1
2
3
4 1 REVISED manuscript MIMET-D-15-00816 on Feb. 8, 2016
5
6

7 2
8
9 3 A computer program for fast and easy typing of a partial *endoglucanase* gene sequence into
10
11 4 genospecies and sequevars 1&2 of the *Ralstonia solanacearum* species complex
12
13
14 5
15
16 6
17

18 7 Michael J Stulberg^{1,2} and Qi Huang^{1*}
19
20
21 8
22

23 9 ¹Floral and Nursery Plants Research Unit, Agricultural Research Service, United States
24
25

26 10 Department of Agriculture, Beltsville, Maryland, United States of America
27
28

29 11 ²Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee, United States of
30
31

32 12 America
33
34 13
35
36 14
37
38 15
39
40
41 16
42

43 17 Key words: *Ralstonia solanacearum* species complex, genospecies, phylotype, sequevar,
44
45

46 18 endoglucanase, sequence typing, quarantine pathogen, select agent, computer program
47
48 19
49

50 20 Short title: Computer program to aid the typing of *Rs* species complex strains
51
52
53 21
54
55 22
56
57

Abstract

The phytopathogen *Ralstonia solanacearum* is a species complex that contains race 3 biovar 2 strains belonging to phylotype IIB sequevars 1 and 2 that are quarantined or select agent pathogens. Recently, the *R. solanacearum* species complex strains have been reclassified into three genospecies: *R. solanacearum*, *R. pseudosolanacearum* and *R. syzyzii*. An unidentified *R. solanacearum* strain is considered a select agent in the US until proven to be a non-race 3 biovar 2 (non-phylotype IIB sequevars 1&2). Currently, sequevars of *R. solanacearum* species complex strains can only be determined by phylogenetic analysis of a partial *endoglucanase* (*egl*) sequence of approximately 700-bp in length. Such analysis, however, requires expert knowledge to properly trim the sequence, include the correct reference strains, and interpret the results. By comparing GenBank *egl* sequences of representative *R. solanacearum* species-complex strains, we identified genospecies- and sequevar 1 and 2-specific single nucleotide polymorphisms (SNPs). We also designed primers to amplify a shorter, 526-bp, *egl* fragment from *R. solanacearum* species complex strains for easy sequencing of the amplicon, and to facilitate direct and specific amplification of *egl* from *R. solanacearum*-infected plant samples without the need of bacterial isolation. We wrote a computer program (*Ralstonia solanacearum* typing program) that analyzes a minimum 400-bp user-input *egl* sequence from a *R. solanacearum* strain for *egl* homology and SNP content to determine 1) whether it belongs to the *R. solanacearum* species complex, 2) if so, to which genospecies, and 3) whether it is of the sequevar type (sequevars 1 and 2) associated with the select agent/quarantined *R. solanacearum* strain. The program correctly typed all 371 tested *egl* sequences with known sequevars, obtained either from GenBank or through personal communication. Additionally, the program successfully typed 25 *R. solanacearum* strains in our collection with no prior sequevar

information, as well as 4 strains in infected plant samples, using their partial *egl* sequences amplified and sequenced with primers designed in this study. The *Ralstonia solanacearum* typing program does not require expertise or specific knowledge to use, gives results in seconds, and provides data interpretation for the user. The program and primers can help expert or non-expert users to quickly type an unknown *R. solanacearum* species-complex strain and determine whether it is a highly regulated *R. solanacearum* strain. The program can also serve as a confirmation method, since it is the only method that can easily and directly determine whether the strain in question is a sequevar 1 or 2 strain of *R. solanacearum*.

Introduction

The *Ralstonia solanacearum* species-complex strains are classified into four phlotypes currently made up of 53 sequevars, determined by a variable part of the *endoglucanase* (*egl*) gene sequence (Albuquerque et al., 2014). Phlotypes I-IV correspond with the strains' geographical origin, and phlotype II is split into groups A and B (Fegan and Prior, 2005; Poussier et al., 2000). Sequevar designation does not adhere to any pattern – the 53 sequevars are distributed among the four phlotypes – but it is widely accepted that sequevars 1 and 2 in phlotype IIB contain the strains previously defined as race 3 biovar 2 (r3b2). The r3b2 strains cause destructive brown rot of potato and are listed as select agents by the U.S. Government and as quarantine pathogens by many other countries (Agricultural Bioterrorism Protection Act of 2002). As a result, once a *R. solanacearum* strain is detected, it is critical to quickly determine whether the strain belongs to r3b2 (phlotype IIB sequevar 1 or 2) to make timely decisions to exclude this select agent/quarantine pathogen from entering and/or becoming established. We have recently developed multiplex PCR and qPCR assays based on unique gene sequences to

1
2
3
4 69 detect and differentiate phylotype IIB sequevar 1 and 2 strains (Stulberg et al., 2015; Stulberg
5
6 70 and Huang, 2015), but to determine the exact sequevar identity of a *R. solanacearum* strain, a
7
8
9 71 phylogenetic analysis must be performed using a partial sequence of the strain's *egl* gene
10
11
12 72 (Albuquerque et al., 2014; Fegan and Prior, 2005; Poussier et al., 2000; Wicker et al., 2007).

