

Supplementary Figure 1. UPGMA dendrogram of sub-module MLST-ID of Genotype-ID. The tree was constructed using 100 bootstrap replicates and MLST analysis of 8 genes (5 housekeeping genes and 3 virulence genes) of *Clavibacter michiganensis*. The query used is a concatenation of a WASH sample using all 8 genes. Queries are represented in blue. Note that all queries are correctly placed amongst samples of its presumptive clonal lineage while also representing relationships between lineages in the reference dataset.

Supplementary Figure 2. UPGMA dendrogram of sub-module Binary-ID of Genotype-ID using 100 bootstrap replicates for *Aphanomyces euteiches.* The queries used for this iteration of Binary-ID are two samples from a presumptive "Athena" origin. Queries are represented in red. Note that all queries were correctly placed amongst samples of its presumptive clonal lineage while also representing relationships between lineages in the reference dataset, indicating both queries are more closely related to the "Athena" population than the "Mt. Vernon" population.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2005v1 | CC-BY 4.0 Open Access | rec: 28 Apr 2016, publ: 28 Apr 2016

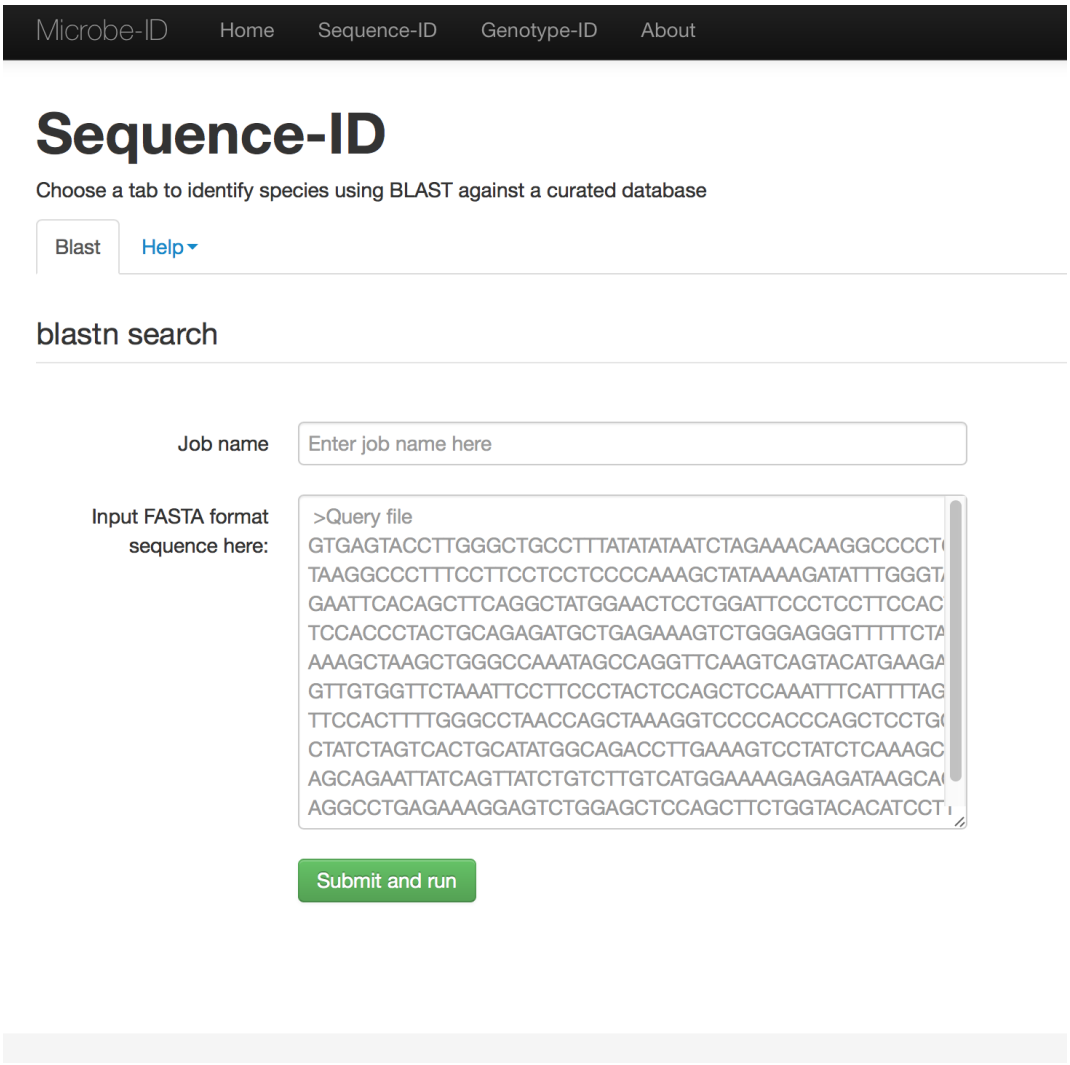

Figure 1. Screen capture of the Sequence-ID user interface of Microbe-ID. The interface includes tabs for input of barcodes and a dropdown help link that contains laboratory protocols, examples data, and logs of each query dataset. The input frame has a control scheme that only permits data in FASTA format.

