Gall-ID: tools for genotyping gall-causing phytopathogenic bacteria

Edward W Davis II, Alexandra J Weisberg, Javier F Tabima, Niklaus J. Grunwald, Jeff Chang

Understanding the population structure and genetic diversity of plant pathogens, as well as the effect of agricultural practices on pathogen evolution, are important for disease management. Developments in molecular methods have contributed to increase the resolution for accurate pathogen identification but those based on analysis of DNA sequences can be less straightforward to use. To address this, we developed Gall-ID, a web-based platform that uses DNA sequence information from 16S rDNA, multilocus sequence analysis and whole genome sequences to group disease-associated bacteria to their taxonomic units. Gall-ID was developed with a particular focus on gall-forming bacteria belonging to *Agrobacterium*, *Pseudomonas savastanoi*, *Pantoea agglomerans*, and *Rhodococcus*. Members of these groups of bacteria cause growth deformation of plants, and some are capable of infecting many species of field, orchard, and nursery crops. Gall-ID also enables the use of high-throughput sequencing reads to search for evidence for homologs of characterized virulence genes, and provides downloadable software pipelines for automating multilocus sequence analysis, analyzing genome sequences for average nucleotide identity, and constructing core genome phylogenies. Lastly, additional databases were included in Gall-ID to help determine the identity of other plant pathogenic bacteria that may be in microbial communities associated with galls or causative agents in other diseased tissues of plants. The URL for Gall-ID is http://gall-id.cgrb.oregonstate.edu/.

Gall-ID: tools for genotyping gall-causing phytopathogenic bacteria

Edward W. Davis II[1,2*], Alexandra J. Weisberg[1*], Javier F. Tabima[1], Niklaus J. Grünwald[1,2,3,4], and Jeff H. Chang[1,2,4]

[1]Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR, 97331, USA
[2]Molecular and Cellular Biology Program, Oregon State University, Corvallis, OR, 97331, USA
[3]Horticultural Crops Research Laboratory, USDA-ARS, Corvallis, OR, 97331, USA
[4]Center for Genome Research and Biocomputing, Oregon State University, Corvallis, OR, 97331, USA

*Equal contribution

Corresponding author
Jeff H. Chang
3096 Cordley Hall, Corvallis, OR, USA
changj@science.oregonstate.edu

ABSTRACT

Understanding the population structure and genetic diversity of plant pathogens, as well as the effect of agricultural practices on pathogen evolution, are important for disease management. Developments in molecular methods have contributed to increase the resolution for accurate pathogen identification but those based on analysis of DNA sequences can be less straightforward to use. To address this, we developed Gall-ID, a web-based platform that uses DNA sequence information from 16S rDNA, multilocus sequence analysis and whole genome sequences to group disease-associated bacteria to their taxonomic units. Gall-ID was developed with a particular focus on gall-forming bacteria belonging to *Agrobacterium*, *Pseudomonas savastanoi*, *Pantoea agglomerans*, and *Rhodococcus*. Members of these groups of bacteria cause growth deformation of plants, and some are capable of infecting many species of field, orchard, and nursery crops. Gall-ID also enables the use of high-throughput sequencing reads to search for evidence for homologs of characterized virulence genes, and provides downloadable software pipelines for automating multilocus sequence analysis, analyzing genome sequences for average nucleotide identity, and constructing core genome phylogenies. Lastly, additional databases were included in Gall-ID to help determine the identity of other plant pathogenic bacteria that may be in microbial communities associated with galls or causative agents in other diseased tissues of plants. The URL for Gall-ID is http://gall-id.cgrb.oregonstate.edu/.

INTRODUCTION

Diagnostics

Determining the identity of the disease causing pathogen, establishing its source of introduction, and/or understanding the genetic diversity of pathogen populations are critical steps for containment and treatment of disease. Proven methods for identification have been developed based on discriminative phenotypic and genotypic characteristics, including presence of antigens, differences in metabolism, or fatty acid methyl esters, and assaying based on polymorphic nucleotide sequences (Alvarez 2004). For the latter, polymerase chain reaction (PCR) amplification-based approaches for amplifying informative regions of the genome can be used. These regions should have broadly conserved sequences that can be targeted for amplification but the intervening sequences need to provide sufficient resolution to infer taxonomic grouping.

The 16S rDNA sequence is commonly used for identification (Stackebrandt & Goebel 1994). Because of highly conserved regions in the gene sequence, a single pair of degenerate oligonucleotide primers can be used to amplify the gene from a diversity of bacteria, and allow for a kingdom-wide comparison. In general, the sequences of the amplified fragments have enough informative polymorphic sites to delineate genera, but do not typically allow for more refined taxonomic inferences at the sub-genus level (Janda & Abbott 2007). Multilocus sequence analysis (MLSA) leverages the phylogenetic signal from four to ten genes to provide increased resolution, and can distinguish between species and sometimes sub-species (Wertz et al. 2003; Zeigler 2003). MLSA however, is more restricted than the use of 16S rDNA sequences and may not allow for comparisons between members of different genera. MLSA also requires more time

and effort to identify informative and taxon-specific genes as well as develop corresponding oligonucleotide primer sets.

Whole genome sequences can also be used. This is practical because of advances in next generation sequencing technologies. The key advantages of this approach is that availability of whole genome sequences obviates the dependency on *a priori* knowledge to provide clues on taxonomic association of a pathogen and the need to select and amplify marker genes. Also, whole genome sequences provide the greatest resolution in terms of phylogenetic signal, and sequences that violate assumptions of phylogenetic analyses (e.g., not vertically inherited) can be removed from studies to allow for robust analyses. Briefly, sequencing reads of genome(s) are compared to a high quality draft or finished reference genome sequence to identify variable positions between the genome sequences (Pearson et al. 2009). The positions core to the compared genome sequences are aligned and used to generate a phylogenetic tree. Alternatively, whole genome sequences can be used to determine average nucleotide identities (ANI) between any sufficiently similar pair, e.g., within the same taxonomic family, of genome sequences to determine genetic relatedness (Goris et al. 2007; Kim et al. 2014). ANI can be used to make taxonomic inferences, as a 95% threshold for ANI has been calibrated to those used to operationally define bacterial species based on 16S rDNA (> 94% similarity) and DNA-DNA hybridization (DDH; 70%) (Goris et al. 2007). Finally, the whole genome sequences can be analyzed to inform on more than just the identity of the causative agent and provide insights into mechanisms and evolution of virulence. However, a non-trivial trade-off is that processing, storing, and analyzing whole genome sequence data sets require familiarity with methods in computational biology.

84 ***Agrobacterium* spp.**

85 Members from several taxa of Gram-negative bacteria are capable of causing abnormal
86 growth of plants. Members of *Agrobacterium* are the most notorious causative agents of
87 deformation of plant growth. These bacteria have been classified according to various schemes
88 that differ in the phenotypic and genetic characteristics that were used. Its taxonomic
89 classification has been a subject of multiple studies (Young et al. 2001; Farrand et al. 2003;
90 Young 2003). Here, we will use the classification scheme that is based on disease phenotype and
91 more commonly encountered in the literature. Within *Agrobacterium* there are four recognized
92 groups of gall-causing bacteria. *A. tumefaciens* (also known as *Rhizobium radiobacter* (Young et
93 al. 2001), formerly *A. radiobacter,* genomovar G8 forms *A. fabrum* (Lassalle et al. 2011)) can
94 cause crown gall disease that typically manifests as tumors on roots or stem tissue (Gloyer 1934;
95 Kado 2014). *A. tumefaciens* can infect a wide variety of hosts and the galls can restrict plant
96 growth and in some cases kill the plant (Gloyer 1934; Schroth 1988). Other gall causing clades
97 include *A. vitis* (restricted to infection of grapevine), *A. rubi* (*Rubus* galls), and *A. larrymoorei*,
98 which is sufficiently different based on results of DNA-DNA hybridization studies to justify a
99 species designation (Hildebrand 1940; Ophel & Kerr 1990; Bouzar & Jones 2001). The genus
100 was also traditionally recognized to include hairy-root inducing bacteria belonging to *A.*
101 *rhizogenes*, as well as non-pathogenic biocontrol isolates belonging to *A. radiobacter* (Young et
102 al. 2001; Velázquez et al. 2010). Members of *Agrobacterium* are atypical in having multipartite
103 genomes, which in some cases include a linear replicon (Allardet-Servent et al. 1993).

104 A Ti (tumor-inducing) plasmid imparts upon members of *Agrobacterium* the ability to
105 genetically modify its host and cause dysregulation of host phytohormone levels and induce gall
106 formation (Sachs 1975). The Ti plasmid contains a region of DNA (T-DNA) that is transferred

and integrated into the genome of the host cell (Van Larebeke et al. 1974; Chilton et al. 1977; Thompson et al. 1988; Ward et al. 1988; Broothaerts et al. 2005). Conjugation is mediated by a type IV secretion system, encoded by genes located outside the borders of the T-DNA on the Ti plasmid (Thompson et al. 1988). Within the T-DNA are key genes that encode for auxin, cytokinin, and opine biosynthesis (Morris 1986; Binns & Costantino 1998). Expression of the former two genes in the plant leads to an increase in plant hormone levels to cause growth deformation whereas the latter set of genes encode enzymes for the synthesis of modified sugars that only organisms with the corresponding opine catabolism genes can use as an energy source (Bomhoff et al. 1976; Montoya et al. 1977). The latter genes are located outside the T-DNA borders on the Ti plasmid (Zhu et al. 2000).

Pseudomonas savastanoi

Pseudomonas savastanoi (formerly *Pseudomonas syringae* pv. *savastanoi,* (Gardan et al. 1992)) is the causal agent of olive knot disease, typically forming as aerial tumors on stems and branches. Phytopathogenicity of *P. savastanoi* is dependent on the *hrp/hrc* genes located in a pathogenicity island on the chromosome (Sisto et al. 2004). These genes encode for a type III secretion system, a molecular syringe that injects type III effector proteins into host cells that collectively function to dampen host immunity (Chang et al. 2014). Phytopathogenicity of *P. savastanoi* is also associated with the production of phytohormones. Indole-3-acetic acid may have an indirect role as a bacterial signaling molecule (Aragon et al. 2014). A cytokinin biosynthesis gene has also been identified on plasmids in *P. savastanoi* and strains cured of the plasmid caused smaller galls but were not affected in growth within the galls (Iacobellis et al. 1994; Bardaji et al. 2011).

130

131 ***Pantoea agglomerans***

132 *Pantoea agglomerans* (formerly *Erwinia herbicola*) is a member of the
133 Enterobacteriaceae family. *P. agglomerans* can induce the formation of galls on diverse species
134 of plants (Cooksey 1986; Burr et al. 1991; Opgenorth et al. 1994; DeYoung et al. 1998;
135 Vasanthakumar & McManus 2004). Phytopathogenicity is dependent on the pPATH plasmid
136 (Manulis & Barash 2003; Weinthal et al. 2007). This plasmid contains a pathogenicity island
137 consisting of an *hrp*/*hrc* cluster and operons encoding for the biosynthesis of cytokinins, indole-
138 3-acetic acid, and type III effectors (Clark et al. 1993; Lichter et al. 1995; Nizan et al. 1997; Mor
139 et al. 2001; Nizan-Koren et al. 2003; Barash & Manulis 2005; Barash et al. 2005; Barash &
140 Manulis-Sasson 2007). As is the case with *P. savastanoi*, mutants of the *hrp*/*hrc* genes abolish
141 pathogenicity whereas mutations in the phytohormone biosynthesis genes led to galls of reduced
142 size (Manulis et al. 1998; Mor et al. 2001; Nizan-Koren et al. 2003; Barash & Manulis-Sasson
143 2007).