13
14 73 The *R. solanacearum* species complex has been reclassified into three genospecies: *R.*
15
16 74 *solanacearum*, *R. pseudosolanacearum*, and *R. sygyii* with multiple subspecies (Safni et al.,
17
18
19 75 2014). The new classification utilized the phylotype divisions, with phylotype II strains
20
21 76 remaining as *R. solanacearum*, phylotypes I and III strains reclassified as *R.*
22
23 77 *pseudosolanacearum*, and phylotype IV strains split into three subspecies of *R. syzygii*. The
24
25
26 78 current methods for determining the genospecies of a strain relies on phylogenetic analysis or
27
28
29 79 phylotyping using a multiplex PCR (Fegan and Prior, 2005). The multiplex PCR is only able to
30
31 80 identify a phylotype and further tests are needed for sequevar identification.

32
33 81 Currently, there is no standard for typing *R. solanacearum* using partial *egl* sequences.
34
35
36 82 Specifically, there is no consensus on 1) the number and type of reference strains that should be
37
38
39 83 used to build an *egl*-based phylogenetic tree (Albuquerque et al., 2014; Guidot et al., 2009;
40
41 84 Mahbou Somo Toukam et al., 2009; Poussier et al., 2000; Santana et al., 2012; Stulberg et al.,
42
43 85 2015; Wicker et al., 2007), 2) which primers to use to amplify the variable *egl* region (Hong et
44
45
46 86 al., 2012; Poussier et al., 2000), and 3) where to trim the sequence once it is obtained (Hong et
47
48 87 al., 2012; Poussier et al., 2000; Santana et al., 2012; Schneider et al., 2011; Wicker et al., 2007).
49
50
51 88 Moreover, *egl* sequences deposited to GenBank are either untrimmed or trimmed to various
52
53 89 lengths depending on the individual studies, meaning that even the simple act of downloading
54
55
56 90 sequences requires expert knowledge to know which sequences can be used as references. Given
57
58 91 these obstacles, specific knowledge and expertise are required for creating and analyzing *egl*-
59
60
61
62
63
64
65

based phylogenetic trees in order to determine sequevars of newly identified *R. solanacearum* strains or strains with undetermined sequevar status. Our goal was to develop a computer program that can type an *egl* sequence into a genospecies and can also determine whether it is sequevar 1 or 2 of the *R. solanacearum* species complex, while standardizing and simplifying the procedure for end users by eliminating the need for trimming *egl* sequences and picking reference strains.

Partial *egl* sequences in *R. solanacearum* species complex strains were analyzed to reveal genospecies/phyloptype- and sequevar 1 or 2-specific single-nucleotide polymorphisms (SNPs) for strain typing, or for conserved regions that could be used for amplifying and sequencing a shorter *egl* sequence. The SNPs were embedded in a computer program written to analyze an *egl* sequence automatically, to identify the strain's genospecies, and to determine whether it belongs to the highly regulated *R. solanacearum* sequevar 1 or 2. The program eliminates the need for trimming of the *egl* sequences, and ensures that all of the proper reference strains are included in analysis of genospecies/sequevar 1 or 2 typing. Furthermore, our computer program allows customizable messages to be programmed with test results, suggesting end users of next steps to take if further identity confirmation is required.

Methods

Design of primers and identification of SNPs:

To design primers for amplifying a shorter *egl* sequence for use in our computer program, partial *egl* sequences were aligned using Multiple Sequence Comparison by Log-Expectation (MUSCLE) (Edgar, 2004) in Mega 6 (Tamura et al., 2013) to look for conserved regions for primer design. Fourteen GenBank deposited *egl* sequences - JT516 (AF295258), UW551

(DQ657596), CFBP4611 (GU294981), LNPV32.37 (GU294932), UW349 (KF701622), Molk2 (EF371841), UW469 (DQ657612), K60 (AF295262), JT519 (GU295032), JT525 (AF295272), CMR15 (EF439743), Psi07 (EF371804), MAFF301552 (GU295042), and R229 (GU295045) - representing each of the three genospecies/four phylotypes (including IIB, IIA) and sequevars 1, 2, 3, 4, 7, 8, 10, 19, 27, 29, 31, and 40 (multiple representatives for 1 and 10) were aligned, and conserved regions were assessed for primer design. For alignment to identify SNPs specific to phylotype IIB sequevar 1 or 2, as well as the different genospecies/phylotypes, three additional biovar 2 *egl* sequences from strains Rs240 (GU049839), Rs115 (GU049809), and Rs205 (GU049821) were added to the above 14 sequences. Sequences were again aligned using MUSCLE in Mega 6 and manually searched for the specific SNPs.

Primer suitability was tested in A plasmid Editor (ApE) (Davis, 2015) with specific criteria (GC = 45 to 60%, T_m = 60-64°C, primer length = 18-26 bp).