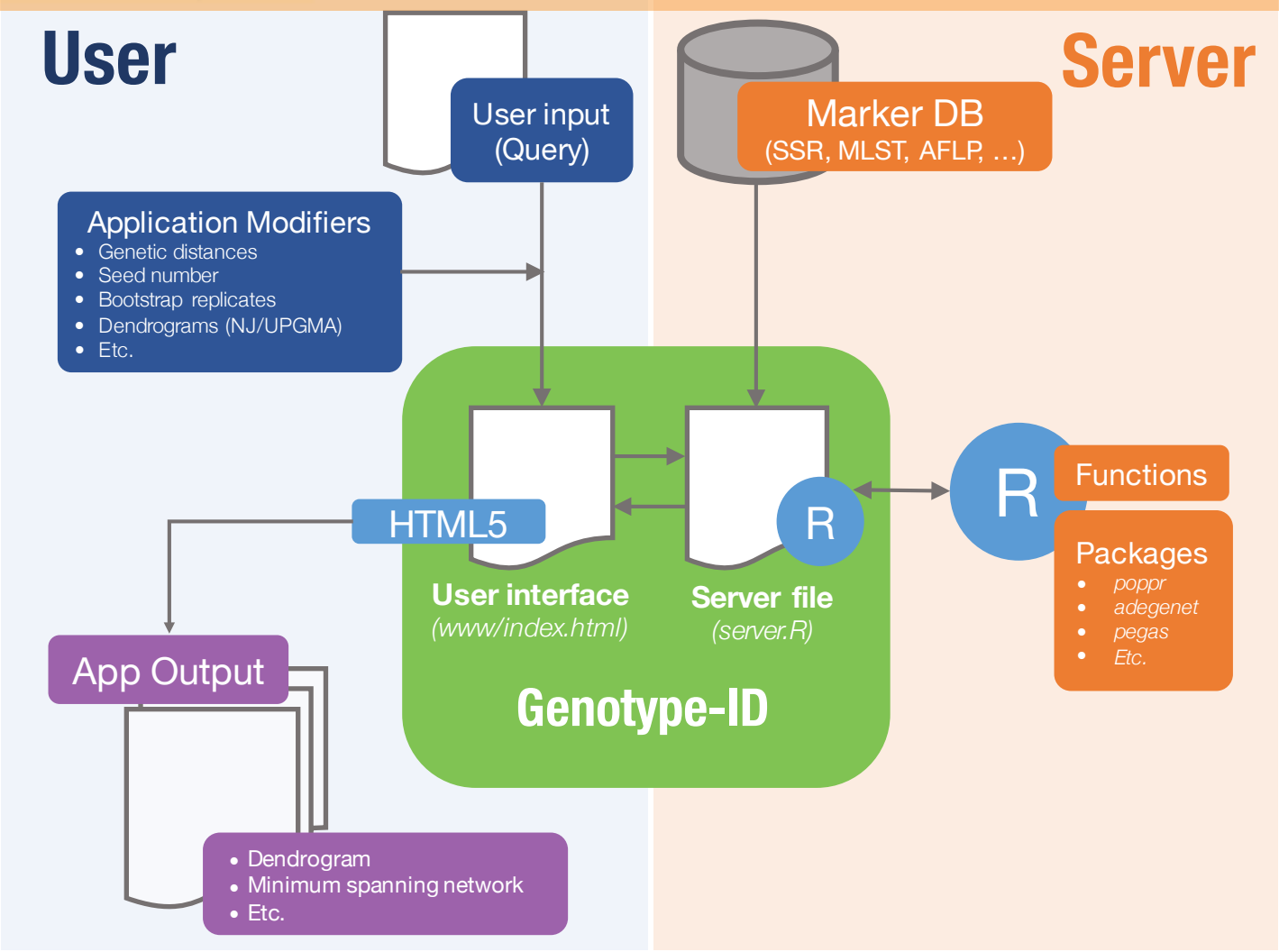

Figure 2. Diagram representing implementation of Genotype-ID, which is comprised of a user interface file (index.html) and a server file (server.R). Each file communicates with the R framework (via shiny) and user (via HTML5). On the user side (left side), user input is provided by copy/paste of a query and selects/specifies the desired application modifiers (seed number, genetic distance calculation). This information is subsequently received and processed by the server file, prompting the application to run in R. On the server side (right side) a database file (Marker DB), R packages, and functions are retrieved and executed. When the run is complete, the server file provides output to the user interface file and displayed on the app output.