144

145 ***Rhodococcus* spp.**

146 Gram-positive bacteria within the *Rhodococcus* genus can cause leafy gall disease to over
147 100 species of plants (Putnam & Miller 2007). The phytopathogenic members of this genus
148 belong to at least two genetically distinct groups of bacteria, with *R. fascians* (formerly
149 *Corynebacterium fascians*) being the original recognized species (Goodfellow 1984; Creason et
150 al. 2014a). It is suggested that *R. fascians* upsets levels of phytohormones of the plant to induce
151 gall formation. However, unlike Ti plasmid-carrying *Agrobacterium*, it is hypothesized that *R.*
152 *fascians* directly synthesizes and secretes the cytokinin phytohormone (Stes et al. 2013; Creason

et al. 2014b). Phytopathogenicity is most often associated with a linear plasmid, which carries a cluster of virulence loci, *att*, *fasR*, and *fas* (Creason et al. 2014b). The functions for the translated products of *att* are unknown but the sequences have homology to proteins involved in amino acid and antibiotic biosynthesis (Maes et al. 2001). The *fasR* gene is necessary for full virulence; the gene encodes a putative transcriptional regulator (Temmerman et al. 2000). Some of the genes within the *fas* operon are necessary for virulence, as many of the *fas* genes encode proteins with demonstrable functions in cytokinin metabolism (Crespi et al. 1992). In rare cases, the virulence loci, or variants therein, are located on the chromosome (Creason et al. 2014b).

We developed Gall-ID to aid in determining the genetic identity of gall-causing members of *Agrobacterium*, *Pseudomonas*, *Pantoea*, and *Rhodococcus*. Users can provide sequences from 16S rDNA or gene sets used in MLSA, and Gall-ID will automatically query curated databases and generate phylogenetic trees to group the query isolate of interest and provide estimates of relatedness to previously characterized species and/or genotypes. Users can also submit short reads from whole genome sequencing projects to query curated databases to search for evidence for known virulence genes of these gall-causing bacteria. Finally, users can download tools that automate the analysis of whole genome sequencing data to infer genetic relatedness based on MLSA, average nucleotide identity (ANI), or single nucleotide polymorphisms (SNPs).

MATERIAL AND METHODS

Website framework and bioinformatics tools

The Gall-ID website and corresponding R shiny server backend are based on the Microbe-ID platform (Tabima et al. 2016) but includes major additions and modifications: Auto MLSA, Auto ANI, BLAST with MAFFT, and the WGS Pipeline. The MLSA framework

website was extended to support building Neighbor-Joining trees using incomplete distance matrices (NJ*) using the function njs() in the R package PHYLOCH (Paradis et al. 2004). The MLSA framework was also modified to use the multiple sequence alignment program MAFFT using the R package PHYLOCH (Katoh & Toh 2008; Heibl 2013; Katoh & Standley 2013). This allows user-submitted sequences to be added to pre-existing sequence alignments using the MAFFT "--add" function, to dramatically reduce the computational time required for analysis.

The server hosting the Gall-ID tools is running Centos Linux release 6.6, MAFFT version 7.221, SRST2 version 0.1.5, Bowtie 2 version 2.2.3, and Samtools version 0.1.18. Gall-ID uses R version 3.1.2 with the following R packages: Poppr version 1.1.0.99 (Kamvar et al. 2014), Ape version 3.1-1 (Paradis et al. 2004), PHYLOCH version 1.5-5, and Shiny version 0.8.0.

The Auto MLSA tool was developed previously (Creason et al. 2014a). Briefly, Auto MLSA does the following: BLAST (either TBLASTN or BLASTN) to query NCBI user-selectable databases and/or local databases and retrieve sequences, filter out incomplete sets of gene sequences, align gene sequence individually, concatenate aligned gene sequences, determine the best substitution model (for amino acid sequences), filter out identical sequences, append key information to sequences, and generate a partition file for RAxML (Stamatakis 2014). Auto MLSA also has the option of using Gblocks to trim alignments (Castresana 2000). Auto MLSA was modified to use the NCBI E-utilities, implemented in BioPerl, to associate accession numbers with taxon IDs, species names, and assembly IDs (Stajich 2002). For organisms without taxon identifiers, Auto MLSA will attempt to extract meaningful genus and species information from the NCBI nucleotide entry. Gene sequences are linked together using assembly IDs, which allows for genomes with multiple chromosomes to be compared, without having to rely on potentially ambiguous organism names. When assembly IDs are unavailable,

whole genome sequences are linked using the four letter WGS codes, and, as a last resort, sequences will be associated using their nucleotide accession number. The disadvantage of using the latter approach is that organisms with multiple replicons, each with its own accession number, will be excluded from analysis. Auto MLSA is available for download from the Gall-ID website. Detailed instructions for using the tools are provided.

The Auto ANI script automates the calculations of ANI for all pairwise combinations for any number of input genome sequences. Each of the supplied genome sequences are chunked into 1020 nt fragments and used as queries in all possible reciprocal pairwise BLAST searches. Parameters for genome chunk size, percent identity, and percent coverage have default values set according to published guidelines but can be changed by the user (Goris et al. 2007; Creason et al. 2014a). BLAST version 2.2.31+ was used with recommended settings and previously described in Creason et al. (2014): –task blastn –dust no –xdrop_gap 150 – penalty −1 –reward 1 –gapopen 5 –gapextend 2 (Goris et al. 2007). BLAST hits above the user-specified cut-offs (30% identity, 70% coverage, by default) are averaged to calculate the pairwise ANI values.

BLAST+ version 2.2.27 has been tested and works, but this version is currently unsupported. Versions 2.2.28-2.2.30 of BLAST+ have an undocumented bug that prevents efficient filtering using -max_hsps and -max_target_seqs and precludes their use in ANI calculation. Hence BLAST 2.2.31+ is the preferred and recommended version.

Sequences downloaded from NCBI are linked using assembly IDs. All accession types from NCBI are supported, assuming accession numbers are provided in the header line of the FASTA file. Locally generated genome sequences are also supported, in FASTA format, provided they follow the specified header format listed in the user guide. Alternatively, an

auxiliary script is provided to rename headers within user-generated FASTA files to the supported format.

The WGS Pipeline was written in bash shell script and Perl. Paired Illumina sequencing reads located in the "reads" folder of the pipeline are processed in pairs. The program SMALT (Ponstingl, 2013) is used to align reads to a reference genome and produce CIGAR format output files (Ponstingl 2013). The SSAHA_pileup program converts the CIGAR format files into individual pileup files (Ning et al. 2001). The pileup output is then combined with any additional supplied pre-computed pileup files and used to produce a core alignment of sites shared by 90% of the represented isolates. The optional "remove_recombination.sh" script runs the program Gubbins (Croucher et al. 2014) to remove sites identified as potentially acquired by recombination. Finally, the program RAxML is used to produce a maximum-likelihood phylogenetic tree with non-parametric bootstrap support (Stamatakis 2014). By default 20 maximum likelihood tree searches are performed, and the "autoMRE" criterion is used to determine the number of non-parametric bootstrap replicates.

The WGS Pipeline test analysis was performed and benchmarked using 10 cores of a cluster server running Centos Linux release 6.6 and containing four AMD Opteron™ 6376 2.3 Ghz processors (64 cores total) and 512 GB of RAM (Table 2). The versions of the tools used in tests of this pipeline were Perl version 5.10.1, SMALT version 0.7.6, SSAHA_pileup version 0.6, Gubbins version 1.1.2, and RAxML version 8.1.17. The default parameters for WGS Pipeline were used (20 ML search trees, "autoMRE" cutoff for bootstrap replicates) with the exception that the maximum-allowed percentage gaps in the Gubbins recombination analysis was increased to 50% in order to retain strain D188. The WGS Pipeline scripts were also

243 modified to not ask for user input on the command line in order to run in a Sun Grid Engine
244 (SGE) cluster environment.
245 Vir-Search uses the program SRST2, which employs Bowtie 2 and Samtools, with the "--
246 gene_db" function to align the reads to custom databases of the virulence genes (Li et al. 2009;
247 Inouye et al. 2012; Langmead & Salzberg 2012; Inouye et al. 2014). The identity of the virulence
248 genes that the reads input by the user align to, the read coverage and depth, and the name of the
249 strain corresponding to the most similar allele are parsed from the SRST2 output and reported to
250 the user as a static webpage. Users are emailed a link to results once the analysis is complete.
251 The submitted sequencing reads are deleted from the server immediately after completion, and
252 results are available only to those with a direct link to the results webpage.
253
254 **Datasets**
255 The 16S and MLSA gene sequences were downloaded from the genome sequences of the
256 following reference strains: *Agrobacterium* strain C58, *Rhodococcus* strain A44a, *P. savastanoi*
257 pv. *phaseolicola* 1448A, and *P. agglomerans* strain LMAE-2, *C. michiganensis* subsp.
258 *nebraskensis* NCPPB 2581, *D. dadantii* strain 3937, *P. atrosepticum* strain 21A, *R.*
259 *solanacearum* strain GMI1000, *X. oryzae* pv. *oryzicola* strain CFBP2286, and *X. fastidiosa*
260 subsp. *fastidiosa* GB514 (NCBI assembly ID: GCF_000092025.1, GCF_000760735.1,
261 GCF_000012205.1, GCF_000814075.1, GCF_000355695.1, GCF_000147055.1,
262 GCF_000740965.1, GCF_000009125.1, GCF_001042735.1, and GCF_000148405.1,
263 respectively). The gene sequences were used as input for the Auto MLSA tool in BLAST
264 searches carried out against complete genome sequences in the NCBI non-redundant (nr) and
265 whole genome sequence (wgs) databases. The Auto MLSA parameters were: minimum query

coverage of 50% (90% for the 16S plant pathogen dataset) and e-value cutoffs of 1e-5 for nr and 1e-50 for wgs. BLAST searches were limited to the genus for the bacteria of interest, with the exceptions of *Agrobacterium*, which was limited to *Rhizobiaceae*, and *P. savastanoi,* which was limited to the *P. syringae* group. BLAST searches were completed in August of 2015. The Auto MLSA tool uses MAFFT aligner to produce multiple sequence alignments for each gene (Katoh & Standley 2013). The Gblocks trimmed alignment output of Auto MLSA was not used because Gall-ID aligns user-submitted gene sequences to each full gene alignment (Castresana 2000).

Bacterial strains, growth conditions, nucleic acid extraction, and genome sequencing

Cultures of *Agrobacterium* were grown overnight in Lysogeny Broth (LB) media at 28°C, with shaking at 250 rpm (Table 3). Cells were pelleted by centrifugation and total genomic DNA was extracted using a DNeasy Blood and Tissue kit (Qiagen, Venlo, Netherlands). DNA was quantified using a QuBit Fluorometer (Thermo Fisher, Eugene, Oregon) and libraries were prepared using an Illumina Nextera XT DNA Library Prep kit, according to the instructions of the manufacturer, with the exception that libraries were normalized based on measurements from an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA). Each library was assigned an individual barcode using an Illumina Nextera XT Index kit. Libraries were multiplexed and sequenced on an Illumina MiSeq to generate 300 bp paired-end reads. Sequencing was done in the Center for Genome Research and Biocomputing Core Facility (Oregon State University, Corvallis, OR). Sickle was used to trim reads based on quality (minimum quality score cutoff of 25, minimum read length 150 bp after trimming) (Joshi & Fass 2011). Read quality was assessed prior to and after trimming using FastQC (Andrews 2010). Paired reads for each library were *de novo* assembled using Velvet version 1.2.10 with the short

289 paired read input option ("-shortPaired"), estimated expected coverage ("-exp_cov auto"), and
290 default settings for other parameters (Zerbino & Birney 2008). Genome sequences were
291 assembled using a range of input hash lengths (k-mer sizes), and the final assembly for each
292 isolate was identified based on those with the best metrics for the following parameters: total
293 assembly length (5.0~7.0 Mb), number of contigs, and N50. Paired reads for each library were
294 error corrected and assembled using SPAdes versions 3.6.2 and 3.7.0, with the careful option ("--
295 careful") and kmers 21, 33, 55, 77, and 99. Scaffolds shorter than 500bp and with coverage less
296 than 5X were removed from the SPAdes assemblies prior to analysis.
297

298 **RESULTS AND DISCUSSION**

299 Gall-ID (http://gall-id.cgrb.oregonstate.edu/) is based on the Microbe-ID platform and
300 uses molecular data to determine the identity of plant pathogenic bacteria (Tabima et al. 2016).
301 Gall-ID is organized into modules shown as tabs that allow users to choose from one of four
302 options for analyzing data (Figure 1).
303

304 **Gall Isolate Typing**

305 The "Gall Isolate Typing" tab provides online tools to use molecular data, either 16S or
306 sequences of marker genes used for MLSA, to group isolates of interest into corresponding
307 taxonomic units that include gall-causing pathogens. Users must first select the appropriate
308 taxonomic group, *Agrobacterium*, *Pseudomonas*, *Pantoea*, or *Rhodococcus* for comparison. For
309 some of these taxonomic groups, multiple gene sets used in MLSA are available, and the user
310 must therefore select the appropriate set for analysis. FASTA formatted gene sequences are
311 input, and after selecting the appropriate options for alignment and tree parameters, a

phylogenetic tree that includes the isolate of interest is generated and displayed. The tree parameters include choice of distance matrix, tree generating algorithm (Neighbor-Joining or UPGMA), and number of bootstrap replicates. A sub-clade of the tree containing only the isolate of interest and its nearest sister taxa is displayed to the right of the full tree. The tree can be saved as a Newick file or as a PDF. An example sequence from *Agrobacterium* can be loaded by clicking the "Demo" button located in the Agro-type tab.