Design of the *R. solanacearum* typing computer program to type *R. solanacearum* strains based on partial *egl* sequences:

The computer program was written in Python. Tkinter and PyInstaller were used to make the graphical user interface and to package the program into one executable file (S1 Program), respectively. Users can input sequence in FASTA format into the “enter sequence” text field using the “paste” button after copying the sequence or by typing it in, or they can input sequence by opening a phd formatted file (*.phd.1), which are provided by many major genome centers (generated by Phred, which returns bases with an assigned quality value from a DNA sequencing trace file). A built-in quality check was included, so if the user imports a phd file, the program will check the sequencing quality of each nucleotide in the sequence for the user and only accept

1
2
3
4 138 a nucleotide if its quality value is above 20 (99% accuracy) (Ewing and Green, 1998), consistent
5
6 139 with the thresholds utilized by most other method. If a nucleotide's quality value does not meet
7
8
9 140 the 20 threshold, it is replaced with an "n" in the sequence.
10

11 Once the user has input sequence, the "execute" button can be pushed to run the analysis
12 141
13
14 142 (on some high-resolution monitors, the window of the program may need to be expanded to see
15
16 143 the "execute" button). The user-input sequence requires a minimum of 400-bp in length to make
17
18
19 144 an accurate phylotype assignment. Larger sequences (e.g., 750-bp *egl* fragments from GenBank)
20
21 145 can also be used without affecting the analysis. After pushing "execute", the program uses the
22
23
24 146 Smith-Waterman algorithm for local alignment (Smith and Waterman, 1981) written in Python
25
26 147 by Ryan Boehning (Boehning, 2013), with minor adjustments to the algorithm to better handle
27
28
29 148 alignment to the starting nucleotide and adapting it to our graphical user interface. The >400-bp
30
31 149 input sequence is compared to a 413-bp fragment of a sequevar 1 strain UW551 *egl*, and if it
32
33 150 matches that fragment with >80% homology, the input sequence is passed to the next series of
34
35
36 151 SNPs tests to determine if it is a *R. solanacearum* genospecies strain and, additionally, whether it
37
38 152 is a sequevar 1 or 2 strain or whether it belongs to one of the other two genospecies. The >80%
39
40
41 153 homology was chosen empirically, since the most distantly related partial-*egl* sequence to the
42
43 154 UW551 reference sequence is about 84% using the Smith-Waterman algorithm, and 80% was
44
45
46 155 chosen as the cutoff to allow for some sequencing errors at the ends. No other sequences
47
48 156 belonging to different species in GenBank were found to match this 400-bp of partial *egl*
49
50
51 157 sequence in the NCBI BLAST database. In each of the SNP tests, the entire input sequence is
52
53 158 compared to sequences that are 5- to 16-bp in length using the Smith-Waterman algorithm, and
54
55
56 159 each must match with 100% nucleotide identity. The program only tests the sequence alignment
57
58 160 in the 5' to 3' orientation, and the user should be able to determine the orientation of the
59
60
61
62
63
64
65

sequence based on whether our F- (for 5' to 3') or R- (for 3' to 5') primer was used for sequencing. If the reverse primer was used, the user should input the sequence into the “enter sequence” field and then push “reverse compliment” followed by “execute”.

DNA extraction, PCR conditions, and sequencing:

Genomic DNA was extracted from the bacterial strains listed in Table 1 following Qiagen’s Blood and Tissue kit (Valencia, CA) protocol. Total DNA from healthy and *R. solanacearum*-infected tomato, geranium, potato, and tobacco was extracted using Qiagen’s Plant kit according to manufacturer’s instructions (Stulberg et al., 2015).

DNA from 26 *R. solanacearum* strains in our collection and ten plant samples infected with strain K60 (sequence (seq) 7), GMI1000 (seq 18), UW551 (seq 1), or UW349 (seq 27) was used as template for PCR reactions with our MJS-amp-egl primer pair. Plant samples were symptomatic tobacco (*Nicotiana tabacum*), tomato (*Lycopersicon esculentum* Mill cv. 'bonnie best'), potato (*Solanum tuberosum* 'Russet Norkotah'), and geranium (*Pelargonium x hortorum* 'zonal geranium') prepared as before (Stulberg et al., 2015). DNA from healthy plant samples was used as PCR negative controls. PCR assays were performed in a 20 µl volume using Qiagen’s *Taq* DNA polymerase and *Taq* PCR core kit (Valencia, CA). Reactions contained 1 x CoralLoad PCR buffer, 200 µM of each dNTP, 1 x Q-solution, 2.5 units of *Taq*, 250 pmol of each primer, and either 3 ng of bacterial DNA or 20 ng of Qiagen-extracted total plant DNA. DNA was quantified using a nanodrop 2000c UV-Vis spectrophotometer (Thermo Scientific, Waltham, MA). Amplicons were visualized on 1% agarose gels, and the 526-bp bands from the twenty-six strains and 4 infected plant samples were purified using a Qiagen Gel Extraction kit. The bands were commercially sequenced by Eurofin (Eurofins MWG Operon LLC, Huntsville,

AL) using the MJS-seq-egl primers by Eurofin (Eurofins MWG Operon LLC, Huntsville, AL). Forward and reverse sequences were assembled using SeqMan Pro in DNASTAR Lasergene suite v11 (DNASTAR, Inc, Madison, WI) and analyzed. The 2x coverage sequences were deposited into GenBank and given accession numbers (Table 1).

