

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 increasing 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/>.

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

2
3 Edward W. Davis II^{1,2*}, Alexandra J. Weisberg^{1*}, Javier F. Tabima¹, Niklaus J. Grünwald^{1,2,3,4},
4 and Jeff H. Chang^{1,2,4}

5
6 ¹Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR, 97331,
7 USA

8 ²Molecular and Cellular Biology Program, Oregon State University, Corvallis, OR, 97331, USA

9 ³Horticultural Crops Research Laboratory, USDA-ARS, Corvallis, OR, 97331, USA

10 ⁴Center for Genome Research and Biocomputing, Oregon State University, Corvallis, OR,
11 97331, USA

12
13 *Equal contribution

14
15 Corresponding author

16 Jeff H. Chang

17 3096 Cordley Hall, Corvallis, OR, USA

18 changj@science.oregonstate.edu

19 **ABSTRACT**

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

37

38

39 INTRODUCTION

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

49 The 16S rDNA sequence is commonly used for identification (Stackebrandt & Goebel
50 1994). Because of highly conserved regions in the gene sequence, a single pair of degenerate
51 oligonucleotide primers can be used to amplify the gene from a diversity of bacteria, and allow
52 for a kingdom-wide comparison. In general, the sequences of the amplified fragments have
53 enough informative polymorphic sites to delineate genera, but do not typically allow for more
54 refined taxonomic inferences at the sub-genus level (Janda & Abbott 2007). Multilocus sequence
55 analysis (MLSA) leverages the phylogenetic signal from four to ten genes to provide increased
56 resolution, and can distinguish between species and sometimes sub-species (Wertz et al. 2003;
57 Zeigler 2003). MLSA however, is more restricted than the use of 16S rDNA sequences and may
58 not allow for comparisons between members of different genera. MLSA also requires more time
59 and effort to identify informative and taxon-specific genes as well as develop corresponding
60 oligonucleotide primer sets.

61 Whole genome sequences can also be used. This is practical because of advances in next
62 generation sequencing technologies. The key advantages of this approach is that availability of
63 whole genome sequences obviates the dependency on *a priori* knowledge to provide clues on
64 taxonomic association of a pathogen and the need to select and amplify marker genes. Also,
65 whole genome sequences provide the greatest resolution in terms of phylogenetic signal, and
66 sequences that violate assumptions of phylogenetic analyses can be removed from studies to
67 allow for robust analyses. Briefly, sequencing reads of genome(s) are compared to a high quality
68 draft or finished reference genome sequence to identify variable positions between the genome
69 sequences (Pearson et al. 2009). The positions core to the compared genome sequences are
70 aligned and used to generate a phylogenetic tree. Alternatively, whole genome sequences can be
71 used to determine average nucleotide identities (ANI) between any sufficiently similar pair, e.g.,
72 within the same taxonomic family, of genome sequences to determine genetic relatedness (Goris
73 et al. 2007; Kim et al. 2014). ANI can be used to make taxonomic inferences, as a 95% threshold
74 for ANI has been calibrated to those used to operationally define bacterial species based on 16S
75 rDNA (> 94% similarity) and DNA-DNA hybridization (DDH; 70%) (Goris et al. 2007). Finally,
76 the whole genome sequences can be analyzed to inform on more than just the identity of the
77 causative agent and provide insights into mechanisms and evolution of virulence. However, a
78 non-trivial trade-off is that processing, storing, and analyzing whole genome sequence data sets
79 require familiarity with methods in computational biology.

80 Members from several taxa of Gram-negative bacteria are capable of causing abnormal
81 growth of plants. Members of *Agrobacterium* are the most notorious causative agents of
82 deformation of plant growth. These bacteria have been classified according to various schemes
83 that differ in the phenotypic and genetic characteristics that were used. Its taxonomic

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

99 A Ti plasmid imparts upon members of *Agrobacterium* the ability to genetically modify
100 its host and cause dysregulation of host phytohormone levels and induce gall formation (Sachs
101 1975). The Ti plasmid contains a region of DNA (T-DNA) that is transferred and integrated into
102 the genome of the host cell (Van Larebeke et al. 1974; Chilton et al. 1977; Thompson et al. 1988;
103 Ward et al. 1988; Broothaerts et al. 2005). Conjugation is mediated by a type IV secretion
104 system, encoded by genes located outside the borders of the T-DNA on the Ti plasmid
105 (Thompson et al. 1988). Within the T-DNA are key genes that encode for auxin, cytokinin, and
106 opine biosynthesis (Morris 1986; Binns & Costantino 1998). Expression of the former two genes

107 in the plant leads to an increase in plant hormone levels to cause growth deformation whereas the
108 latter set of genes encode enzymes for the synthesis of modified sugars that only organisms with
109 the corresponding opine catabolism genes can use as an energy source (Bomhoff et al. 1976;
110 Montoya et al. 1977). The latter genes are located outside the T-DNA borders on the Ti plasmid
111 (Zhu et al. 2000).