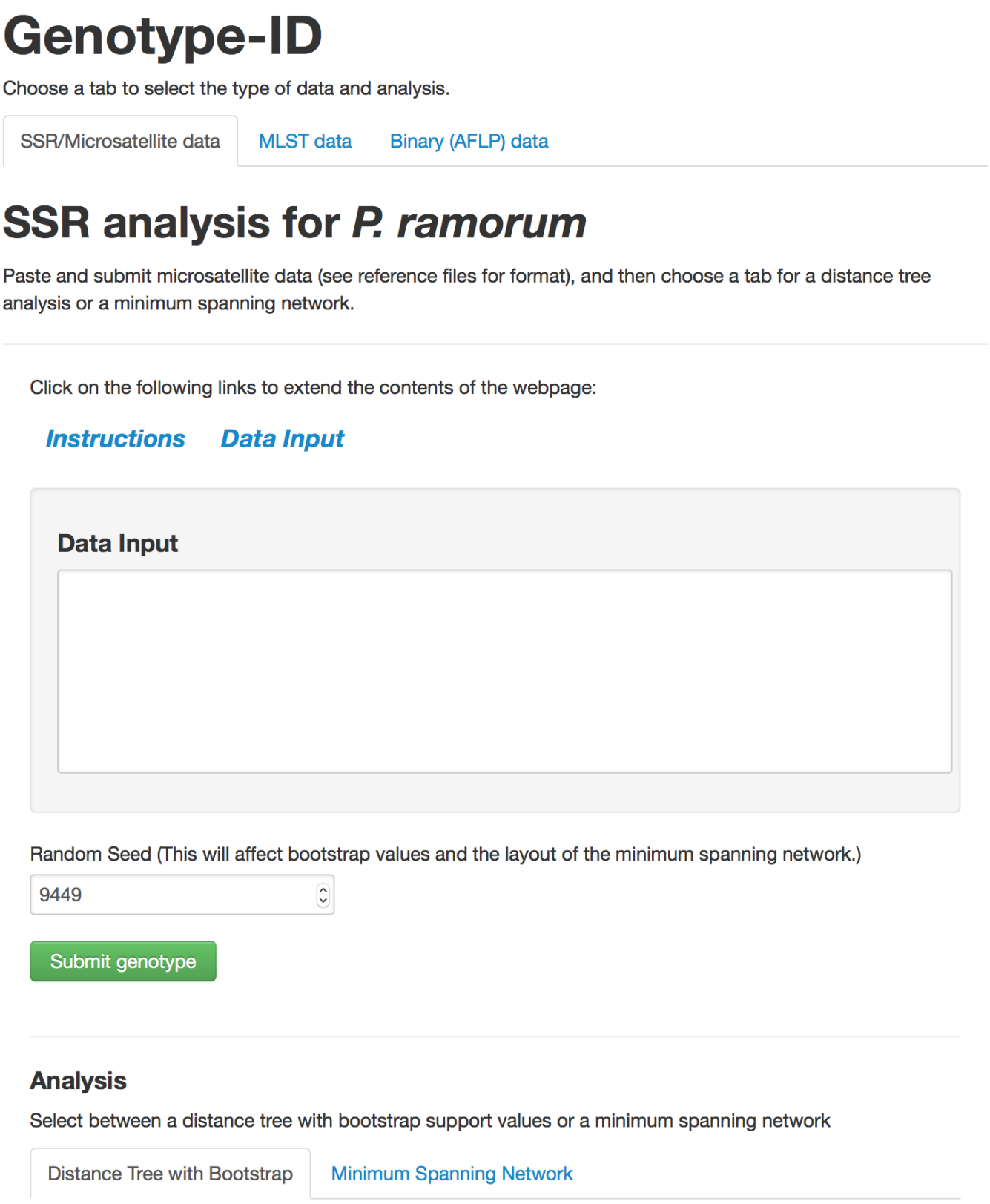

Figure 3. Screen capture of the Genotype-ID user interface of Microbe-ID. Shown is the SSR-ID module for P. ramorum. Genotype-ID includes an Intstructions link with an example Excel file that the user can modify, copy, and paste into the data input form. There is no limit to the number of queries submitted. In the Analysis section, tabs are provided to select either between a distance tree or a minimum spanning network rendering.