Validation and utilization of the *R. solanacearum* typing program:

The computer program was validated using 371 *egl* sequences deposited in GenBank or received through personal communication with known phylotype/sequevar or r3b2 (phylotype IIB sequevars 1 and 2) identities (S2 Table). It was used on 30 *egl* sequences from 26 *R. solanacearum* species-complex strains in our collection and from four *R. solanacearum* species complex-infected plant samples by importing their phd.1 files directly into our computer program for analysis.

Results

Primers were designed for amplifying a shorter 526-bp *egl* sequence from *R. solanacearum* species-complex strains and for sequencing the resulting amplicon (Fig. 1). The 5' to 3' primers for amplifying *egl* are MJS-amp-egl-F, CGGACACGGACACCACGACTCTGAA and MJS-amp-egl-R, TAGCGCGCATAGTTGTGCGGATCGAGC; and the sequencing primers are MJS-seq-egl-F, ACACGGACACCACGACTCTG and MJS-seq-egl-R, GCATAGTTGTGCGGATCGAGC. Originally, the MJS-seq-egl primers were designed to amplify *egl* from *R. solanacearum* species-complex strains. It was observed, however, that they could not amplify all members of *R. pseudosolanacearum* (phylotype I specifically) equally well as other species-complex strains, prompting redesign of these primers by adding three and two

nucleotides on the 5' and 3' ends of MJS-seq-egl-F, respectively, and five nucleotides on the 5' end of MJS-seq-egl-R to obtain the MJS-amp-egl-F and MJS-amp-egl-R primers for better amplification of *egl* sequences from all *R. solanacearum* species-complex strains. The *egl* amplification primers MJS-amp-egl successfully amplified *egl* sequences from DNA isolated from 28 strains in our collection representing each of the three genospecies. Most importantly, they also successfully amplified *egl* directly from Qiagen-extracted plant DNA from *R. solanacearum* species complex-infected tomato, potato, tobacco, and geranium, and non-specific bands near the target size were not amplified by the MJS-amp-egl primer pair in the presence of plant DNA (Fig. 2). The MJS-seq-egl primers were used as sequencing primers for 26 of the 28 bacterial and four selected infected plant *egl* amplicons (Table 1). Amplicons from CMR15 and Psi07 were not sequenced since sequence for these strains was already deposited in GenBank (Mahbou Somo Toukam et al., 2009; Wicker et al., 2007).

Additional sequence alignment of 17 *egl* sequences deposited in GenBank revealed specific SNPs that could be used to identify each of the genospecies and sequevars 1 and 2. Identification of SNPs was limited to the region that could be consistently sequenced with 2x coverage by the MJS-seq-egl primers (a 413-bp region within the 526-bp amplified region) (Fig. 1). This 413-bp region serves as the target region for the computer program decision tree based on identified SNPs (Fig. 3).

The *R. solanacearum* typing program matches a user-input sequence to the entire 413-bp target region (S1 Program). If the input sequence has more than 80% identity with the 413-bp sequence of sequevar 1 strain UW551 (the 413-bp is highly conserved in *R. solanacearum* species complex strains), it is considered to belong to the *R. solanacearum* species complex, and the result is included in the display of the first test result (Fig. 4). The input sequence is then

aligned to a *R. solanacearum* SNPs (underlined nucleotide)-containing sequence (SNPs sequence) (gttccggcc) and proceeds through the program decision tree (Fig. 3). If the sequence fails the first *R. solanacearum*-specific sequence it is then compared to a *R. syzygii* SNPs sequence (gaaagag). If it matches, it is determined to be a “*Ralstonia syzygii* strain (formerly phylotype IV of *R. solanacearum* species complex)”, but if it fails, it is compared to two *R. pseudosolanacearum* SNPs sequence (gtggcgaa, gtgcagaa). If it fails all of these tests then it is determined to be a “*Ralstonia syzygii*” because the *egl* sequence for *R. syzygii* strain R058, a *R. solanacearum* species-complex phylotype IV strain with no sequevar assignment, fell into this category (S2 Table). If the input sequence passes the first *R. solanacearum* test, it then goes to a second *R. solanacearum* test (gtggaaa). If the sequence fails the second test, presumably it is still *R. solanacearum* although none of the *egl* sequences tested fell into this category, otherwise it is passed to the four “sequevar 1&2” tests. There are three sequences that must match to be considered sequevars 1&2 (gcggcacacgcacggt, gcgctgtc, cgata), and if the sequence fails any of these tests it is designated “*Ralstonia solanacearum* strain (formerly *Rs* phylotype II), not sequevar 1 or 2”. If the input sequence matches all three of those with 100% nucleotide identity, it is then compared to a sequence that differentiates sequevar 2 strains from sequevar 1 (ggtggg). If it matches this sequence, the program designates it “*Ralstonia solanacearum* strain (formerly *Rs* phylotype II), sequevar 2,” and if it does not, then it is designated “*Ralstonia solanacearum* strain (formerly *Rs* phylotype II), sequevar 1” (Fig 4). If the sequence is categorized into either of these two sequevars, the user is notified of the identification of a select agent/quarantine pathogen and is instructed to perform further testing to confirm the typing result.