112 *Pseudomonas savastanoi* (formerly *Pseudomonas syringae* pv. *savastanoi*, (Gardan et al.
113 1992)) is the causal agent of olive knot disease, typically forming as aerial tumors on stems and
114 branches. Phytopathogenicity of *P. savastanoi* is dependent on the *hrp/hrc* genes located on the
115 chromosome (Sisto et al. 2004). These genes encode for a type III secretion system, a molecular
116 syringe that injects type III effector proteins into host cells that collectively function to dampen
117 host immunity (Chang et al. 2014). Phytopathogenicity of *P. savastanoi* is also associated with
118 the production of phytohormones. Indole-3-acetic acid may have an indirect role as a bacterial
119 signaling molecule (Aragon et al. 2014). A cytokinin biosynthesis gene has also been identified
120 on plasmids in *P. savastanoi* and strains cured of the plasmid caused smaller galls but were not
121 affected in growth within the galls (Iacobellis et al. 1994; Bardaji et al. 2011).

122 *Pantoea agglomerans* (formerly *Erwinia herbicola*) is a member of the
123 Enterobacteriaceae family. *P. agglomerans* can induce the formation of galls on diverse species
124 of plants (Cooksey 1986; Burr et al. 1991; Opgenorth et al. 1994; DeYoung et al. 1998;
125 Vasanthakumar & McManus 2004). Phytopathogenicity is dependent on the pPATH plasmid
126 (Manulis & Barash 2003; Weinthal et al. 2007). This plasmid contains a pathogenicity island
127 consisting of an *hrp/hrc* cluster and operons encoding for the biosynthesis of cytokinins, indole-
128 3-acetic acid, and type III effectors (Clark et al. 1993; Lichter et al. 1995; Nizan et al. 1997; Mor
129 et al. 2001; Nizan-Koren et al. 2003; Barash & Manulis 2005; Barash et al. 2005; Barash &

130 Manulis-Sasson 2007). As is the case with *P. savastanoi*, mutants of the *hrp/hrc* genes abolish
131 pathogenicity whereas mutations in the phytohormone biosynthesis genes led to galls of reduced
132 size (Manulis et al. 1998; Mor et al. 2001; Nizan-Koren et al. 2003; Barash & Manulis-Sasson
133 2007).

134 Gram-positive bacteria within the *Rhodococcus* genus can cause leafy gall disease to over
135 100 species of plants (Putnam & Miller 2007). The phytopathogenic members of this genus
136 belong to at least two genetically distinct groups of bacteria, with *R. fascians* (formerly
137 *Corynebacterium fascians*) being the original recognized species (Goodfellow 1984; Creason et
138 al. 2014a). It is suggested that *R. fascians* upsets levels of phytohormones of the plant to induce
139 gall formation. However, unlike Ti plasmid-carrying *Agrobacterium*, it is hypothesized that *R.*
140 *fascians* directly synthesizes and secretes the cytokinin phytohormone (Stes et al. 2013; Creason
141 et al. 2014b). Phytopathogenicity is most often associated with a linear plasmid, which carries a
142 cluster of virulence loci, *att*, *fasR*, and *fas* (Creason et al. 2014b). The functions for the translated
143 products of *att* are unknown but the sequences have homology to proteins involved in amino acid
144 and antibiotic biosynthesis (Maes et al. 2001). The *fasR* gene is necessary for full virulence; the
145 gene encodes a putative transcriptional regulator (Temmerman et al. 2000). Some of the genes
146 within the *fas* operon are necessary for virulence, as many of the *fas* genes encode proteins with
147 demonstrable functions in cytokinin metabolism (Crespi et al. 1992). In rare cases, the virulence
148 loci, or variants therein, are located on the chromosome (Creason et al. 2014b).

149 We developed Gall-ID to aid in determining the genetic identity of gall-causing members
150 of *Agrobacterium*, *Pseudomonas*, *Pantoea*, and *Rhodococcus*. Users can provide sequences from
151 16S rDNA or gene sets used in MLSA, and Gall-ID will automatically query curated databases
152 and generate phylogenetic trees to group the query isolate of interest and provide estimates of

153 relatedness to previously characterized species and/or genotypes. Users can also submit short
154 reads from whole genome sequencing projects to query curated databases to search for evidence
155 for known virulence genes of these gall-causing bacteria. Finally, users can download tools that
156 automate the analysis of whole genome sequencing data to infer genetic relatedness based on
157 MLSA, average nucleotide identity (ANI), or single nucleotide polymorphisms (SNPs).

158

159 **RESULTS AND DISCUSSION**

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

164 The “Gall Isolate Typing” tab provides online tools to use molecular data, either 16S or
165 sequences of marker genes used for MLSA, to group isolates of interest into corresponding
166 taxonomic units that include gall-causing pathogens. Users must first select the appropriate
167 taxonomic group, *Agrobacterium*, *Pseudomonas*, *Pantoea*, or *Rhodococcus* for comparison. For
168 some of these taxonomic groups, multiple gene sets used in MLSA are available, and the user
169 must therefore select the appropriate set for analysis. FASTA formatted gene sequences are
170 input, and after selecting the appropriate options for alignment and tree parameters, a
171 phylogenetic tree that includes the isolate of interest is generated and displayed. The tree
172 parameters include choice of distance matrix, tree generating algorithm (Neighbor-Joining or
173 UPGMA), and number of bootstrap replicates. A sub-clade of the tree containing only the isolate
174 of interest and its nearest sister taxa is displayed to the right of the full tree. The tree can be

175 saved as a Newick file or as a PDF. An example sequence from *Agrobacterium* can be loaded by
176 clicking the "Demo" button located in the Agro-type tab.