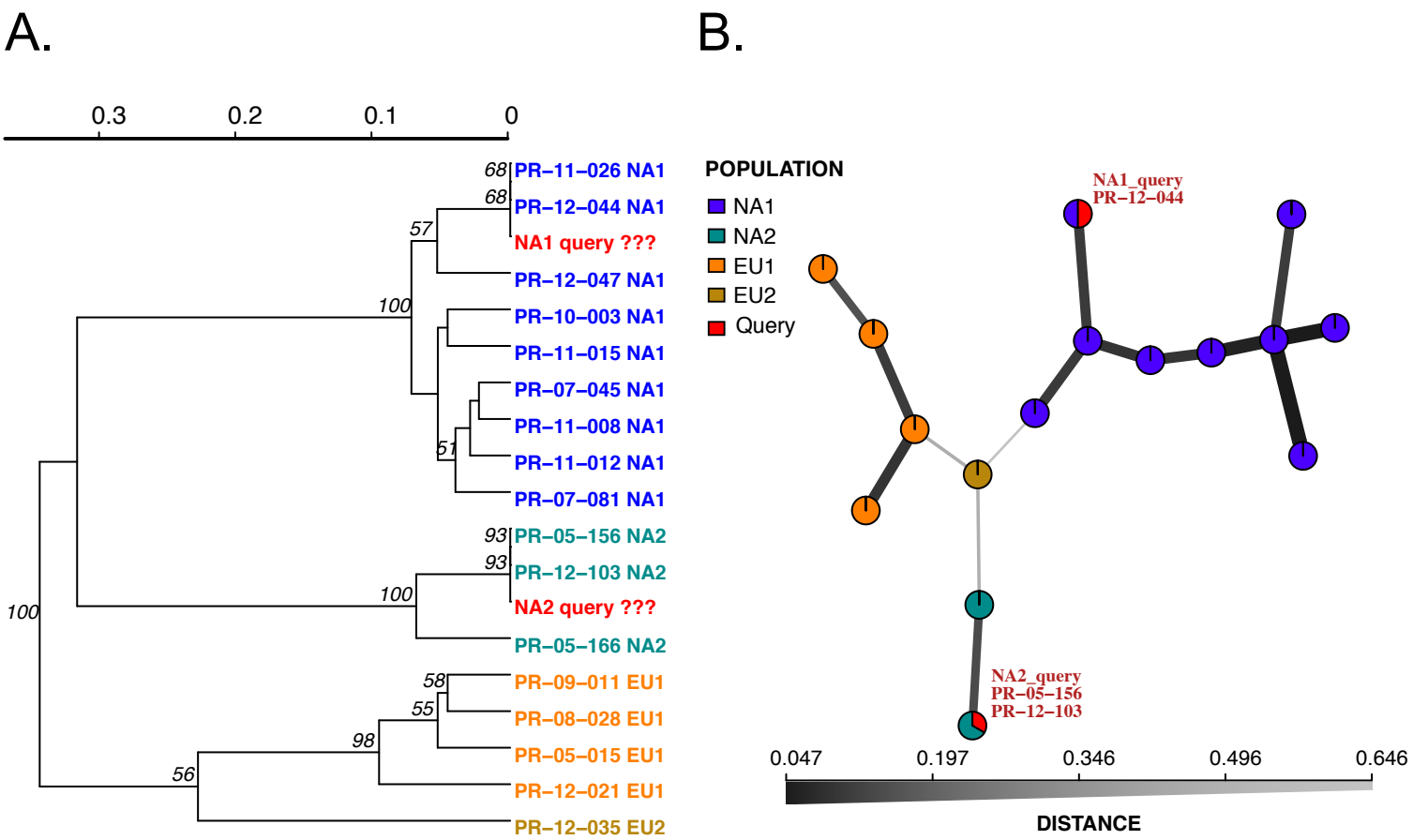

Figure 4. Results of SSR-ID for NA1 and NA2 queries of P. ramorum provided in the example data file. Each color represents a clonal lineage pre-assigned to each reference sample (NA1, NA2, EU1, EU2) with queries colored in red. A. UPGMA tree with 1,000 bootstrap replicates and support values above branches. Queries are represented in red and all are correctly placed with reference samples of the presumptive clonal lineage while also representing the relationship between clonal lineages in the reference dataset. B. Minimum spanning network reconstruction. Edge shade and width are inversely proportional to Bruvo's distance as shown in the horizontal scale bar. Queries are represented in red and placed in nodes with the most similar reference sample in the dataset, indicating the NA1 query is most similar to the PR-12-044 reference sample and the NA2 query is more closely related to the PR-05-156 and PR-12-103 samples, which also belong to the NA2 clonal lineage.