The computer program was tested using 371 partial *egl* sequences deposited in GenBank and from Dr. Jason Hong (personal communication), representing all of the 53 known sequevars

(S2 Table). The program correctly identified sequences from 102 phylotype IIB sequevar 1 strains and 4 sequevar 2 strains, as well as sequences from 182 non-sequevars 1 and 2 strains in the *R. solanacearum* genospecies (phylotype II) (Fig. 3). The majority of the sequences from *R. solanacearum* strains were identified as “*Ralstonia solanacearum* (formerly *Rs* phylotype II) not sequevar 1 or 2” at test #1 (n = 171), but more closely related *R. solanacearum* sequences needed tests #2 (n = 10) and #3 (n = 1) to filter from the sequevars 1 and 2 (Fig. 3). The program also correctly identified sequences from 70 *R. pseudosolanacearum* strains (53 phylotype I and 17 phylotype III strains) and 13 *R. syzygii* strains (Fig. 3) representing each subspecies (*R. syzygii* subsp. *syzygii*, *R. syzygii* subsp. *indonesiensis*, and *R. syzygii* subsp. *celebesensis*).

The computer program was used to determine the sequevar identity of 26 *R. solanacearum* species-complex strains in our collection, 25 with unknown sequevar identities, including 20 r3b2 strains and 5 non-r3b2 strains (Table 1). The program identified all of the 20 r3b2 strains as *R. solanacearum* sequevar 1, and it correctly identified the 5 non-r3b2 strains as *R. solanacearum* non-sequevar 1 or 2 strains (n = 2) or as *R. pseudosolanacearum* strains (n = 3) (Table 1). The previously identified phylotype I sequevar 13 strain Rs121 (Stulberg and Huang, 2015) was confirmed to be a *R. pseudosolanacearum* strain by our computer program (Table 1). The 4 *egl* sequences amplified directly from infected plant DNA samples were also tested by the computer program (Fig. 2), and the genospecies/sequevar 1 or 2 status of the *R. solanacearum* strains used for infection were correctly identified (Table 1).

Discussion

We wrote the *Ralstonia solanacearum* typing program that types a partial *egl* gene sequence into one of the three genospecies in the *R. solanacearum* species complex and also

identifies the highly regulated sequevar 1&2 strains in *R. solanacearum*. The program accurately typed *egl* sequences collected from GenBank or through personal communication (n=371) as well as sequences from our collection and infected plant samples (n=30). The program is a user-friendly, stand-alone executable file that can determine in seconds whether an *egl* sequence belongs to the highly regulated sequevars, and no taxonomic and phylogenetic expertise is needed to use the program or to interpret results. The steps to use the computer program to type a *R. solanacearum* species-complex strain include: 1) Obtaining a partial *egl* sequence, if it is not available in GenBank. This can be done by amplifying the 526-bp *egl* sequence directly from infected plant DNA samples or isolated bacterial cells with our MJS-amp-egl primers by PCR and sequencing the amplicon using our MJS-seq-egl primers. Partial *egl* sequences can also be obtained using previously published *egl* primers (Hong et al., 2012; Poussier et al., 2000), although additional bands have been noted for the Endo primers (Poussier et al., 2000) in plant and bacterial DNA samples using established PCR conditions (Mahbou Somo Toukam et al., 2009). 2) The user then inputs the *egl* sequence into the “enter sequence” text field by either copying and pasting the sequence or by clicking the phd1 file button to directly use a Phred file (a file provided by the sequencing company), which allows the user a streamlined method of inputting sequence (Fig. 4). 3) The user presses the “execute” button, and the results are displayed in the “results” field. The “results” box was designed to be immutable to prevent manipulation. 5) Based on the results, users are advised what to do next to confirm the results, especially if a sequevar 1 or 2 strain is identified.

There are a number of DNA-based amplification assays for the detection of r3b2 strains of the *R. solanacearum* species complex (Fegan et al., 1998; Guidot et al., 2009; Ha et al., 2012; Kubota et al., 2011; Kubota and Jenkins, 2015; Stulberg et al., 2015; Stulberg and Huang, 2015;

Weller et al., 2000). These assays target unique DNA/gene sequences supposedly conserved in all r3b2, phylotype IIB sequevar 1 or 2 strains, although a few of the assays (Ha et al., 2012; Kubota et al., 2011; Weller et al., 2000) also detect at least one non-sequevar 1 or 2 strain. There is also a multiplex PCR assay that determines only the phylotype of a given strain (Fegan and Prior, 2005). All of these assays do not detect *R. solanacearum* directly at the sequevar level, and require positive controls, which can be difficult to obtain given the regulations surrounding this pathogen. The *Ralstonia solanacearum* typing program serves to complement the DNA-based amplification assays by providing direct sequence verification of a sequevar 1 or 2 strain and also providing the new genospecies in one analysis. The program also allows users without access to positive controls an easy method to determine sequevar and genospecies. The program is easier to use than building a phylogenetic tree because it does not require any specialized experience, avoids the issues of sequence trimming and use of proper reference strains, and interprets the results for the user. Additionally, our program directly identifies whether the strain is sequevar 1 or 2, which the DNA-based amplification assays cannot yet resolve.