177 The "Phytopath-Type" tab provides online tools for the analysis of other non-gall-causing
178 pathogens important in agriculture (Mansfield et al. 2012). This tool is similar in function to the
179 "Gall Isolate Typing" tools, except it is not limited to a single taxon of pathogen. A database of
180 16S rDNA sequences from genera of important bacterial phytopathogens (*Pseudomonas*
181 *syringae* group, *Ralstonia*, *Agrobacterium*, *Rhodococcus*, *Xanthomonas*, *Pantoea*, *Xylella*,
182 *Dickeya*, *Pectobacterium*, and *Clavibacter*) is available for associating a bacterial pathogen to its
183 genus. Additionally, for *Clavibacter*, *Dickeya*, *Pectobacterium*, *Ralstonia*, *Xanthomonas*, and
184 *Xylella*, the user can use MLSA to genotype isolates of interest. As is the case with Gall Isolate
185 Typing, a phylogenetic tree will be generated and displayed, associating the isolate of interest to
186 the most closely related genotype in the curated databases.

187 The "Vir-Search" tab provides an online tool for using user-input read sequences of a
188 genome to search for the presence of homologs of known virulence genes. Users select a
189 taxonomic group (*Agrobacterium*, *P. savastanoi*, *P. agglomerans*, or *Rhodococcus*) to designate
190 the set of virulence genes to search against. Users also determine a minimum percent gene
191 coverage and maximum allowed percent identity divergence, and upload single or paired read
192 files in FASTQ format. The user-supplied read sequences are then aligned to the chosen
193 virulence gene dataset on the Gall-ID server. Once the search is complete, a link to the final
194 results is sent to a user-provided email address. Results display the percent coverage of the
195 virulence genes and the percent similarity of the covered sequences. If the query identifies
196 multiple alleles of virulence genes from different sequenced strains, the Vir-Search tool will
197 report the strain name associated with the best-mapped allele. User-submitted data and results are

198 confidential and submitted sequencing reads are deleted from the Gall-ID server upon
199 completion of the analysis.

200 A key component of the tools associated with the aforementioned tabs is the manually
201 curated databases of gene sequences. The literature was reviewed to identify validated taxon-
202 specific sets of genes for MLSA of taxa with gall-causing bacteria as well as other pathogens that
203 affect agriculture (Table 1) (Sarkar & Guttman 2004; Hwang et al. 2005; Castillo & Greenberg
204 2007; Alexandre et al. 2008; Young et al. 2008; Delétoile et al. 2009; Kim et al. 2009; Adékambi
205 et al. 2011; Jacques et al. 2012; Parker et al. 2012; Marrero et al. 2013; Pérez-Yépez et al. 2014;
206 Tancos et al. 2015). The sequences for the corresponding genes were subsequently extracted
207 from the whole genome sequences of reference strains. Auto MLSA was employed to use the
208 gene sequences as queries in BLAST searches. Auto MLSA is based on a previously developed
209 set of Perl scripts to automate retrieving, filtering, aligning, concatenating, determining of best
210 substitution models, appending of key identifiers to sequences, and generating files for tree
211 construction (Creason et al. 2014a). Gene sets in which there were less than 50% query sequence
212 coverage for all of the genes were excluded to ensure that the databases contained only
213 taxonomically informative sequences. Each gene set database was manually checked for
214 duplicate strains, large gaps in gene sequences, poor sequence alignment, and mis-annotated
215 taxonomic information. Each of the MLSA databases used in the Gall-ID tools is also available
216 for download on the "Database Downloads" tab of the Gall ID website. The 16S rDNA databases
217 were populated in a similar manner, with one small exception. For the Phytopath-Type tool, the
218 sequence of the 16S rRNA-encoding gene from C58 of *Agrobacterium* was used as a query to
219 retrieve corresponding sequences from 345 isolates distributed across the different genera of
220 plant pathogenic bacteria.

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

232 The analyses of whole genome sequence datasets can be computationally intensive,
233 which is prohibitive for online tools. Therefore, the "Whole Genome Analysis" tab provides
234 downloadable software pipeline tools for users to employ their institutional infrastructure or a
235 cloud computing service to analyze whole genome sequencing reads (Illumina HiSeq or MiSeq).
236 There are two options in this tab, the first, "WGS Pipeline: Core Genome Analysis," provides a
237 download link and instructions for using the WGS Pipeline tool to generate a phylogeny based
238 on the core genome sequence or core set of single nucleotide polymorphisms (SNPs). The
239 second option, "Auto ANI: Average Nucleotide Identity Analysis," provides a download link for
240 the Auto ANI tool and detailed instructions for its use in calculating all possible pairwise ANI
241 within a set of genome sequences.