Phytopath-Type

The "Phytopath-Type" tab provides online tools for the analysis of other non-gall-causing pathogens important in agriculture (Mansfield et al. 2012). This tool is similar in function to the "Gall Isolate Typing" tools, except it is not limited to a single taxon of pathogen. A database of 16S rDNA sequences from genera of important bacterial phytopathogens (*Pseudomonas syringae* group, *Ralstonia*, *Agrobacterium*, *Rhodococcus*, *Xanthomonas*, *Pantoea*, *Xylella*, *Dickeya*, *Pectobacterium*, and *Clavibacter*) is available for associating a bacterial pathogen to its genus. Additionally, for *Clavibacter*, *Dickeya*, *Pectobacterium*, *Ralstonia*, *Xanthomonas*, and *Xylella*, the user can use MLSA to genotype isolates of interest. As is the case with Gall Isolate Typing, a phylogenetic tree will be generated and displayed, associating the isolate of interest to the most closely related genotype in the curated databases.

Vir-Search

The "Vir-Search" tab provides an online tool for using user-input read sequences of a genome to search for the presence of homologs of known virulence genes. Users select a taxonomic group (*Agrobacterium*, *P. savastanoi*, *P. agglomerans*, or *Rhodococcus*) to designate

335 the set of virulence genes to search against. Users also determine a minimum percent gene
336 coverage and maximum allowed percent identity divergence, and upload single or paired read
337 files in FASTQ format. The user-supplied read sequences are then aligned to the chosen
338 virulence gene dataset on the Gall-ID server. Once the search is complete, a link to the final
339 results is sent to a user-provided email address. Results display the percent coverage of the
340 virulence genes and the percent similarity of the covered sequences. If the query identifies
341 multiple alleles of virulence genes from different sequenced strains, the Vir-Search tool will
342 report the strain name associated with the best-mapped allele. User-submitted data and results are
343 confidential and submitted sequencing reads are deleted from the Gall-ID server upon
344 completion of the analysis.
345
346 **Databases of Gall-ID**
347 A key component of the tools associated with the aforementioned tabs is the manually
348 curated databases of gene sequences. The literature was reviewed to identify validated taxon-
349 specific sets of genes for MLSA of taxa with gall-causing bacteria as well as other pathogens that
350 affect agriculture (Table 1) (Sarkar & Guttman 2004; Hwang et al. 2005; Castillo & Greenberg
351 2007; Alexandre et al. 2008; Young et al. 2008; Delétoile et al. 2009; Kim et al. 2009; Adékambi
352 et al. 2011; Jacques et al. 2012; Parker et al. 2012; Marrero et al. 2013; Pérez-Yépez et al. 2014;
353 Tancos et al. 2015). The sequences for the corresponding genes were subsequently extracted
354 from the whole genome sequences of reference strains. Auto MLSA was employed to use the
355 gene sequences as queries in BLAST searches. Auto MLSA is based on a previously developed
356 set of Perl scripts to automate retrieving, filtering, aligning, concatenating, determining of best
357 substitution models, appending of key identifiers to sequences, and generating files for tree

construction (Creason et al. 2014a). Gene sets in which there were less than 50% query sequence coverage for all of the genes were excluded to ensure that the databases contained only taxonomically informative sequences. Each gene set database was manually checked for duplicate strains, large gaps in gene sequences, poor sequence alignment, and mis-annotated taxonomic information. Each of the MLSA databases used in the Gall-ID tools is also available for download on the “Database Downloads" tab of the Gall ID website. The 16S rDNA databases were populated in a similar manner, with one small exception. For the Phytopath-Type tool, the sequence of the 16S rRNA-encoding gene from C58 of *Agrobacterium* was used as a query to retrieve corresponding sequences from 345 isolates distributed across the different genera of plant pathogenic bacteria.

To populate the database for virulence genes, the literature was reviewed to identify genes with demonstrably necessary functions for the pathogenicity of *Agrobacterium* spp., *R. fascians*, *P. savastanoi*, and *P. agglomerans* (Thompson et al. 1988; Ward et al. 1988; Lichter et al. 1995; Nizan et al. 1997; Manulis et al. 1998; Zhu et al. 2000; Maes et al. 2001; Vereecke et al. 2002; Nizan-Koren et al. 2003; Sisto et al. 2004; Nissan et al. 2006; Barash & Manulis-Sasson 2007; Matas et al. 2012). The gene sequences were downloaded from corresponding type strains in the NCBI nucleotide (nr) database or from nucleotide sequences in the NCBI nr database. The downloaded virulence gene sequences were then used as input for the Auto MLSA tool to retrieve sequenced alleles from other isolates of the same taxa. The downloaded alleles were manually inspected to ensure only pathogenic strains were represented. The database was formatted for SRST2.

Whole Genome Analysis

381 The analyses of whole genome sequence datasets can be computationally intensive,
382 which is prohibitive for online tools. Therefore, the "Whole Genome Analysis" tab provides
383 downloadable software pipeline tools for users to employ their institutional infrastructure or a
384 cloud computing service to analyze whole genome sequencing reads (Illumina HiSeq or MiSeq).
385 There are two options in this tab, the first, "WGS Pipeline: Core Genome Analysis," provides a
386 download link and instructions for using the WGS Pipeline tool to generate a phylogeny based
387 on the core genome sequence or core set of single nucleotide polymorphisms (SNPs). The
388 second option, “Auto ANI: Average Nucleotide Identity Analysis," provides a download link for
389 the Auto ANI tool and detailed instructions for its use in calculating all possible pairwise ANI
390 within a set of genome sequences.
391 The WGS Pipeline is a set of scripts that automates the use of sequences from Illumina-
392 based paired reads derived from whole genome sequencing projects to determine core genome
393 sequences or core SNPs and generate phylogenetic trees (Figure 2). This pipeline uses SMALT
394 and SSAHA2 pile-up pipeline to align sequencing reads to an indexed reference genome
395 sequence and generate a pileup file, respectively (Ning et al. 2001; Ponstingl 2013). The WGS
396 Pipeline then combines the pileup files along with other pre-computed pileup files to derive a
397 core genome alignment defined based on regions that are shared between at least 90% of the
398 compared genome sequences. Users have the option of using Gubbins to remove regions that are
399 flagged as potentially derived from recombination (Croucher et al. 2014). Invoking Gubbins will
400 also remove all non-polymorphic sites from the alignments, thus yielding a SNP alignment that
401 is based only on polymorphic sites that are identified as vertically inherited and shared between
402 at least 90% of the compared genome sequences. Finally, the user can use either the core genome

403 sequence or core SNP alignment and RAxML to generate a maximum likelihood (ML)
404 phylogeny (Stamatakis 2014).

405 Users must place their input files in the correspondingly named folders in order to run the
406 WGS pipeline. The pipeline down weights reads with a Q score of < 30, requires a minimum
407 depth of 12 and relies on a minimum threshold of 75% for consensus base calling. Users
408 concerned with sequencing quality may, prior to running the WGS pipeline, run programs such
409 as FastQC, Trimmomatic, Sickle, and/or BBDuk to filter reads based on quality threshold
410 (Andrews 2010; Joshi & Fass 2011; Bolger et al. 2014; Bushnell 2016). Paired read sequences
411 for each genome are read from the "reads" folder, while a SMALT index named as "reference"
412 and placed in the "index" folder will be used as a reference to align to. Identifiers are taken from
413 the prefix of the read pair file names and used to name the output pileup files and taxa in the
414 phylogeny. The read pair file names must have the suffixes ".1.fastq" and ".2.fastq" for files with
415 forward and reverse read sequences, respectively. The read sequences must be in FASTQ format
416 and because of requirements of the SMALT program, each paired read name must end in ".p1k"
417 and ".q1k" for forward and reverse reads, respectively. If the input read sequences are not in the
418 proper format, the user may run the included optional script "prepare_for_pileup.sh" to format
419 read names. If the user has pre-computed SMALT pileup files prepared using the same SMALT
420 index, the files may be placed in the "pileup" folder and will also be included in the analysis. The
421 user may be prompted to input the length of the inserts for each sequencing library. Users also
422 have the option of changing the number of ML searches or non-parametric bootstrap replicates
423 when building a phylogeny (default values are 20 ML searches, autoMRE cutoff criterion for
424 bootstrap replicates).

Pre-built SMALT indices for reference genome sequences from strain C58 of *Agrobaterium* and strain A44a of *Rhodococcus*, as well as pre-computed pileup files for 17 publicly available *Rhodococcus* genome sequences, are available for download on the Gall-ID website. Detailed usage instructions and download links for the pipeline scripts are located in the "WGS Pipeline: Core Genome Analysis" tool in the "Whole Genome Analysis" tab of Gall-ID.

Previously developed scripts for ANI analysis were rewritten and named Auto ANI. The current version of these scripts alleviates dependencies on our institutional computational infrastructure and increases the scalability of analyses (Creason et al. 2014a). Results are saved in a manner that enables analyzing additional genomes without having to re-compute ANI values for previously calculated comparisons. All BLAST searches are done in a modular manner and can be modified to run on a computer cluster with a queuing system such as the Sun Grid Engine. There are no inherent restrictions on the numbers of pairwise comparisons that can be performed. The data are output as a tab delimited matrix of all pairwise comparisons and can also easily be sorted and resorted based on any reference within the output. Additionally, genome sequences with evidence for poor quality assemblies can be easily filtered out. A distance dendrogram based on ANI divergence can also be generated; a python script is available for download (Chan et al. 2012; Creason et al. 2014a).

Validation of tools available in Gall-ID

We validated the efficacy of the online tools available from the Gall Isolate Typing, and Vir-Search tabs. DNA from 14 isolates were prepared, barcoded, and sequenced on an Illumina MiSeq (Table 3). Of these isolates, the identities of 11 were previously verified as *Agrobacterium*. The remaining three were associated with plant tissues showing symptoms of

crown gall disease but were not tested or had results inconsistent with being a pathogenic member of the *Agrobacterium* genus. The reads were trimmed for quality and first *de novo* assembled within each library using the Velvet assembler (Zerbino & Birney 2008). The 16S gene sequences were identified and extracted from the assemblies and used as input for the Agro-type tool. The 16S gene sequences from each of the 11 isolates originally typed as *Agrobacterium* clustered accordingly; isolate 13-2099-1-2 is shown as an example (Figure 3A). The 16S sequence from isolates AC27/96, AC44/96, and 14-2641 were more distant from the 16S sequences of *Agrobacterium* (Table 3). The isolates AC27/96 and AC44/96 grouped more closely with various *Rhizobium* species, while subsequent analysis using the Phytopath-Type tool suggested isolate 14-2641 was more closely related to members of *Erwinia*, *Dickeya*, and *Pectobacterium* (Table 3, Supplementary Figure 1). A search against the NCBI nr database revealed similarities to members of *Serratia*.