New primers were designed to amplify a shorter region of *egl* for use in the *R. solanacearum* typing program, since they are more specific than the Endo primers using Qiagen's polymerase (data not shown). The MJS-amp-egl primers are especially useful for amplifying a specific *egl* product directly from *R. solanacearum*-infected plant samples, such as its major hosts of potato, tomato and geranium, saving time by eliminating the more laborious bacterial isolation step. Furthermore, DNA-amplification inhibitors can accompany some plant samples (especially geranium and potato) (Ha et al., 2012), and a shorter amplicon helps reduce the risk of PCR inhibition. Our MJS-amp-egl primers, therefore, will be especially useful for difficult samples where the Endo primers may produce non-target bands or the amplification is

partially/fully inhibited in plant extracts. While the primers work well for strains in phylotypes II, III, and IV of the *R. solanacearum* species complex, phylotype I strains are amplified less efficiently, and there is an additional band that is amplified at a much higher molecular weight (Fig. 2, lane 5). The optimization of annealing temperatures with a given polymerase is necessary for the best amplification of *egl* in phylotype I strains.

In conclusion, the computer program is useful for end users who want to type a *R. solanacearum* species-complex strain into the three new genospecies and to determine whether it belongs to the highly regulated sequevar 1 or 2. Our primers are useful for those interested in amplifying partial *egl* directly from an infected plant sample for use in the typing program. The program is also especially useful as a secondary confirmation of a PCR r3b2-positive sample, or to help researchers who do not have access to the proper sequevar 1 or 2 positive controls to determine whether they have a sequevar 1 or 2 strain. The program is not designed to assign a sequevar other than sequevar 1 or 2 or replace phylogenetic analysis, but it can be used to help supply quick and accurate genospecies/sequevar 1 or 2 typing. It will simplify *R. solanacearum* sequevar 1 or 2 analysis in lieu of set standards for phylogenetic analysis and can also be used to convey other, relevant information, since interpretations of the results and next steps can be coded for a variety of outcomes. In addition to sequevars 1 and 2, this program architecture could easily be adapted to typing other sequevars of *R. solanacearum*, if needed in the future, or other important plant pathogens.

Acknowledgements

This research was financially supported by the U.S. Department of Agriculture (USDA), Agricultural Research Service (ARS) and Animal and Plant Health Inspection Service. It was supported in part by an appointment to the ARS Research Participation Program administered by

the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and the USDA. ORISE is managed by ORAU under DOE contract number DE-AC05-06OR23100. We thank Daniel Winograd-Cort for his help in learning computer science. We thank Jason Hong for providing *egl* sequences and Philippe Prior for giving us insight into *R. solanacearum egl* phylogenetic analysis. We also thank John Hartung for critical review of our manuscript.

References

- Agricultural Bioterrorism Protection Act of 2002: Possession, use, and transfer of biological agents and toxins; interim rule and request for comments., 2002. . 67 Fed. Regist. 240, 76908–76938.
- Albuquerque, G.M.R., Santos, L.A., Felix, K.C.S., Rollemberg, C.L., Silva, A.M.F., Souza, E.B., Cellier, G., Prior, P., Mariano, R.L.R., 2014. Moko Disease-Causing Strains of *Ralstonia solanacearum* from Brazil Extend Known Diversity in Paraphyletic Phylotype II. *Phytopathology* 104, 1175–1182. doi:10.1094/PHYTO-12-13-0334-R
- Boehning, R., 2013. Smith Waterman Implementation in Python [WWW Document]. URL <https://gist.github.com/radaniba/11019717>
- Davis, M.W., 2015. ApE - A plasmid Editor v.2.0.47 [WWW Document]. URL <http://biologylabs.utah.edu/jorgensen/wayned/ap/>
- Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi:10.1093/nar/gkh340
- Ewing, B., Green, P., 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 8, 186–194.