242 The WGS Pipeline is a set of scripts that automates the use of sequences from Illumina-
243 based paired reads derived from whole genome sequencing projects to determine core genome

244 sequences or core SNPs and generate phylogenetic trees (Figure 2). This pipeline uses SMALT
245 and SSAHA2 pile-up pipeline to align sequencing reads to an indexed reference genome
246 sequence and generate a pileup file, respectively (Ning et al. 2001; Ponstingl 2013). The WGS
247 Pipeline then combines the pileup files along with other pre-computed pileup files to derive a
248 core genome alignment defined based on regions that are shared between at least 90% of the
249 compared genome sequences. Users have the option of using Gubbins to remove regions that are
250 flagged as potentially derived from recombination (Croucher et al. 2014). Invoking Gubbins will
251 also remove all non-polymorphic sites from the alignments, thus yielding a SNP alignment that
252 is based only on polymorphic sites that are identified as vertically inherited and shared between
253 at least 90% of the compared genome sequences. Finally, the user can use either the core genome
254 sequence or core SNP alignment and RAxML to generate a maximum likelihood (ML)
255 phylogeny (Stamatakis 2014).

256 Users must place their input files in the correspondingly named folders in order to run the
257 WGS pipeline. Paired read sequences for each genome are read from the "reads" folder, while a
258 SMALT index named as "reference" and placed in the "index" folder will be used as a reference
259 to align to. Identifiers are taken from the prefix of the read pair file names and used to name the
260 output pileup files and taxa in the phylogeny. The read pair file names must have the suffixes
261 ".1.fastq" and ".2.fastq" for files with forward and reverse read sequences, respectively. The read
262 sequences must be in FASTQ format and because of requirements of the SMALT program, each
263 paired read name must end in ".p1k" and ".q1k" for forward and reverse reads, respectively. If
264 the input read sequences are not in the proper format, the user may run the included optional
265 script "prepare_for_pileup.sh" to format read names. If the user has pre-computed SMALT
266 pileup files prepared using the same SMALT index, the files may be placed in the "pileup" folder

267 and will also be included in the analysis. The user may be prompted to input the length of the
268 inserts for each sequencing library. Users also have the option of changing the number of ML
269 searches or non-parametric bootstrap replicates when building a phylogeny (default values are 20
270 ML searches, autoMRE cutoff criterion for bootstrap replicates).

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

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

288 We validated the efficacy of the online tools available from the Gall Isolate Typing, and
289 Vir-Search tabs. DNA from 14 isolates were prepared, barcoded, and sequenced on an Illumina

290 MiSeq (Table 3). Of these isolates, the identities of 11 were previously verified as
291 *Agrobacterium*. The remaining three were associated with plant tissues showing symptoms of
292 crown gall disease but were not tested or had results inconsistent with being a pathogenic
293 member of the *Agrobacterium* genus. The reads were trimmed for quality and first *de novo*
294 assembled within each library using the Velvet assembler (Zerbino & Birney 2008). The 16S
295 gene sequences were identified and extracted from the assemblies and used as input for the
296 Agro-type tool. The 16S gene sequences from each of the 11 isolates originally typed as
297 *Agrobacterium* clustered accordingly; isolate 13-2099-1-2 is shown as an example (Figure 3A).
298 The 16S sequence from isolates AC27/96, AC44/96, and 14-2641 were more distant from the
299 16S sequences of *Agrobacterium* (Table 3). The isolates AC27/96 and AC44/96 grouped more
300 closely with various *Rhizobium* species, while subsequent analysis using the Phytopath-Type tool
301 suggested isolate 14-2641 was more closely related to members of *Erwinia*, *Dickeya*, and
302 *Pectobacterium* (Table 3, Supplementary Figure 1). A search against the NCBI nr database
303 revealed similarities to members of *Serratia*.

304 The trimmed read sequences were used as input for the Vir-Search tool as an additional
305 step to confirm the identity of these isolates. Paired read sequences for each of the 14 isolates
306 were individually uploaded to the Gall-ID server. The *Agrobacterium* virulence gene database
307 was selected, with the minimum gene length coverage set to 80% and maximum allowed
308 sequence divergence set to 20%. The time for each Vir-Search analysis ranged from 2-5 minutes.
309 Results suggested that the genome sequences for nearly all of the *Agrobacterium* isolates had
310 homologs of virulence genes demonstrably necessary for pathogenicity by *Agrobacterium*, while
311 the genome sequences for the isolates AC27/96, AC44/96, and 14-2641 did not (Figure 3B, data
312 for isolate 13-2099-1-2 shown). Contrary to the results from molecular diagnostics tests, the

313 reads from isolate 13-626 failed to align to any virulence genes except for two (*nocM*, *nocP*)
314 involved in nopaline transport. This isolate had the fewest number of useable sequencing reads
315 and the highest number of contigs compared to the others, and results could have been a
316 consequence of a poor assembly of the Ti plasmid.