The trimmed read sequences were used as input for the Vir-Search tool as an additional step to confirm the identity of these isolates. Paired read sequences for each of the 14 isolates were individually uploaded to the Gall-ID server. The *Agrobacterium* virulence gene database was selected, with the minimum gene length coverage set to 80% and maximum allowed sequence divergence set to 20%. The time for each Vir-Search analysis ranged from 2-5 minutes. Results suggested that the genome sequences for nearly all of the *Agrobacterium* isolates had homologs of virulence genes demonstrably necessary for pathogenicity by *Agrobacterium*, while the genome sequences for the isolates AC27/96, AC44/96, and 14-2641 did not (Figure 3B, data for isolate 13-2099-1-2 shown). Contrary to the results from molecular diagnostics tests, the reads from isolate 13-626 failed to align to any virulence genes except for two (*nocM*, *nocP*) involved in nopaline transport. This isolate had the fewest number of useable sequencing reads

and the highest number of contigs compared to the others, and results could have been a consequence of a poor assembly of the Ti plasmid.

Indeed, the qualities of the 14 assemblies were highly variable, likely reflecting the multi-partite structure of the agrobacterial genomes, presence of a linear replicon, and/or variation in depth of sequencing. We therefore used SPAdes v. 3.6.2 to *de novo* assemble each of the genome sequences, with the exception of isolate 14-2641 (Bankevich et al. 2012). The total sizes of the assemblies were similar to those generated using Velvet and the qualities of the assemblies were high. But assemblies generated using SPAdes had proliferations in errors with palindromic sequence that appeared to be unique to isolates expected to have linear replicons. We informed the developers of the SPAdes software who immediately resolved the issue in SPAdes 3.7.0. Inspection of the summary statistics of the assemblies derived using this latest version of SPAdes suggested that relative to Velvet-based assemblies, there were improvements to all assemblies, witht the most dramatic to those with the lowest read coverage (Supplementary Table 1, Supplementary Figure 2). To further verify the quality of assemblies generated using SPAdes 3.7.0, we used Mauve to align Velvet and SPAdes assembled genome sequences of isolate 13-626 to the finished genome sequence of the reference sequence of *A. radiobacter* K84 (Darling et al. 2004; Slater et al. 2009). The SPAdes-based assembly was superior in being less fragmented and we were able to elevate the quality of the assembly from “unusable” to “high quality” (Supplementary Figures 2 and 3). Therefore, there is greater confidence in concluding that isolate 13-626 lacks the *vir* genes and T-DNA sequence. It does however have an ~200 kb plasmid sequence which encodes *nocM* and *nocP*; this contig also encodes sequences common to replication origins of plasmids. We therefore suggest that because an isolate from the same pear

493 gall sample originally tested positive for *virD2*, we mistakenly sequenced a non-pathogenic
494 isolate.

495 To validate the WGS Pipeline tool of Gall-ID, Illumina paired end read sequences
496 derived from previously generated genome sequences of 20 *Rhodococcus* isolates were used to
497 construct a phylogeny based on SNPs (Creason et al. 2014a). Using default parameters, the entire
498 process, from piling up reads to generating the final phylogenetic tree, took 16 hours (Table 2).
499 A total of 855,355 sites (out of a total of 5,947,114 sites in the A44a reference sequence) were
500 shared in at least 18 of the 20 *Rhodococcus* genome sequences. Of the shared sites, 177,961 sites
501 were polymorphic, of which 3,142 were removed because they were identified as potentially
502 acquired by recombination. The final core SNP alignment was therefore represented by 174,819
503 polymorphic sites and used to construct a maximum likelihood tree (Figure 4). Most of the nodes
504 were well supported, with all exceeding 68% bootstrap support and most having 100% support.
505 The topology of the tree was consistent with that derived from a multi-gene phylogeny (Creason
506 et al. 2014a). As previously reported, the 20 isolates formed two well-supported and distinct
507 clades, and could explain the relatively low number of shared SNPs. The substructure that was
508 previously observed in clade I was also evident in the ML tree based on core SNPs. The one
509 noticeable difference between the trees was that the tips of the tree based on the core SNPs had
510 substantially more resolution, and in particular, revealed a greater genetic distance between
511 isolates A76 and 05-339-1, than previously appreciated based on the multi-gene phylogeny.

512 The amount of time to run Auto ANI was determined by comparing genome assemblies
513 of the same 20 *Rhodococcus* isolates. The entire process was completed in five hours.

514

515 **Conclusions**

Gall-ID provides simplified and straightforward methods to rapidly and efficiently characterize gall-causing pathogenic bacterial isolates using Sanger sequencing or Illumina sequencing. Though Gall-ID was developed with a particular focus on these types of bacteria, it can be used for some of the more common and important agricultural bacterial pathogens. Additionally, the downloadable tools can be used for any taxa of bacteria, regardless of whether or not they are pathogens.

ACKNOWLEDGEMENTS

We thank Melodie Putnam for providing the 14 bacterial isolates and for critical reading of the manuscript. We thank Dr. Pankaj Jaiswal for organizing and inviting us to participate in the STEM DNA Biology and Bioinformatics summer camps (Oregon State University). Camp participants, Ana Bechtel, Mason Hall, Reagan Hunt, Pranav Kolluri, Benjamin Phelps, Joshua Phelps, Aravind Sriram, Megan Thorpe, and eight others, prepared genomic DNA and libraries for whole genome sequencing and analyzed the data. We thank Charlie DuBois of Illumina for providing kits for library preparation as well as sequencing, and Mark Dasenko, Matthew Peterson, and Chris Sullivan of the Center for Genome Research and Biocomputing for sequencing, data processing, and computing services. Finally, we thank the Department of Botany and Plant Pathology for supporting the computational infrastructure. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the U. S. Department of Agriculture or National Science Foundation.

538 **FIGURE LEGENDS**

539 **Figure 1. Overview of Gall-ID diagnostic tools.** DNA sequence information can be used to
540 reveal the identity of the causative agent (unknown isolate) of disease. Tools associated with
541 "Gall Isolate Typing" and “Phytopath-type” use 16S rDNA or pathogen-specific MLSA gene
542 sequences to infer the identity of the isolate by comparing the sequences to manually curated
543 sequence databases. Tools associated with “Whole Genome Analysis” and “Vir-Search” use
544 Illumina short sequencing reads to characterize pathogenic isolates. The former tab provides
545 downloadable tools to infer genetic relatedness based on SNPs (WGS Pipeline) or average
546 nucleotide identity (Auto ANI). The “Vir-Search” tab provides an on-line tool to quickly map
547 short reads against a database of sequences of virulence genes.
548
549 **Figure 2. Flowchart for the WGS Pipeline.** Scripts and the programs that each script runs are
550 boxed and presented along the left. The logic flow of the WGS Pipeline tool is presented along
551 the right. Rectangles with rounded corners = inputs and outputs; boxes outlined in red =
552 processes. The inputs, outputs, and processes are matched to the corresponding script and
553 program.
554
555 **Figure 3. Validation of the Agro-type and Vir-Search tools. A)** An unrooted Neighbor Joining
556 phylogenetic tree based on 16s rDNA sequences from *Agrobacterium* spp. The 16S rDNA
557 sequence was identified and extracted from the genome assembly of *Agrobacterium* isolate 13-
558 2099-1-2 and analyzed using the tool available in the Agro-type tab. The isolate is labeled in red,
559 as “query_isolate”; inset shows the clade that circumscribes the isolate. **B)** Screenshot of output
560 results from Vir-Search. Paired 2x300 bp MiSeq short reads from *Agrobacterium* isolate 13-
561 2099-1-2 were analyzed using the Vir-Search tool in Gall-ID. Reference virulence gene
562 sequences that were aligned are indicated with a green plus ("+") icon and the lengths and depths
563 of the read coverage are reported (must exceed user-specified cutoffs, which were designated as
564 90% minimum coverage and 20% maximum sequence divergence). Virulence genes that failed
565 to exceed user-specific cutoffs for read alignment parameters are indicated with a red "X".
566 Virulence genes are grouped into categories based on their function in virulence.
567
568 **Figure 4. Maximum likelihood tree based on vertically inherited polymorphic sites core to**
569 **20 *Rhodococcus* isolates.** WGS Pipeline was used to automate the processing of paired end short
570 reads from 20 previously sequenced *Rhodococcus* isolates, and generate a maximum likelihood
571 unrooted tree. Sequencing reads were aligned, using *R. fascians* strain A44a as a reference.
572 SNPs potentially acquired via recombination were removed. The tree is midpoint-rooted. Scale
573 bar = 0.05 average substitutions per site; non-parametric bootstrap support as percentages are
574 indicated for each node. Major clades and sub-clades are labeled in a manner consistent with
575 previous labels.
576

577

REFERENCES

Adékambi T, Butler RW, Hanrahan F, Delcher AL, Drancourt M, and Shinnick TM. 2011. Core gene set as the basis of multilocus sequence analysis of the subclass *Actinobacteridae*. *PLoS One* 6:e14792. 10.1371/journal.pone.0014792

Alexandre A, Laranjo M, Young JP, and Oliveira S. 2008. *dnaJ* is a useful phylogenetic marker for *Alphaproteobacteria*. *International Journal of Systematic and Evolutionary Microbiology* 58:2839-2849. 10.1099/ijs.0.2008/001636-0

Allardet-Servent A, Michaux-Charachon S, Jumas-Bilak E, Karayan L, and Ramuz M. 1993. Presence of one linear and one circular chromosome in the *Agrobacterium tumefaciens* C58 genome. *Journal of Bacteriology* 175:7869-7874.

Alvarez AM. 2004. Integrated approaches for detection of plant pathogenic bacteria and diagnosis of bacterial diseases. *Annual Review of Phytopathology* 42:339-366. doi:10.1146/annurev.phyto.42.040803.140329

Andrews S. 2010. FastQC A quality control tool for high throughput sequence data. Available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Aragon IM, Perez-Martinez I, Moreno-Perez A, Cerezo M, and Ramos C. 2014. New insights into the role of indole-3-acetic acid in the virulence of *Pseudomonas savastanoi* pv. *savastanoi*. *FEMS Microbiol Letters* 356:184-192. 10.1111/1574-6968.12413

Bankevich A, Nurk S, Antipov D, Gurevich A, Dvorkin M, Kulikov A, Lesin V, Nikolenko S, Pham S, Prjibelski A, Pyshkin A, Sirotkin A, Vyahhi N, Tesler G, Alekseyev M, and Pevzner P. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19:455-477. 10.1089/cmb.2012.0021

Barash I, and Manulis S. 2005. Hrp-dependent biotrophic mechanism of virulence: How has it evolved in tumorigenic bacteria? *Phytoparasitica* 33:317-324. 10.1007/BF02981296

Barash I, and Manulis-Sasson S. 2007. Virulence mechanisms and host specificity of gall-forming *Pantoea agglomerans*. *Trends in Microbiology* 15:538-545. 10.1016/j.tim.2007.10.009

Barash I, Panijel M, Gurel F, Chalupowicz L, and Manulis S. 2005. Transformation of *Pantoea agglomerans* into a tumorigenic pathogen. In: Sorvari S, and Toldi O, editors. Proceedings of the 1st International Conference on Plant-Microbe Interactions: Endophytes and Biocontriol Agents. Lapland, Finland: Saariselka. p 10-19.

Bardaji L, Perez-Martinez I, Rodriguez-Moreno L, Rodriguez-Palenzuela P, Sundin GW, Ramos C, and Murillo J. 2011. Sequence and role in virulence of the three plasmid complement of the model tumor-inducing bacterium *Pseudomonas savastanoi* pv. *savastanoi* NCPPB 3335. *PLoS One* 6:e25705. 10.1371/journal.pone.0025705

Binns AN, and Costantino P. 1998. The *Agrobacterium* oncogenes. In: Spaink HP, Kondorosi A, and Hooykaas PJJ, eds. *The Rhizobiaceae*. Netherlands: Springer, 251-166.

Bolger A, Lohse M, and Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114-2120. 10.1093/bioinformatics/btu170

Bomhoff G, Klapwijk PM, Kester HC, Schilperoort RA, Hernalsteens JP, and Schell J. 1976. Octopine and nopaline synthesis and breakdown genetically controlled by a plasmid of *Agrobacterium tumefaciens*. *Molecular & General Genetics* 145:177-181.