- Fegan, M., Holoway, G., Hayward, A.C., Timmis, J., 1998. Development of a Diagnostic Test Based on the Polymerase Chain Reaction (PCR) to Identify Strains of *Ralstonia solanacearum* Exhibiting the Biovar 2 Genotype, in: Prior, P., Allen, C., Elphinstone, J.G. (Eds.), Bacterial Wilt Disease: Molecular and Ecological Aspects. Springer, pp. 34–43.
- Fegan, M., Prior, P., 2005. How complex is the *Ralstonia solanacearum* species complex?, in: Allen, C., Prior, P., Hayward, A.C. (Eds.), Bacterial Wilt Disease and the *Ralstonia solanacearum* Species Complex. APS Press, Minnesota, pp. 449–461.
- Guidot, A., Elbaz, M., Carrère, S., Siri, M.I., Pianzzola, M.J., Prior, P., Boucher, C., 2009. Specific genes from the potato brown rot strains of *Ralstonia solanacearum* and their potential use for strain detection. *Phytopathology* 99, 1105–1112.
- Ha, Y., Kim, J.-S., Denny, T.P., Schell, M.A., 2012. A rapid, sensitive assay for *Ralstonia solanacearum* race 3 biovar 2 in plant and soil samples using magnetic beads and real-time PCR. *Plant Dis.* 96, 258–264.
- Hong, J.C., Norman, D.J., Reed, D.L., Momol, M.T., Jones, J.B., 2012. Diversity Among *Ralstonia solanacearum* Strains Isolated from the Southeastern United States. *Phytopathology* 102, 924–936.
- Kubota, R., Jenkins, D., 2015. Real-Time Duplex Applications of Loop-Mediated AMPLification (LAMP) by Assimilating Probes. *Int. J. Mol. Sci.* 16, 4786–4799.
doi:10.3390/ijms16034786
- Kubota, R., Schell, M.A., Peckham, G.D., Rue, J., Alvarez, A.M., Allen, C., Jenkins, D.M., 2011. *In silico* genomic subtraction guides development of highly accurate, DNA-based

diagnostics for *Ralstonia solanacearum* race 3 biovar 2 and blood disease bacterium. J. Gen. Plant Pathol. 77, 182–193. doi:10.1007/s10327-011-0305-2

Mahbou Somo Toukam, G., Cellier, G., Wicker, E., Guilbaud, C., Kahane, R., Allen, C., Prior, P., 2009. Broad diversity of *Ralstonia solanacearum* strains in Cameroon. Plant Dis. 93, 1123–1130.

Poussier, S., Prior, P., Luisetti, J., Hayward, C., Fegan, M., 2000. Partial Sequencing of the *hrpB* and Endoglucanase Genes Confirms and Expands the Known Diversity within the *Ralstonia solanacearum* Species Complex. Syst. Appl. Microbiol. 23, 479–486. doi:10.1016/S0723-2020(00)80021-1

Safni, I., Cleenwerck, I., De Vos, P., Fegan, M., Sly, L., Kappler, U., 2014. Polyphasic taxonomic revision of the *Ralstonia solanacearum* species complex: proposal to emend the descriptions of *Ralstonia solanacearum* and *Ralstonia syzygii* and reclassify current *R. syzygii* strains as *Ralstonia syzygii* subsp. *syzygii* subsp. nov., *R. solanacearum* phylotype IV strains as *Ralstonia syzygii* subsp. *indonesiensis* subsp. nov., banana blood disease bacterium strains as *Ralstonia syzygii* subsp. *celebesensis* subsp. nov. and *R. solanacearum* phylotype I and III strains as *Ralstonia pseudosolanacearum* sp. nov. Int. J. Syst. Evol. Microbiol. 64, 3087–3103. doi:10.1099/ijms.0.066712-0

Santana, B.G., Lopes, C.A., Alvarez, E., Barreto, C.C., Allen, C., Quirino, B.F., 2012. Diversity of Brazilian biovar 2 strains of *Ralstonia solanacearum*. J. Gen. Plant Pathol. 78, 190–200. doi:10.1007/s10327-012-0369-7

Schneider, K.L., Marrero, G., Alvarez, A.M., Presting, G.G., 2011. Classification of Plant Associated Bacteria Using RIF, a Computationally Derived DNA Marker. PLoS ONE 6, e18496. doi:10.1371/journal.pone.0018496

- Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195–197.
- Stulberg, M.J., Huang, Q., 2015. A TaqMan-Based Multiplex qPCR Assay and DNA Extraction Method for Phylotype IIB Sequevars 1&2 (Select Agent) Strains of *Ralstonia solanacearum*. *PLOS ONE* 10, e0139637. doi:10.1371/journal.pone.0139637
- Stulberg, M.J., Shao, J., Huang, Q., 2015. A multiplex PCR assay to detect and differentiate select agent strains of *Ralstonia solanacearum*. *Plant Dis.* 99, 333–341. doi:10.1094/PDIS-05-14-0483-RE
- Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi:10.1093/molbev/mst197
- Weller, S.A., Elphinstone, J.G., Smith, N.C., Boonham, N., Stead, D.E., 2000. Detection of *Ralstonia solanacearum* Strains with a Quantitative, Multiplex, Real-Time, Fluorogenic PCR (TaqMan) Assay. *Appl. Environ. Microbiol.* 66, 2853–2858. doi:10.1128/AEM.66.7.2853-2858.2000
- Wicker, E., Grassart, L., Coranson-Beaudu, R., Mian, D., Guilbaud, C., Fegan, M., Prior, P., 2007. *Ralstonia solanacearum* Strains from Martinique (French West Indies) Exhibiting a New Pathogenic Potential. *Appl. Environ. Microbiol.* 73, 6790–6801. doi:10.1128/AEM.00841-07

Figure legends

Fig. 1. Graphical representation of partial *egl* amplicons and trimmed products used by different research groups for sequevar analysis. Black lines indicate PCR products amplified by the

indicated primers, while gray lines immediately below are their trimmed products. The numbers on each line indicate basepair position relative to the 850-bp fragment of the phylotype IIB sequevar 3 strain Molk2, serving as the reference partial *egl* sequence (top line).