317 Indeed, the qualities of the 14 assemblies were highly variable, likely reflecting the multi-
318 partite structure of the agrobacterial genomes, presence of a linear replicon, and/or variation in
319 depth of sequencing. We therefore used SPAdes v. 3.6.2 to *de novo* assemble each of the genome
320 sequences, with the exception of isolate 14-2641 (Bankevich et al. 2012). The total sizes of the
321 assemblies were similar to those generated using Velvet and the qualities of the assemblies were
322 high. But assemblies generated using SPAdes had proliferations in errors with palindromic
323 sequence that appeared to be unique to isolates expected to have linear replicons. We informed
324 the developers of the SPAdes software who immediately resolved the issue in SPAdes 3.7.0.
325 Inspection of the summary statistics of the assemblies derived using this latest version of SPAdes
326 suggested that relative to Velvet-based assemblies, there were improvements to all assemblies,
327 with the most dramatic to those with the lowest read coverage (Supplementary Table 1,
328 Supplementary Figure 2). To further verify the quality of assemblies generated using SPAdes
329 3.7.0, we used Mauve to align Velvet and SPAdes assembled genome sequences of isolate 13-
330 626 to the finished genome sequence of the reference sequence of *A. radiobacter* K84 (Darling et
331 al. 2004; Slater et al. 2009). The SPAdes-based assembly was superior in being less fragmented
332 and we were able to elevate the quality of the assembly from “unusable” to “high quality”
333 (Supplementary Figures 2 and 3). Therefore, there is greater confidence in concluding that
334 isolate 13-626 lacks the *vir* genes and T-DNA sequence. It does however have an ~200 kb
335 plasmid sequence which encodes *nocM* and *nocP*; this contig also encodes sequences common to

336 replication origins of plasmids. We therefore suggest that because an isolate from the same pear
337 gall sample originally tested positive for *virD2*, we mistakenly sequenced a non-pathogenic
338 isolate.

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

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

358

359 **Conclusions**

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

366

367 **MATERIAL AND METHODS**

368

369 **Website framework and bioinformatics tools**

370 The Gall-ID website and corresponding R shiny server backend are based on the
371 Microbe-ID platform (Tabima et al. 2016) but includes major additions and modifications: Auto
372 MLSA, Auto ANI, BLAST with MAFFT, and the WGS Pipeline. The MLSA framework
373 website was extended to support building Neighbor-Joining trees using incomplete distance
374 matrices (NJ*) using the function `njs()` in the R package PHYLOCH (Paradis et al. 2004). The
375 MLSA framework was also modified to use the multiple sequence alignment program MAFFT
376 using the R package PHYLOCH (Kato & Toh 2008; Heibl 2013; Kato & Standley 2013). This
377 allows user-submitted sequences to be added to pre-existing sequence alignments using the
378 MAFFT "--add" function, to dramatically reduce the computational time required for analysis.

379 The server hosting the Gall-ID tools is running Centos Linux release 6.6, MAFFT version
380 7.221, SRST2 version 0.1.5, Bowtie 2 version 2.2.3, and Samtools version 0.1.18. Gall-ID uses

381 R version 3.1.2 with the following R packages: Poppr version 1.1.0.99 (Kamvar et al. 2014), Ape
382 version 3.1-1 (Paradis et al. 2004), PHYLOCH version 1.5-5, and Shiny version 0.8.0.

383 The Auto MLSA tool was developed previously (Creason et al. 2014a). Briefly, Auto
384 MLSA does the following: BLAST (either TBLASTN or BLASTN) to query NCBI user-
385 selectable databases and/or local databases and retrieve sequences, filter out incomplete sets of
386 gene sequences, align gene sequence individually, concatenate aligned gene sequences,
387 determine the best substitution model (for amino acid sequences), filter out identical sequences,
388 append key information to sequences, and generate a partition file for RAxML (Stamatakis
389 2014). Auto MLSA also has the option of using Gblocks to trim alignments (Castresana 2000).
390 Auto MLSA was modified to use the NCBI E-utilities, implemented in BioPerl, to associate
391 accession numbers with taxon IDs, species names, and assembly IDs (Stajich 2002). For
392 organisms without taxon identifiers, Auto MLSA will attempt to extract meaningful genus and
393 species information from the NCBI nucleotide entry. Gene sequences are linked together using
394 assembly IDs, which allows for genomes with multiple chromosomes to be compared, without
395 having to rely on potentially ambiguous organism names. When assembly IDs are unavailable,
396 whole genome sequences are linked using the four letter WGS codes, and, as a last resort,
397 sequences will be associated using their nucleotide accession number. The disadvantage of using
398 the latter approach is that organisms with multiple replicons, each with its own accession
399 number, will be excluded from analysis. Auto MLSA is available for download from the Gall-ID
400 website. Detailed instructions for using the tools are provided.

401 The Auto ANI script automates the calculations of ANI for all pairwise combinations for
402 any number of input genome sequences. Each of the supplied genome sequences are chunked
403 into 1020 nt fragments and used as queries in all possible reciprocal pairwise BLAST searches.