Bouzar H, and Jones JB. 2001. *Agrobacterium larrymoorei* sp. nov., a pathogen isolated from aerial tumours of *Ficus benjamina*. *International Journal of Systematic and Evolutionary Microbiology* 51:1023-1026. 10.1099/00207713-51-3-1023

623 Broothaerts W, Mitchell HJ, Weir B, Kaines S, Smith LMA, Yang W, Mayer JE, Roa-Rodríguez
624 C, and Jefferson RA. 2005. Gene transfer to plants by diverse species of bacteria. *Nature*
625 433:629-633. 10.1038/nature03309
626 Burr T, Katz B, Abawi G, and Crosier D. 1991. Comparison of tumorigenic strains of *Erwinia*
627 *herbicola* isolated from table beet with *E. h. gypsophilae*. *Plant Disease* 75:855-858.
628 Bushnell B. 2016. BBMap. Available at http://sourceforge.net/projects/bbmap/.
629 Castillo JA, and Greenberg JT. 2007. Evolutionary dynamics of *Ralstonia solanacearum*.
630 *Applied and Environmental Microbiology* 73:1225-1238. 10.1128/AEM.01253-06
631 Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in
632 phylogenetic analysis. *Molecular Biology and Evolution* 17:540-552.
633 Chan JZ-M, Halachev MR, Loman NJ, Constantinidou C, and Pallen MJ. 2012. Defining
634 bacterial species in the genomic era: insights from the genus *Acinetobacter*. *BMC*
635 *Microbiology* 12:302. 10.1186/1471-2180-12-302
636 Chang J-M, Di Tommaso P, and Notredame C. 2014. TCS: A new multiple sequence alignment
637 reliability measure to estimate alignment accuracy and improve phylogenetic tree
638 reconstruction. *Molecular Biology and Evolution* 31:1625-1637. 10.1093/molbev/msu117
639 Chilton MD, Drummond MH, Merio DJ, Sciaky D, Montoya AL, Gordon MP, and Nester EW.
640 1977. Stable incorporation of plasmid DNA into higher plant cells: the molecular basis of
641 crown gall tumorigenesis. *Cell* 11:263-271.
642 Clark E, Manulis S, Ophir Y, Barash I, and Y G. 1993. Cloning and characterization of *iaaM* and
643 *iaaH* from *Erwinia herbicola* pathovar *gypsophilae*. *Phytopathology* 83:234-240.
644 10.1094/Phyto-83-234
645 Cooksey DA. 1986. Galls of *Gypsophila paniculata* caused by *Erwinia herbicola*. *Plant Disease*
646 70:464. 10.1094/PD-70-464
647 Creason AL, Davis EW, Putnam ML, Vandeputte OM, and Chang JH. 2014a. Use of whole
648 genome sequences to develop a molecular phylogenetic framework for *Rhodococcus*
649 *fascians* and the *Rhodococcus* genus. *Frontiers in Plant Science* 5:406.
650 10.3389/fpls.2014.00406
651 Creason AL, Vandeputte OM, Savory EA, Davis EW, Putnam ML, Hu E, Swader-Hines D, Mol
652 A, Baucher M, Prinsen E, Zdanowska M, Givan SA, Jaziri ME, Loper JE, Mahmud T,
653 and Chang JH. 2014b. Analysis of genome sequences from plant pathogenic
654 *Rhodococcus* reveals genetic novelties in virulence loci. *PLoS One* 9:e101996.
655 10.1371/journal.pone.0101996
656 Crespi M, Messens E, Caplan AB, van Montagu M, and Desomer J. 1992. Fasciation induction
657 by the phytopathogen *Rhodococcus fascians* depends upon a linear plasmid encoding a
658 cytokinin synthase gene. *The EMBO Journal* 11:795-804.
659 Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, and Harris
660 SR. 2014. Rapid phylogenetic analysis of large samples of recombinant bacterial whole
661 genome sequences using Gubbins. *Nucleic Acids Research*:gku1196-.
662 10.1093/nar/gku1196
663 Darling A, Mau B, Blattner F, and Perna N. 2004. Mauve: multiple alignment of conserved
664 genomic sequence with rearrangements. *Genome Res* 14:1394-1403. 10.1101/gr.2289704
665 Delétoile A, Decré D, Courant S, Passet V, Audo J, Grimont P, Arlet G, and Brisse S. 2009.
666 Phylogeny and identification of *Pantoea* species and typing of *Pantoea agglomerans*
667 strains by multilocus gene sequencing. *Journal of Clinical Microbiology* 47:300-310.
668 10.1128/JCM.01916-08

669 DeYoung RM, Copeman RJ, and Hunt RS. 1998. Two strains in the genus *Erwinia* cause galls
670 on Douglas-fir in southwestern British Columbia. *Canadian Journal of Plant Pathology*
671 20:194-200. 10.1080/07060669809500427
672 Farrand SK, Van Berkum PB, and Oger P. 2003. *Agrobacterium* is a definable genus of the
673 family *Rhizobiaceae*. *International Journal of Systematic and Evolutionary Microbiology*
674 53:1681-1687. 10.1099/ijs.0.02445-0
675 Gardan L, Bollet C, Abu Ghorrah M, Grimont F, and Grimont PAD. 1992. DNA relatedness
676 among the pathovar strains of *Pseudomonas syringae* subsp. *savastanoi* Janse (1982) and
677 proposal of *Pseudomonas savastanoi* sp. nov. *International Journal of Systematic*
678 *Bacteriology* 42:606-612. 10.1099/00207713-42-4-606
679 Gloyer WO. 1934. *Crown gall and hairy root of apples in nursery and orchard*. Geneva, NY:
680 New York Agricultural Experiment Station Bulletin 638.
681 Goodfellow M. 1984. Reclassification of *Corynebacterium fascians* (Tilford) Dowson in the
682 genus *Rhodococcus*, as *Rhodococcus fascians* comb. nov. *Systematic and Applied*
683 *Microbiology* 5:225-229. 10.1016/S0723-2020(84)80023-5
684 Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, and Tiedje JM. 2007.
685 DNA-DNA hybridization values and their relationship to whole-genome sequence
686 similarities. *International Journal of Systematic and Evolutionary Microbiology* 57:81-
687 91. 10.1099/ijs.0.64483-0
688 Heibl C. 2013. PHYLOCH: interfaces and graphic tools for phylogenetic data in R. Available at
689 http://www.christophheibl.de/Rpackages.html.
690 Hildebrand EM. 1940. Cane gall of Brambles caused by *Phytomonas rubi* n.sp. *Journal of*
691 *Agricultural Research* 61:685--696 pp.
692 Hwang MSH, Morgan RL, Sarkar SF, Wang PW, and Guttman DS. 2005. Phylogenetic
693 characterization of virulence and resistance phenotypes of *Pseudomonas syringae*.
694 *Applied and Environmental Microbiology* 71:5182-5191. 10.1128/AEM.71.9.5182-
695 5191.2005
696 Iacobellis NS, Sisto A, Surico G, Evidente A, and DiMaio E. 1994. Pathogenicity of
697 *Pseudomonas syringae* subsp. *savastanoi* mutants defective in phytohormone production.
698 *Journal of Phytopathology* 140:238-248. 10.1111/j.1439-0434.1994.tb04813.x
699 Inouye M, Conway TC, Zobel J, and Holt KE. 2012. Short read sequence typing (SRST): multi-
700 locus sequence types from short reads. *BMC Genomics* 13:338. 10.1186/1471-2164-13-
701 338
702 Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ, Tomita T, Zobel J, and Holt KE.
703 2014. SRST2: Rapid genomic surveillance for public health and hospital microbiology
704 labs. *Genome Medicine* 6:90. 10.1186/s13073-014-0090-6
705 Jacques M-A, Durand K, Orgeur G, Balidas S, Fricot C, Bonneau S, Quillévéré A, Audusseau C,
706 Olivier V, Grimault V, and Mathis R. 2012. Phylogenetic analysis and polyphasic
707 characterization of *Clavibacter michiganensis* strains isolated from tomato seeds reveal
708 that nonpathogenic strains are distinct from *C. michiganensis* subsp. *michiganensis*.
709 *Applied and Environmental Microbiology* 78:8388-8402. 10.1128/AEM.02158-12
710 Janda JM, and Abbott SL. 2007. 16S rRNA gene sequencing for bacterial identification in the
711 diagnostic laboratory: pluses, perils, and pitfalls. *Journal of Clinical Microbiology*
712 45:2761-2764. 10.1128/JCM.01228-07
713 Joshi N, and Fass J. 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for
714 FastQ files. Available at https://github.com/najoshi/sickle.

715 Kado CI. 2014. Historical account on gaining insights on the mechanism of crown gall
716 tumorigenesis induced by *Agrobacterium tumefaciens*. *Frontiers in Microbiology* 5:340.
717 10.3389/fmicb.2014.00340
718 Kamvar ZN, Tabima JF, and Grünwald NJ. 2014. Poppr: an R package for genetic analysis of
719 populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* 2:e281.
720 10.7717/peerj.281
721 Katoh K, and Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
722 Improvements in performance and usability. *Molecular Biology and Evolution* 30:772-
723 780. 10.1093/molbev/mst010
724 Katoh K, and Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment
725 program. *Briefings in Bioinformatics* 9:286-298. 10.1093/bib/bbn013
726 Kim H-S, Ma B, Perna NT, and Charkowski AO. 2009. Phylogeny and virulence of naturally
727 occurring type III secretion system-deficient *Pectobacterium* strains. *Applied and*
728 *Environmental Microbiology* 75:4539-4549. 10.1128/AEM.01336-08
729 Kim M, Oh H-S, Park S-C, and Chun J. 2014. Towards a taxonomic coherence between average
730 nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of
731 prokaryotes. *International Journal of Systematic and Evolutionary Microbiology* 64:346-
732 351. 10.1099/ijs.0.059774-0
733 Langmead B, and Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature*
734 *Methods* 9:357-359. 10.1038/nmeth.1923
735 Lassalle F, Campillo T, Vial L, Baude J, Costechareyre D, Chapulliot D, Shams M, Abrouk D,
736 Lavire C, Oger-Desfeux C, Hommais F, Guéguen L, Daubin V, Muller D, and Nesme X.
737 2011. Genomic species are ecological species as revealed by comparative genomics in
738 *Agrobacterium tumefaciens*. *Genome Biology and Evolution* 3:762-781.
739 10.1093/gbe/evr070
740 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R,
741 and Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and
742 SAMtools. *Bioinformatics* 25:2078-2079.
743 Lichter A, Barash I, Valinsky L, and Manulis S. 1995. The genes involved in cytokinin
744 biosynthesis in *Erwinia herbicola* pv. *gypsophilae*: characterization and role in gall
745 formation. *Journal of Bacteriology* 177:4457-4465.
746 Maes T, Vereecke D, Ritsema T, Cornelis K, Thu HN, Van Montagu M, Holsters M, and
747 Goethals K. 2001. The *att* locus of *Rhodococcus fascians* strain D188 is essential for full
748 virulence on tobacco through the production of an autoregulatory compound. *Molecular*
749 *microbiology* 42:13-28.
750 Mansfield J, Genin S, Magori S, Citovsky V, Sriariyanum M, Ronald P, Dow M, Verdier V,
751 Beer SV, Machado MA, Toth I, Salmond G, and Foster GD. 2012. Top 10 plant
752 pathogenic bacteria in molecular plant pathology. *Molecular Plant Pathology* 13:614-
753 629. 10.1111/j.1364-3703.2012.00804.x
754 Manulis S, and Barash I. 2003. The molecular basis for transformation of an epiphyte into a gall-
755 forming pathogen as exemplified by *Erwinia herbicola* pv. *gypsophilae*. *Plant-Microbe*
756 *Interactions* 6:19-52.
757 Manulis S, Haviv-Chesner A, Brandl MT, Lindow SE, and Barash I. 1998. Differential
758 involvement of indole-3-acetic acid biosynthetic pathways in pathogenicity and epiphytic
759 fitness of *Erwinia herbicola* pv. *gypsophilae*. *Molecular Plant-Microbe Interactions*
760 11:634-642. 10.1094/MPMI.1998.11.7.634