Fig. 2. Amplification of the 526-bp *egl* fragment using MJS-amp-egl primers from symptomatic plants infected with *R. solanacearum*. Lanes 1 and 10: 100-bp ladders; Lanes 2, 7, 11 and 14: DNA from healthy tomato, potato, tobacco and geranium, respectively. Lanes 3-6: tomato plants infected with strains UW349, UW551, GMI1000, and K60, respectively. Lanes 8-9: potato plants infected with strains UW551 and K60, respectively. Lanes 12 and 13: tobacco infected with K60. Lanes 15 and 16: geranium plants infected with strains UW551 and K60, respectively. Amplicons from lanes 5, 8, 12, and 15 were sequenced using MJS-seq-egl primers and analyzed by our *R. solanacearum* typing program. The numbers on the left side of the gels indicate the size (bp) of the corresponding DNA fragments in the DNA ladder.

Fig. 3. Decision tree of our *R. solanacearum* typing program for the 371 partial *egl* sequences used for validation. The nucleotides highlighted in gray are the identified SNPs from *egl* sequence comparisons. Numbers in parentheses refer to the 3 SNPs-containing sequences necessary to distinguish sequevars 1 and 2 from other sequevars in *R. solanacearum*. n: number of the *egl* sequences identified by the program at each designated step.

Fig. 4. Screenshots of the *Ralstonia solanacearum* typing program. The top shows the program ready for sequence input or after the “refresh” button was pressed. The bottom is an example of test results when a *R. solanacearum* sequevar 1 strain is identified.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

- 460
- 461 S1 Program file. An executable file for *Ralstonia solanacearum* typing program.
- 462
- 463 S2 Table. List of the 371 partial *egl* sequences used for validation of our *Ralstonia solanacearum*
- 464 typing program.

Table 1: Computer typing of *Ralstonia solanacearum* species-complex strains in our collection and isolated from artificially infected plants based on partial *egl* sequences determined or confirmed in this study.

GenBank Accession # or host	Strain name ^a	Phylotype/sequevar identified previously	Genospecies/sequevar determined by our typing program
KT861444	UW120 (S214)	II	<i>R. solanacearum</i> /1
KT861445	UW260	II	<i>R. solanacearum</i> /1
KT861446	UW425 (O249)	II	<i>R. solanacearum</i> /1
KT861447	UW449 (CIP259)	II	<i>R. solanacearum</i> /1
KT861448	UW492 (CIP302)	II	<i>R. solanacearum</i> /1
KT861449	UW501 (CIP181)	II	<i>R. solanacearum</i> /1
KT861450	UW552	II	<i>R. solanacearum</i> /1
KT861437	Pss1370	II	<i>R. solanacearum</i> /1
KT861438	Pss1475	II	<i>R. solanacearum</i> /1
KT861439	Pss1586	II	<i>R. solanacearum</i> /1
KT861426	CMR24 (CFBP7027)	II	<i>R. solanacearum</i> /1
KT861440	RE	II	<i>R. solanacearum</i> /1
KT861430	IVIA1602	II	<i>R. solanacearum</i> /1
KT861434	NCPPB4155	II	<i>R. solanacearum</i> /1
KT861433	NCPPB4153	II	<i>R. solanacearum</i> /1
KT861432	NCPPB2505	II	<i>R. solanacearum</i> /1
KT861431	NCPPB1584	II	<i>R. solanacearum</i> /1
KT861428	CSL Pr 3467	II	<i>R. solanacearum</i> /1
KT861429	CSL Pr 3468	II	<i>R. solanacearum</i> /1
KT861427	CLS Pr 1328	II	<i>R. solanacearum</i> /1
KT861442	UW9 (S147)	II	<i>R. solanacearum</i> /not 1 or 2
KT861441	Rs116	II	<i>R. solanacearum</i> /not 1 or 2
KT861435	Pss32	I	<i>R. pseudosolanacearum</i>
KT861443	UW119 (S213)	I	<i>R. pseudosolanacearum</i>
KT861436	Pss901	I	<i>R. pseudosolanacearum</i>
KT355485	Rs121	I/13	<i>R. pseudosolanacearum</i>
Tomato	GMI1000	I/18	<i>R. pseudosolanacearum</i>
Potato	UW551	II/1	<i>R. solanacearum</i> /1
Tobacco	K60	II/7	<i>R. solanacearum</i> /not 1 or 2
Geranium	UW551	II/1	<i>R. solanacearum</i> /1

^aStrains obtained from C. Allen, J.F. Wang, P. Prior, NCPPB, M.M. López, or J. Jones and characterized in Stulberg et al. (2015).

Figure(s) 2

Figure(s) 3

74 *Ralstonia solanacearum* typing program

phd1 file

Refresh

Reverse Complement

Paste

Enter Sequence

Results

Execute

74 *Ralstonia solanacearum* typing program

phd1 file

Refresh

Reverse Complement

Paste

Enter Sequence

Results

Execute