404 Parameters for genome chunk size, percent identity, and percent coverage have default values set
405 according to published guidelines but can be changed by the user (Goris et al. 2007; Creason et
406 al. 2014a). BLAST version 2.2.31+ was used with recommended settings and previously
407 described in Creason et al. (2014): `-task blastn -dust no -xdrop_gap 150 -penalty -1 -reward 1`
408 `-gapopen 5 -gapextend 2` (Goris et al. 2007). BLAST hits above the user-specified cut-offs (30%
409 identity, 70% coverage, by default) are averaged to calculate the pairwise ANI values.

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

414 Sequences downloaded from NCBI are linked using assembly IDs. All accession types
415 from NCBI are supported, assuming accession numbers are provided in the header line of the
416 FASTA file. Locally generated genome sequences are also supported, in FASTA format,
417 provided they follow the specified header format listed in the user guide. Alternatively, an
418 auxiliary script is provided to rename headers within user-generated FASTA files to the
419 supported format.

420 The WGS Pipeline was written in bash shell script and Perl. Paired Illumina sequencing
421 reads located in the "reads" folder of the pipeline are processed in pairs. The program SMALT
422 (Ponstingl, 2013) is used to align reads to a reference genome and produce CIGAR format output
423 files (Ponstingl 2013). The SSAHA_pileup program converts the CIGAR format files into
424 individual pileup files (Ning et al. 2001). The pileup output is then combined with any additional
425 supplied pre-computed pileup files and used to produce a core alignment of sites shared by 90%
426 of the represented isolates. The optional "remove_recombination.sh" script runs the program

427 Gubbins (Croucher et al. 2014) to remove sites identified as potentially acquired by
428 recombination. Finally, the program RAxML is used to produce a maximum-likelihood
429 phylogenetic tree with non-parametric bootstrap support (Stamatakis 2014). By default 20
430 maximum likelihood tree searches are performed, and the "autoMRE" criterion is used to
431 determine the number of non-parametric bootstrap replicates.

432 The WGS Pipeline test analysis was performed and benchmarked using 10 cores of a
433 cluster server running Centos Linux release 6.6 and containing four AMD Opteron™ 6376 2.3
434 Ghz processors (64 cores total) and 512 GB of RAM (Table 2). The versions of the tools used in
435 tests of this pipeline were Perl version 5.10.1, SMALT version 0.7.6, SSAHA_pileup version
436 0.6, Gubbins version 1.1.2, and RAxML version 8.1.17. The default parameters for WGS
437 Pipeline were used (20 ML search trees, "autoMRE" cutoff for bootstrap replicates) with the
438 exception that the maximum-allowed percentage gaps in the Gubbins recombination analysis
439 was increased to 50% in order to retain strain D188. The WGS Pipeline scripts were also
440 modified to not ask for user input on the command line in order to run in a Sun Grid Engine
441 (SGE) cluster environment.

442 Vir-Search uses the program SRST2, which employs Bowtie 2 and Samtools, with the "--
443 gene_db" function to align the reads to custom databases of the virulence genes (Li et al. 2009;
444 Inouye et al. 2012; Langmead & Salzberg 2012; Inouye et al. 2014). The identity of the virulence
445 genes that the reads input by the user align to, the read coverage and depth, and the name of the
446 strain corresponding to the most similar allele are parsed from the SRST2 output and reported to
447 the user as a static webpage. Users are emailed a link to results once the analysis is complete.
448 The submitted sequencing reads are deleted from the server immediately after completion, and
449 results are available only to those with a direct link to the results webpage.

450

451 **Datasets**

452 The 16S and MLSA gene sequences were downloaded from the genome sequences of the
453 following reference strains: *Agrobacterium* strain C58, *Rhodococcus* strain A44a, *P. savastanoi*
454 *pv. phaseolicola* 1448A, and *P. agglomerans* strain LMAE-2, *C. michiganensis* subsp.
455 *nebraskensis* NCPPB 2581, *D. dadantii* strain 3937, *P. atroscopicum* strain 21A, *R.*
456 *solanacearum* strain GMI1000, *X. oryzae* *pv. oryzicola* strain CFBP2286, and *X. fastidiosa*
457 subsp. *fastidiosa* GB514 (NCBI assembly ID: GCF_000092025.1, GCF_000760735.1,
458 GCF_000012205.1, GCF_000814075.1, GCF_000355695.1, GCF_000147055.1,
459 GCF_000740965.1, GCF_000009125.1, GCF_001042735.1, and GCF_000148405.1,
460 respectively). The gene sequences were used as input for the Auto MLSA tool in BLAST
461 searches carried out against complete genome sequences in the NCBI non-redundant (nr) and
462 whole genome sequence (wgs) databases. The Auto MLSA parameters were: minimum query
463 coverage of 50% (90% for the 16S plant pathogen dataset) and e-value cutoffs of 1e-5 for nr and
464 1e-50 for wgs. BLAST searches were limited to the genus for the bacteria of interest, with the
465 exceptions of *Agrobacterium*, which was limited to *Rhizobiaceae*, and *P. savastanoi*, which was
466 limited to the *P. syringae* group. BLAST searches were completed in August of 2015. The Auto
467 MLSA tool uses MAFFT aligner to produce multiple sequence alignments for each gene (Kato
468 & Standley 2013). The Gblocks trimmed alignment output of Auto MLSA was not used because
469 Gall-ID aligns user-submitted gene sequences to each full gene alignment (Castresana 2000).