761 Marrero G, Schneider KL, Jenkins DM, and Alvarez AM. 2013. Phylogeny and classification of
762 *Dickeya* based on multilocus sequence analysis. *International Journal of Systematic and*
763 *Evolutionary Microbiology* 63:3524-3539. 10.1099/ijs.0.046490-0
764 Matas IM, Lambertsen L, Rodríguez-Moreno L, and Ramos C. 2012. Identification of novel
765 virulence genes and metabolic pathways required for full fitness of *Pseudomonas*
766 *savastanoi* pv. *savastanoi* in olive (Olea europaea) knots. *New Phytologist* 196:1182-
767 1196. 10.1111/j.1469-8137.2012.04357.x
768 Montoya AL, Chilton MD, Gordon MP, Sciaky D, and Nester EW. 1977. Octopine and nopaline
769 metabolism in *Agrobacterium tumefaciens* and crown gall tumor cells: role of plasmid
770 genes. *Journal of Bacteriology* 129:101-107.
771 Mor H, Manulis S, Zuck M, Nizan R, Coplin DL, and Barash I. 2001. Genetic organization of
772 the *hrp* gene cluster and *dspAE*/*BF* operon in *Erwinia herbicola* pv. *gypsophilae*.
773 *Molecular Plant-Microbe Interactions* 14:431-436. 10.1094/MPMI.2001.14.3.431
774 Morris RO. 1986. Genes specifying auxin and cytokinin biosynthesis in phytopathogens. *Annual*
775 *Review of Plant Physiology* 37:509-538. 10.1146/annurev.pp.37.060186.002453
776 Ning Z, Cox AJ, and Mullikin JC. 2001. SSAHA: a fast search method for large DNA databases.
777 *Genome Research* 11:1725-1729. 10.1101/gr.194201
778 Nissan G, Manulis-Sasson S, Weinthal D, Mor H, Sessa G, and Barash I. 2006. The type III
779 effectors HsvG and HsvB of gall-forming *Pantoea agglomerans* determine host
780 specificity and function as transcriptional activators. *Molecular microbiology* 61:1118-
781 1131. 10.1111/j.1365-2958.2006.05301.x
782 Nizan R, Barash I, Valinsky L, Lichter A, and Manulis S. 1997. The presence of *hrp* genes on
783 the pathogenicity-associated plasmid of the tumorigenic bacterium *Erwinia herbicola* pv.
784 *gypsophilae*. *Molecular Plant-Microbe Interactions* 10:677-682.
785 10.1094/MPMI.1997.10.5.677
786 Nizan-Koren R, Manulis S, Mor H, Iraki NM, and Barash I. 2003. The regulatory cascade that
787 activates the Hrp regulon in *Erwinia herbicola* pv. *gypsophilae*. *Molecular Plant-*
788 *Microbe Interactions* 16:249-260. 10.1094/MPMI.2003.16.3.249
789 Opgenorth DC, Hendson M, and Clark E. 1994. First report of bacterial gall of *Wisteria sinensis*
790 caused by *Erwinia herbicola* pv. *milletiae* in California. *Plant Disease* 78:1217C.
791 10.1094/PD-78-1217C
792 Ophel K, and Kerr A. 1990. *Agrobacterium vitis* sp. nov. for Strains of *Agrobacterium* biovar 3
793 from Grapevines. *International Journal of Systematic Bacteriology* 40:236-241.
794 10.1099/00207713-40-3-236
795 Paradis E, Claude J, and Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R
796 language. *Bioinformatics* 20:289-290. 10.1093/bioinformatics/btg412
797 Parker JK, Havird JC, and De La Fuente L. 2012. Differentiation of *Xylella fastidiosa* strains via
798 multilocus sequence analysis of environmentally mediated genes (MLSA-E). *Applied and*
799 *Environmental Microbiology* 78:1385-1396. 10.1128/AEM.06679-11
800 Pearson T, Okinaka RT, Foster JT, and Keim P. 2009. Phylogenetic understanding of clonal
801 populations in an era of whole genome sequencing. *Infection, genetics and evolution :*
802 *journal of molecular epidemiology and evolutionary genetics in infectious diseases*
803 9:1010-1019. 10.1016/j.meegid.2009.05.014
804 Pérez-Yépez J, Armas-Capote N, Velázquez E, Pérez-Galdona R, Rivas R, and León-Barrios M.
805 2014. Evaluation of seven housekeeping genes for multilocus sequence analysis of the
806 genus *Mesorhizobium*: Resolving the taxonomic affiliation of the *Cicer canariense*

807 rhizobia. *Systematic and Applied Microbiology* 37:553-559.
808 10.1016/j.syapm.2014.10.003
809 Ponstingl H. 2013. SMALT. Available at https://www.sanger.ac.uk/resources/software/smalt/.
810 Putnam ML, and Miller ML. 2007. *Rhodococcus fascians* in herbaceous perennials. *Plant*
811 *Disease* 91:1064-1076. 10.1094/PDIS-91-9-1064
812 Sachs T. 1975. Plant tumors resulting from unregulated hormone synthesis. *Journal of*
813 *Theoretical Biology* 55:445-453.
814 Sarkar SF, and Guttman DS. 2004. Evolution of the core genome of *Pseudomonas syringae*, a
815 highly clonal, endemic plant pathogen. *Applied and Environmental Microbiology*
816 70:1999-2012.
817 Schroth MN. 1988. Reduction in yield and vigor of grapevine caused by crown gall disease.
818 *Plant Disease* 72:241. 10.1094/PD-72-0241
819 Sisto A, Cipriani MG, and Morea M. 2004. Knot Formation caused by *Pseudomonas syringae*
820 subsp. *savastanoi* on olive plants is *hrp*-dependent. *Phytopathology* 94:484-489.
821 10.1094/PHYTO.2004.94.5.484
822 Slater S, Goldman B, Goodner B, Setubal J, Farrand S, Nester E, Burr T, Banta L, Dickerman A,
823 Paulsen I, Otten L, Suen G, Welch R, Almeida N, Arnold F, Burton O, Du Z, Ewing A,
824 Godsy E, Heisel S, Houmiel K, Jhaveri J, Lu J, Miller N, Norton S, Chen Q,
825 Phoolcharoen W, Ohlin V, Ondrusek D, Pride N, Stricklin S, Sun J, Wheeler C, Wilson
826 L, Zhu H, and Wood D. 2009. Genome sequences of three agrobacterium biovars help
827 elucidate the evolution of multichromosome genomes in bacteria. *J Bacteriol* 191:2501-
828 2511. 10.1128/JB.01779-08
829 Stackebrandt E, and Goebel BM. 1994. Taxonomic Note: A place for DNA-DNA reassociation
830 and 16S rRNA sequence analysis in the present species definition in bacteriology.
831 *International Journal of Systematic Bacteriology* 44:846-849. 10.1099/00207713-44-4-
832 846
833 Stajich JE. 2002. The Bioperl Toolkit: Perl modules for the life sciences. *Genome Research*
834 12:1611-1618. 10.1101/gr.361602
835 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
836 large phylogenies. *Bioinformatics* 30:1312-1313. 10.1093/bioinformatics/btu033
837 Stes E, Francis I, Pertry I, Dolzblasz A, Depuydt S, and Vereecke D. 2013. The leafy gall
838 syndrome induced by *Rhodococcus fascians*. *FEMS Microbiol Lett* 342:187-195.
839 10.1111/1574-6968.12119
840 Tancos MA, Lange HW, and Smart CD. 2015. Characterizing the genetic diversity of the
841 *Clavibacter michiganensis* subsp. *michiganensis* population in New York.
842 *Phytopathology* 105:169-179. 10.1094/PHYTO-06-14-0178-R
843 Temmerman W, Vereecke D, Dreesen R, Van Montagu M, Holsters M, and Goethals K. 2000.
844 Leafy gall formation is controlled by *fasR*, an AraC-type regulatory gene in *Rhodococcus*
845 *fascians*. *Journal of Bacteriology* 182:5832-5840.
846 Thompson DV, Melchers LS, Idler KB, Schilperoort RA, and Hooykaas PJ. 1988. Analysis of
847 the complete nucleotide sequence of the *Agrobacterium tumefaciens virB* operon. *Nucleic*
848 *Acids Research* 16:4621-4636.
849 Van Larebeke N, Engler G, Holsters M, Van den Elsacker S, Zaenen I, Schilperoort RA, and
850 Schell J. 1974. Large plasmid in *Agrobacterium tumefaciens* essential for crown gall-
851 inducing ability. *Nature* 252:169-170.

852 Vasanthakumar A, and McManus PS. 2004. Indole-3-acetic acid-producing bacteria are
853 associated with cranberry stem gall. *Phytopathology* 94:1164-1171.
854 10.1094/PHYTO.2004.94.11.1164
855 Velázquez E, Palomo JL, Rivas R, Guerra H, Peix A, Trujillo ME, García-Benavides P, Mateos
856 PF, Wabiko H, and Martínez-Molina E. 2010. Analysis of core genes supports the
857 reclassification of strains *Agrobacterium radiobacter* K84 and *Agrobacterium*
858 *tumefaciens* AKE10 into the species *Rhizobium rhizogenes*. *Systematic and Applied*
859 *Microbiology* 33:247-251. 10.1016/j.syapm.2010.04.004
860 Vereecke D, Cornelis K, Temmerman W, Jaziri M, Van Montagu M, Holsters M, and Goethals
861 K. 2002. Chromosomal locus that affects pathogenicity of *Rhodococcus fascians*. *Journal*
862 *of Bacteriology* 184:1112-1120.
863 Ward JE, Akiyoshi DE, Regier D, Datta A, Gordon MP, and Nester EW. 1988. Characterization
864 of the *virB* operon from an *Agrobacterium tumefaciens* Ti plasmid. *Journal of Biological*
865 *Chemistry* 263:5804-5814.
866 Weinthal DM, Barash I, Panijel M, Valinsky L, Gaba V, and Manulis-Sasson S. 2007.
867 Distribution and replication of the pathogenicity plasmid pPATH in diverse populations
868 of the gall-forming bacterium *Pantoea agglomerans*. *Applied and Environmental*
869 *Microbiology* 73:7552-7561. 10.1128/AEM.01511-07
870 Wertz JE, Goldstone C, Gordon DM, and Riley MA. 2003. A molecular phylogeny of enteric
871 bacteria and implications for a bacterial species concept. *Journal of Evolutionary Biology*
872 16:1236-1248.
873 Young JM. 2003. Classification and nomenclature of *Agrobacterium* and *Rhizobium* - a reply to
874 Farrand et al. (2003). *International Journal of Systematic and Evolutionary Microbiology*
875 53:1689-1695. 10.1099/ijs.0.02762-0
876 Young JM, Kuykendall LD, Martínez-Romero E, Kerr A, and Sawada H. 2001. A revision of
877 *Rhizobium* Frank 1889, with an emended description of the genus, and the inclusion of all
878 species of *Agrobacterium* Conn 1942 and *Allorhizobium* undicola de Lajudie et al. 1998
879 as new combinations: *Rhizobium* radiobacter, R. rhizogenes, R. rubi, R. undicola and R.
880 vitis. *International Journal of Systematic and Evolutionary Microbiology* 51:89-103.
881 doi:10.1099/00207713-51-1-89
882 Young JM, Park D-C, Shearman HM, and Fargier E. 2008. A multilocus sequence analysis of the
883 genus Xanthomonas. *Systematic and Applied Microbiology* 31:366-377.
884 10.1016/j.syapm.2008.06.004
885 Zeigler DR. 2003. Gene sequences useful for predicting relatedness of whole genomes in
886 bacteria. *International Journal of Systematic and Evolutionary Microbiology* 53:1893-
887 1900.
888 Zerbino DR, and Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de
889 Bruijn graphs. *Genome Research* 18:821-829. 10.1101/gr.074492.107
890 Zhu J, Oger PM, Schrammeijer B, Hooykaas PJ, Farrand SK, and Winans SC. 2000. The bases
891 of crown gall tumorigenesis. *Journal of Bacteriology* 182:3885-3895.
892

Table 1(on next page)