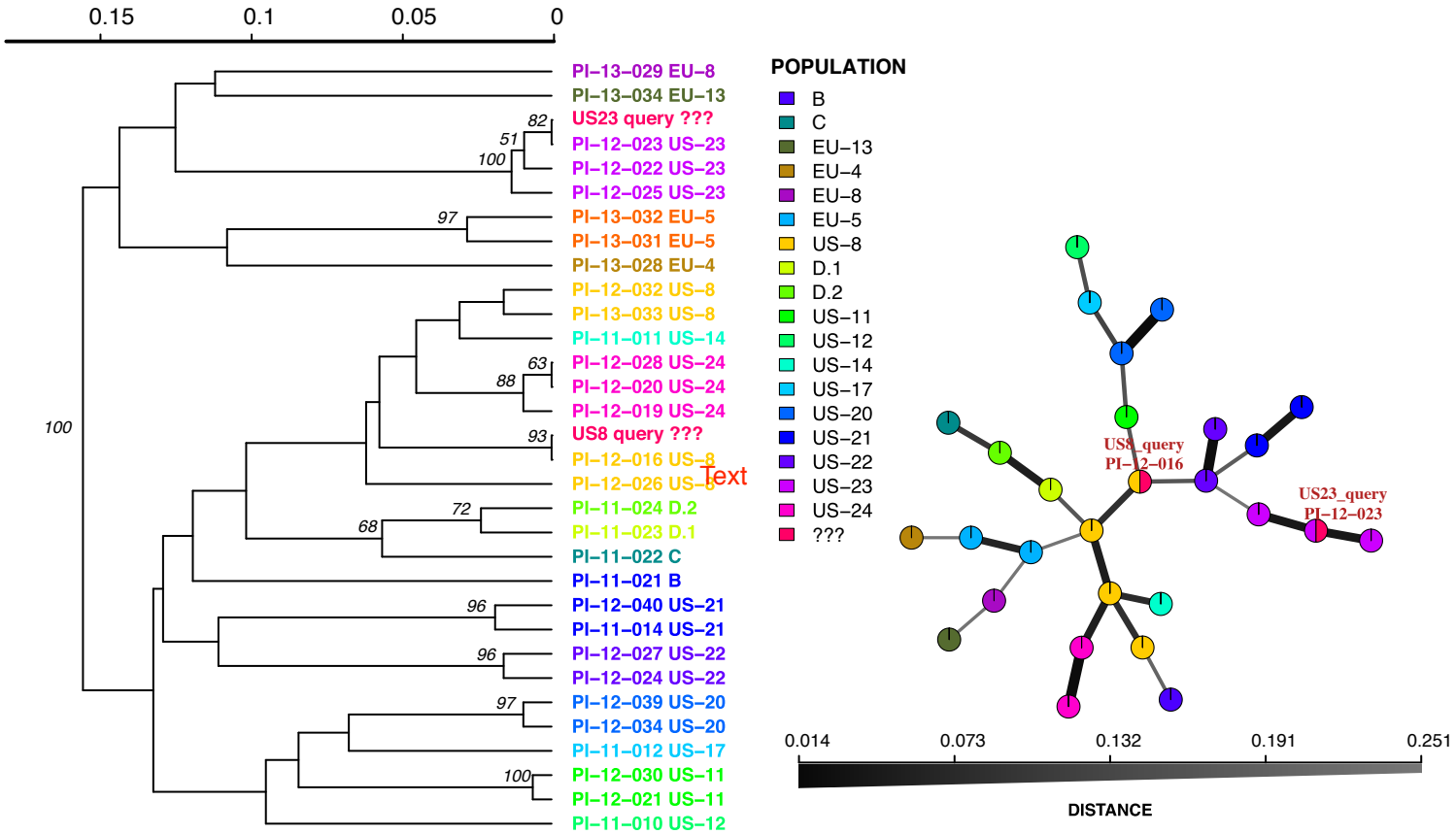

Figure 5. Results of SSR-ID queries for strains placed into the US8 and US23 clonal lineages of the potato late blight pathogen, P. infestans. Colors correspond to clonal lineages assigned to each reference sample (B, C, EU-13, EU-14, etc.) except for the queries which are colored in red. A. UPGMA tree with 1,000 bootstrap replicates with support values above branches. Queries are represented in red and all are correctly placed with samples of the presumptive clonal lineage while also representing relationships between clonal lineages in the reference dataset. B. Minimum spanning network reconstruction. Edge shade and width are proportional to Bruvo's distance shown in the horizontal scale bar. Queries are represented in red nodes and appear in legend as '???'. Queries placed in nodes with the most similar reference sample, indicating that the US8 query is most similar to the PI-12-016 reference sample (US-8 clonal lineage) and the US23 query is most closely related to the PI-12-023 sample, part of the US-23 lineage.