470

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

472 Cultures of *Agrobacterium* were grown overnight in Lysogeny Broth (LB) media at
473 28°C, with shaking at 250 rpm (Table 3). Cells were pelleted by centrifugation and total genomic
474 DNA was extracted using a DNeasy Blood and Tissue kit (Qiagen, Venlo, Netherlands). DNA
475 was quantified using a QuBit Fluorometer (Thermo Fisher, Eugene, Oregon) and libraries were
476 prepared using an Illumina Nextera XT DNA Library Prep kit, according to the instructions of
477 the manufacturer, with the exception that libraries were normalized based on measurements from
478 an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA). Each library was
479 assigned an individual barcode using an Illumina Nextera XT Index kit. Libraries were
480 multiplexed and sequenced on an Illumina MiSeq to generate 300 bp paired-end reads.
481 Sequencing was done in the Center for Genome Research and Biocomputing Core Facility
482 (Oregon State University, Corvallis, OR). Sickle was used to trim reads based on quality
483 (minimum quality score cutoff of 25, minimum read length 150 bp after trimming) (Joshi & Fass
484 2011). Read quality was assessed prior to and after trimming using FastQC (Andrews 2010).
485 Paired reads for each library were *de novo* assembled using Velvet version 1.2.10 with the short
486 paired read input option (“-shortPaired”), estimated expected coverage (“-exp_cov auto”), and
487 default settings for other parameters (Zerbino & Birney 2008). Genome sequences were
488 assembled using a range of input hash lengths (k-mer sizes), and the final assembly for each
489 isolate was identified based on those with the best metrics for the following parameters: total
490 assembly length (5.0~7.0 Mb), number of contigs, and N50. Paired reads for each library were
491 error corrected and assembled using SPAdes versions 3.6.2 and 3.7.0, with the careful option (“--
492 careful”) and kmers 21, 33, 55, 77, and 99. Scaffolds shorter than 500bp and with coverage less
493 than 5X were removed from the SPAdes assemblies prior to analysis.
494

495 ACKNOWLEDGEMENTS

496 We thank Melodie Putnam for providing the 14 bacterial isolates and for critical reading
497 of the manuscript. We thank Dr. Pankaj Jaiswal for organizing and inviting us to participate in
498 the STEM DNA Biology and Bioinformatics summer camps (Oregon State University). Camp
499 participants, Ana Bechtel, Mason Hall, Reagan Hunt, Pranav Kolluri, Benjamin Phelps, Joshua
500 Phelps, Aravind Sriram, Megan Thorpe, and eight others, prepared genomic DNA and libraries
501 for whole genome sequencing and analyzed the data. We thank Charlie DuBois of Illumina for
502 providing kits for library preparation as well as sequencing, and Mark Dasenko, Matthew
503 Peterson, and Chris Sullivan of the Center for Genome Research and Biocomputing for
504 sequencing, data processing, and computing services. Finally, we thank the Department of
505 Botany and Plant Pathology for supporting the computational infrastructure.

506

507 FUNDING STATEMENT

508 This work was supported by the National Institute of Food and Agriculture, U.S.
509 Department of Agriculture award 2014-51181-22384 (JHC and NJG). Partial support also was
510 provided by the USDA Agricultural Research Service Grant 5358-22000-039-00D (NJG),
511 USDA National Institute of Food and Agriculture Grant 2011-68004-30154 (NJG), and the
512 USDA ARS Floriculture Nursery Research Initiative (NJG). EWD is supported by a Provost's
513 Distinguished Graduate Fellowship awarded by Oregon State University. This material is based
514 upon work supported by the National Science Foundation Graduate Research Fellowship under
515 Grant No. DGE-1314109 to EDW. The summer camps were supported by National Science
516 Foundation awards IOS #1340112 and #1127112 to Pankaj Jaiswal. Any opinion, findings, and
517 conclusions or recommendations expressed in this material are those of the authors(s) and do not

518 necessarily reflect the views of the U. S. Department of Agriculture or National Science
519 Foundation.

520

521 **FIGURE LEGENDS**

522 **Figure 1. Overview of Gall-ID diagnostic tools.** DNA sequence information can be used to
523 reveal the identity of the causative agent (unknown isolate) of disease. Tools associated with
524 "Gall Isolate Typing" and "Phytopath-type" use 16S rDNA or pathogen-specific MLSA gene
525 sequences to infer the identity of the isolate by comparing the sequences to manually curated
526 sequence databases. Tools associated with "Whole Genome Analysis" and "Vir-Search" use
527 Illumina short sequencing reads to characterize pathogenic isolates. The former tab provides
528 downloadable tools to infer genetic relatedness based on SNPs (WGS Pipeline) or average
529 nucleotide identity (Auto ANI). The "Vir-Search" tab provides an on-line tool to quickly map
530 short reads against a database of sequences of virulence genes.