Manually curated datasets developed for Gall-ID

1 **Table 1. Manually curated datasets developed for Gall-ID**

Database	Bacterial group	# of isolates used in Gall-ID	References
"Agro-type" tool (*Agrobacterium*)			
MLSA (*dnaK, glnA, gyrB, recA, rpoB, thrA, truA*)	Rhizobiaceae	199	Perez-Yepez et al.,
MLSA (*atpD, gapA, gyrB, recA, rplB*)	Rhizobiaceae	188	Alexandre et al., 2008
dnaJ	Rhizobiaceae	198	Alexandre et al., 2008
16S rDNA	Rhizobiaceae	245	
"Rhodo-type" tool (*Rhodococcus*)			
MLSA (*ftsY, infB, rpoB, rsmA, secY, tsaD, ychF*)	*Rhodococcus*	85	Adekambi et al., 2011
16S rDNA	*Rhodococcus*	66	
"Panto-type" tool (*Pantoea agglomerans*)			
MLSA (*fusA, gyrB, leuS, pyrG, rplB, rpoB*)	*Pantoea*, *Erwinia*	356	Delétoile et al., 2009
16S rDNA	*Pantoea*, *Erwinia*	352	
"Pseudo-type" tool (*Pseudomonas savastanoi*)			
MLSA (*gapA, gltA, gyrB, rpoD*)	*Pseudomonas syringae*	158	Hwang et al., 2005
MLSA (*acnB, fruK, gapA, gltA, gyrB, pgi, rpoD*)	*Pseudomonas syringae*	153	Sarkar et al., 2004
16S rDNA	*Pseudomonas syringae*	161	
"Phytopath-type" tool			
16S rDNA	*Rhodococcus*, *Agrobacterium*, *Pseudomonas syringae*, *Ralstonia*, *Xanthomonas*, *Pantoea*, *Erwinia*, *Xylella*, *Dickeya*, *Pectobacterium*,	345	

	Clavibacter, *Rathayibacter*		
MLSA (*atpD, dnaK, gyrB, ppK, recA, rpoB*)	*Clavibacter*	7	Jacques et al., 2012
MLSA (*dnaA, gyrB, kdpA, ligA, sdhA*)	*Clavibacter*	7	Tancos et al., 2015
MLSA (*dnaA, dnaJ, dnaX, gyrB, recN*)	*Dickeya*	40	Marrero et al., 2013
MLSA (*acnA, gapA, icdA, mdh, pgi*)	*Pectobacterium*	54	Kim et al., 2009
MLSA (*adk, egl, fliC, gapA, gdhA, gyrB, hrpB, ppsA*)	*Ralstonia*	28	Castillo et al., 2007
MLSA (*dnaK, fyuA, gyrB, rpoD*)	*Xanthomonas*	348	Young et al., 2008
MLSA (*acvB, copB, cvaC, fimA, gaa, pglA, pilA, rpfF, xadA*)	*Xylella*	17	Parker et al., 2012

2

Table 2(on next page)

Statistics for the WGS Pipeline

Table 2. Statistics for the WGS Pipeline

WGS Pipeline step	Statistic	Value
generate_pileup.sh (1 cpu)	Number of input paired read sets	19
	Average runtime per pileup (hh:mm:ss)	00:42:01
	Total runtime (hh:mm:ss)	13:18:14
generate_core_alignment.sh (1 cpu)	Total pileup alignment length	5,947,114 bp
	90%-shared core alignment length	855,355 bp
	Total runtime (hh:mm:ss)	00:15:32
remove_recombination.sh (10 cpus)	Number of core polymorphic sites	177,961 bp
	core SNP alignment length (w/o putative recombinant SNPs)	174,819 bp
	Computational time (hh:mm:ss)	04:25:32
	Actual runtime (hh:mm:ss)	00:29:28
	Figure output runtime (hh:mm:ss)	00:13:03
generate_phylogeny.sh (raxmlHPC-PTHREADS-AVX, 10 cpus)	Time to optimize RAxML parameters (hh:mm:ss)	00:02:32
	Time to compute 20 ML searches (hh:mm:ss)	00:34:53
	Number of bootstrap replicates (RAxML autoMRE)	50
	Time to compute 50 bootstrap searches (hh:mm:ss)	01:02:09
	Total runtime (hh:mm:ss)	01:39:34
All	Total runtime (hh:mm:ss)	15:55:51

Table 3(on next page)

Strain identity of 14 isolates associated with crown gall

Table 3. Strain identity of 14 isolates associated with crown gall

Isolate name	Host	Positive ID based on	# high quality read pairs	Clade (based on 16S rDNA)	# of virulence genes ID'ed
13-2099-1-2	Quaking Aspen	*virD2* PCR	1,244,074	*Agrobacterium*	63
13-626	Pear	*virD2* PCR	220,903	*Agrobacterium*	2 (*nocM*, *nocP*)
AC27/96	Pieris	Not pathogenic	826,690	*Rhizobium*	1 (*tssD*)
AC44/96	Pieris	No reaction to hybridization probes	1,404,002	*Rhizobium*	0
B131/95	Peach/Almond Rootstock	Pathogenicity assay	539,283	*Agrobacterium*	46
B133/95	Peach/Almond Rootstock	Pathogenicity assay	1,199,902	*Agrobacterium*	46
B140/95	Peach/Almond Rootstock	Response to 20 different biochemical and physiological tests	448,314	*A. tumefaciens*	51
N2/73	Cranberry gall	Response to 20 different biochemical and physiological tests	1,345,404	*A. tumefaciens*	64
W2/73	Euonymus	Response to 20 different biochemical and physiological tests	1,244,159	*A. rubi*	51
15-1187-1-2a	Yarrow	*virD2* PCR	508,223	*A. tumefaciens*	39
15-1187-1-2b	Yarrow	*virD2* PCR	299,970	*A. tumefaciens*	38
14-2641	Rose	No data	698,756	*Serratia*	0
15-172	Leucanthemum	Colony morphology on selective media	384,308	*A. tumefaciens*	56
15-174	Leucanthemum	Colony morphology on selective media	753,570	*A. tumefaciens*	58

Figure 1(on next page)

Overview of Gall-ID diagnostic tools.

DNA sequence information can be used to reveal the identity of the causative agent (unknown isolate) of disease. Tools associated with "Gall Isolate Typing" and “Phytopath-type” use 16S rDNA or pathogen-specific MLSA gene sequences to infer the identity of the isolate by comparing the sequences to manually curated sequence databases. Tools associated with “Whole Genome Analysis” and “Vir-Search” use Illumina short sequencing reads to characterize pathogenic isolates. The former tab provides downloadable tools to infer genetic relatedness based on SNPs (WGS Pipeline) or average nucleotide identity (Auto ANI). The “Vir-Search” tab provides an on-line tool to quickly map short reads against a database of sequences of virulence genes.

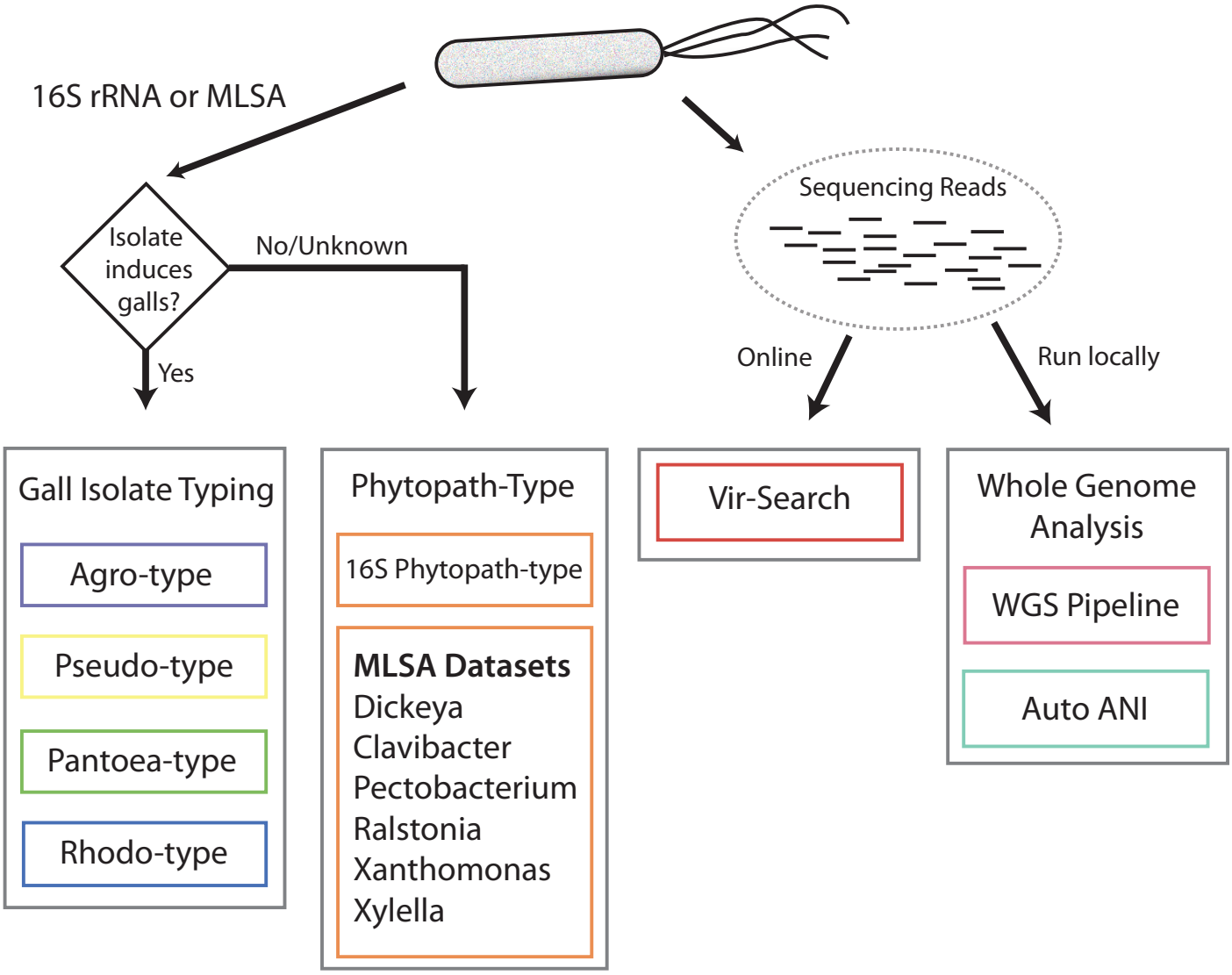

Unknown isolate
16S rRNA or MLSA
Isolate induces galls?
No/Unknown
Yes
Sequencing Reads
Online
Run locally
Gall Isolate Typing
Agro-type
Pseudo-type
Pantoea-type
Rhodo-type
Phytopath-Type
16S Phytopath-type
MLSA Datasets
Dickeya
Clavibacter
Pectobacterium
Ralstonia
Xanthomonas
Xylella
Vir-Search
Whole Genome Analysis
WGS Pipeline
Auto ANI

Figure 2(on next page)

Flowchart for the WGS Pipeline.

Scripts and the programs that each script runs are boxed and presented along the left. The logic flow of the WGS Pipeline tool is presented along the right. Rectangles with rounded corners = inputs and outputs; boxes outlined in red = processes. The inputs, outputs, and processes are matched to the corresponding script and program.

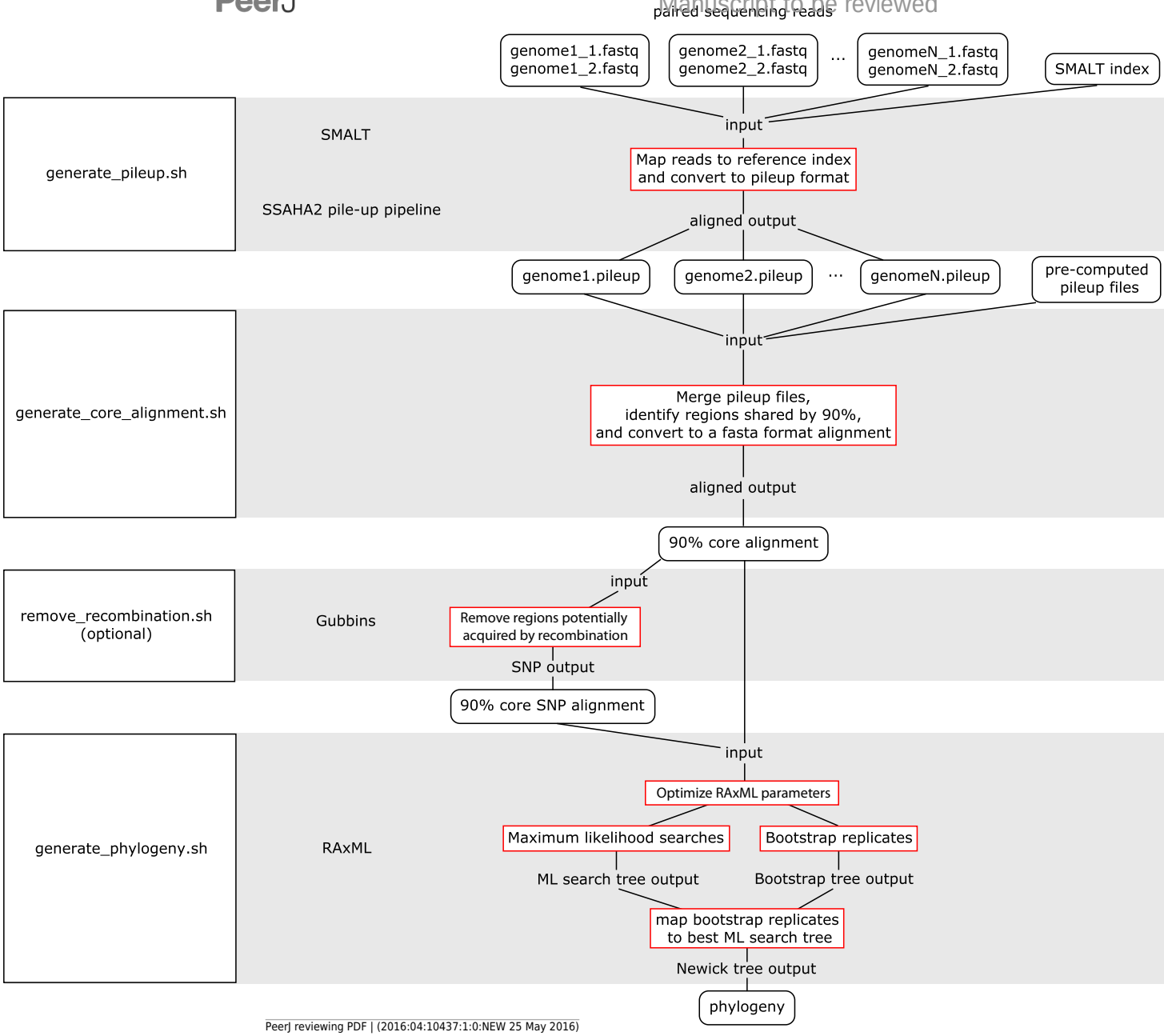

paired sequencing reads
genome1_1.fastq
genome1_2.fastq
genome2_1.fastq
genome2_2.fastq
...
genomeN_1.fastq
genomeN_2.fastq
SMALT index
generate_pileup.sh
SMALT
input
Map reads to reference index
and convert to pileup format
SSAHA2 pile-up pipeline
aligned output
genome1.pileup
genome2.pileup
...
genomeN.pileup
pre-computed
pileup files
input
generate_core_alignment.sh
Merge pileup files,
identify regions shared by 90%,
and convert to a fasta format alignment
aligned output
90% core alignment
input
remove_recombination.sh
(optional)
Gubbins
Remove regions potentially
acquired by recombination
SNP output
90% core SNP alignment
input
generate_phylogeny.sh
RAxML
Optimize RAxML parameters
Maximum likelihood searches
Bootstrap replicates
ML search tree output
Bootstrap tree output
map bootstrap replicates
to best ML search tree
Newick tree output
phylogeny

Figure 3(on next page)

Validation of the Agro-type and Vir-Search tools.

A) An unrooted Neighbor Joining phylogenetic tree based on 16s rDNA sequences from *Agrobacterium* spp. The 16S rDNA sequence was identified and extracted from the genome assembly of *Agrobacterium* isolate 13-2099-1-2 and analyzed using the tool available in the Agro-type tab. The isolate is labeled in red, as “query_isolate”; inset shows the clade that circumscribes the isolate. **B)** Screenshot of output results from Vir-Search. Paired 2x300 bp MiSeq short reads from *Agrobacterium* isolate 13-2099-1-2 were analyzed using the Vir-Search tool in Gall-ID. Reference virulence gene sequences that were aligned are indicated with a green plus ("+") icon and the lengths and depths of the read coverage are reported (must exceed user-specified cutoffs, which were designated as 90% minimum coverage and 20% maximum sequence divergence). Virulence genes that failed to exceed user-specific cutoffs for read alignment parameters are indicated with a red "X". Virulence genes are grouped into categories based on their function in virulence.

A.

B.

Virulence Gene Search results

Submitted job name: test16s
Organism: Agrobacterium
Forward/Single reads file: Atu_JL5258_trimmed_1.fastq
Reverse reads file:
Minimum % coverage of database gene: 90%
Maximum % divergence from database gene: 10%
Output file prefix: job_1441843284_6208
Email: aweisberg@gmail.com

Result table

SRST2 full gene output file: http://gall-id.cgrb.oregonstate.edu/files/tmp/job_1441843284_6208/job_1441843284_6208.out__fullgenes__agrovirdb__results.txt

Agrobacterium tumefaciens C58 genes were used as reference for all virulence genes except for GALLS which came from *Agrobacterium rhizogenes* strain K599

Oncogenes

Found	Gene	Coverage (%)	Depth	Closest Allele
✓	tms2	100.0	31.699	Rhizobium_rubi_NBRC_13261
✓	tms1	100.0	27.473	Rhizobium_rubi_NBRC_13261
✓	ipt	100.0	36.349	Rhizobium_rubi_NBRC_13261
✗	galls	-	-	-

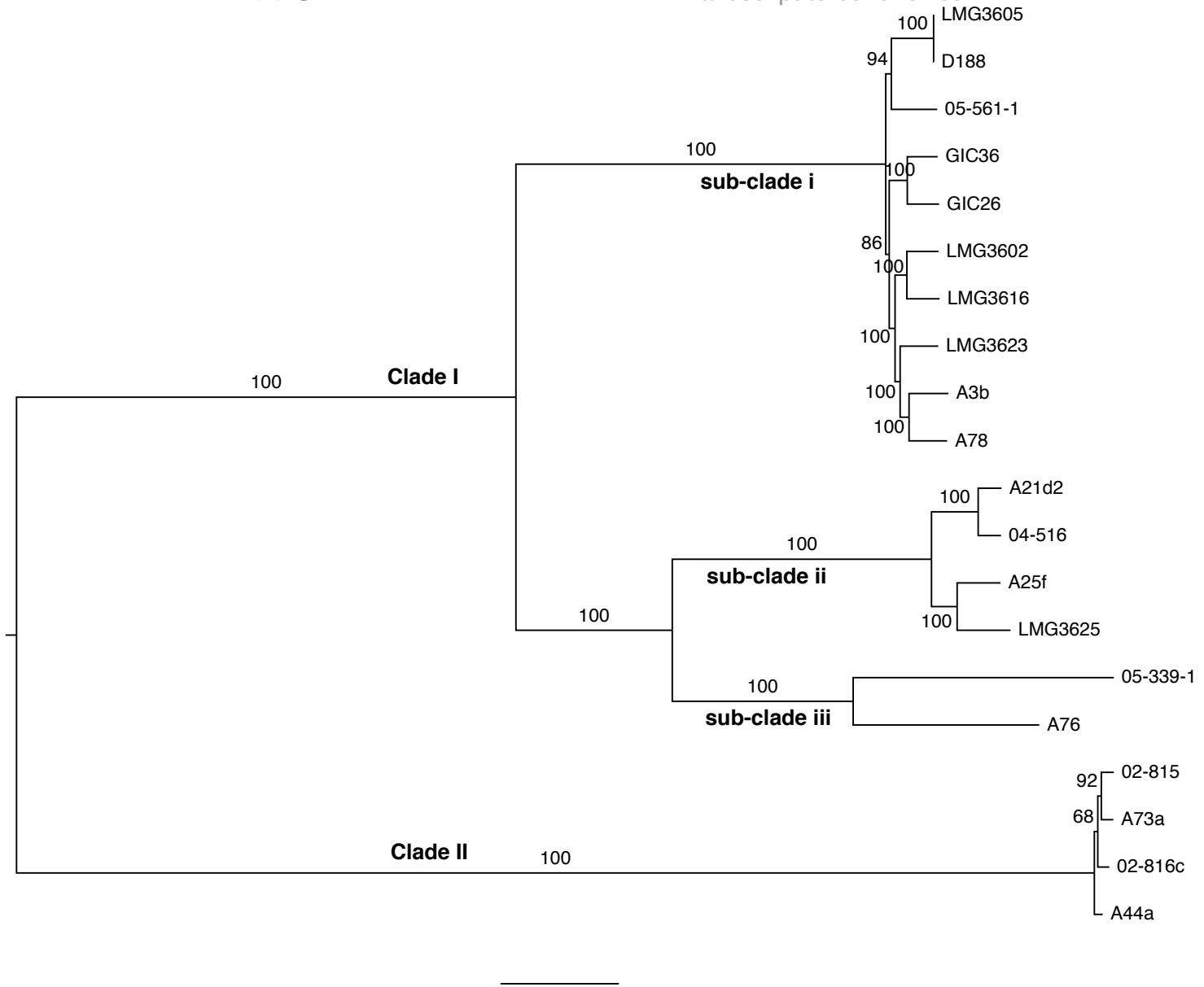
Opine Synthesis/Transport/Catabolism

Found	Gene	Coverage (%)	Depth	Closest Allele
✓	accA	100.0	24.815	Rhizobium_rubi_NBRC_13261
✓	accB	100.0	21.128	Rhizobium_rubi_NBRC_13261
✓	virB3	100.0	28.21	Rhizobium_rubi_NBRC_13261
✓	virB4	100.0	30.705	Rhizobium_rubi_NBRC_13261
✓	virB5	100.0	23.976	Agrobacterium_arsenijevicii_strain_KFB_330
✓	virB6	100.0	31.687	Agrobacterium_arsenijevicii_strain_KFB_330
✓	virB7	100.0	28.272	Agrobacterium_arsenijevicii_strain_KFB_330
✓	virB8	100.0	35.109	Rhizobium_rubi_NBRC_13261
✓	virB9	100.0	27.213	Rhizobium_rubi_NBRC_13261
✓	virB10	100.0	27.482	Rhizobium_rubi_NBRC_13261
✓	nocT	100.0	26.191	Rhizobium_rubi_NBRC_13261
✓	nos	100.0	27.164	Rhizobium_rubi_NBRC_13261
✓	noxA	100.0	24.153	Rhizobium_rubi_NBRC_13261
✓	noxB	100.0	25.276	Rhizobium_rubi_NBRC_13261
✓	occJ	100.0	27.164	Rhizobium_rubi_NBRC_13261
✓	occM	100.0	63.642	Rhizobium_sp._YR060
✓	occP	100.0	66.848	Rhizobium_tropici_strain_YR530
✓	occQ	100.0	59.199	Agrobacterium_radiobacter_K84

Figure 4(on next page)

Maximum likelihood tree based on vertically inherited polymorphic sites core to 20 *Rhodococcus* isolates.

WGS Pipeline was used to automate the processing of paired end short reads from 20 previously sequenced *Rhodococcus* isolates, and generate a maximum likelihood unrooted tree. Sequencing reads were aligned, using *R. fascians* strain A44a as a reference. SNPs potentially acquired via recombination were removed. The tree is midpoint-rooted. Scale bar = 0.05 average substitutions per site; non-parametric bootstrap support as percentages are indicated for each node. Major clades and sub-clades are labeled in a manner consistent with previous labels.


LMG3605
D188
05-561-1
GIC36
GIC26
LMG3602
LMG3616
LMG3623
A3b
A78
A21d2
04-516
A25f
LMG3625
05-339-1
A76
02-815
A73a
02-816c
A44a
100
94
100
100
86
100
100
100
100
100
100
100
100
100
100
100
100
92
68
100
sub-clade i
sub-clade ii
sub-clade iii
Clade I
Clade II
0.05