531

532 **Figure 2. Flowchart for the WGS Pipeline.** Scripts and the programs that each script runs are
533 boxed and presented along the left. The logic flow of the WGS Pipeline tool is presented along
534 the right. Rectangles with rounded corners = inputs and outputs; boxes outlined in red =
535 processes. The inputs, outputs, and processes are matched to the corresponding script and
536 program.

537

538 **Figure 3. Validation of the Agro-type and Vir-Search tools. A)** An unrooted Neighbor Joining
539 phylogenetic tree based on 16s rDNA sequences from *Agrobacterium* spp. The 16S rDNA
540 sequence was identified and extracted from the genome assembly of *Agrobacterium* isolate 13-
541 2099-1-2 and analyzed using the tool available in the Agro-type tab. The isolate is labeled in red,
542 as "query_isolate"; inset shows the clade that circumscribes the isolate. **B)** Screenshot of output
543 results from Vir-Search. Paired 2x300 bp MiSeq short reads from *Agrobacterium* isolate 13-
544 2099-1-2 were analyzed using the Vir-Search tool in Gall-ID. Reference virulence gene
545 sequences that were aligned are indicated with a green plus "+" icon and the lengths and depths
546 of the read coverage are reported (must exceed user-specified cutoffs, which were designated as
547 90% minimum coverage and 20% maximum sequence divergence). Virulence genes that failed
548 to exceed user-specific cutoffs for read alignment parameters are indicated with a red "X".
549 Virulence genes are grouped into categories based on their function in virulence.

550

551 **Figure 4. Maximum likelihood tree based on vertically inherited polymorphic sites core to**
552 **20 *Rhodococcus* isolates.** WGS Pipeline was used to automate the processing of paired end short
553 reads from 20 previously sequenced *Rhodococcus* isolates, and generate a maximum likelihood
554 unrooted tree. Sequencing reads were aligned, using *R. fascians* strain A44a as a reference.
555 SNPs potentially acquired via recombination were removed. The tree is midpoint-rooted. Scale
556 bar = 0.05 average substitutions per site; non-parametric bootstrap support as percentages are
557 indicated for each node. Major clades and sub-clades are labeled in a manner consistent with
558 previous labels.

559

560 REFERENCES

- 561 Adékambi T, Butler RW, Hanrahan F, Delcher AL, Drancourt M, and Shinnick TM. 2011. Core
562 gene set as the basis of multilocus sequence analysis of the subclass *Actinobacteridae*.
563 *PLoS One* 6:e14792. 10.1371/journal.pone.0014792
- 564 Alexandre A, Laranjo M, Young JP, and Oliveira S. 2008. *dnaJ* is a useful phylogenetic marker
565 for *Alphaproteobacteria*. *International Journal of Systematic and Evolutionary*
566 *Microbiology* 58:2839-2849. 10.1099/ijs.0.2008/001636-0
- 567 Allardet-Servent A, Michaux-Charachon S, Jumas-Bilak E, Karayan L, and Ramuz M. 1993.
568 Presence of one linear and one circular chromosome in the *Agrobacterium tumefaciens*
569 C58 genome. *Journal of Bacteriology* 175:7869-7874.
- 570 Alvarez AM. 2004. Integrated approaches for detection of plant pathogenic bacteria and
571 diagnosis of bacterial diseases. *Annual Review of Phytopathology* 42:339-366.
572 doi:10.1146/annurev.phyto.42.040803.140329
- 573 Andrews S. 2010. FastQC A quality control tool for high throughput sequence data. Available at
574 <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
- 575 Aragon IM, Perez-Martinez I, Moreno-Perez A, Cerezo M, and Ramos C. 2014. New insights
576 into the role of indole-3-acetic acid in the virulence of *Pseudomonas savastanoi* pv.
577 *savastanoi*. *FEMS Microbiol Letters* 356:184-192. 10.1111/1574-6968.12413
- 578 Bankevich A, Nurk S, Antipov D, Gurevich A, Dvorkin M, Kulikov A, Lesin V, Nikolenko S,
579 Pham S, Prjibelski A, Pyshkin A, Sirotkin A, Vyahhi N, Tesler G, Alekseyev M, and
580 Pevzner P. 2012. SPAdes: a new genome assembly algorithm and its applications to
581 single-cell sequencing. *J Comput Biol* 19:455-477. 10.1089/cmb.2012.0021
- 582 Barash I, and Manulis S. 2005. Hrp-dependent biotrophic mechanism of virulence: How has it
583 evolved in tumorigenic bacteria? *Phytoparasitica* 33:317-324. 10.1007/BF02981296
- 584 Barash I, and Manulis-Sasson S. 2007. Virulence mechanisms and host specificity of gall-
585 forming *Pantoea agglomerans*. *Trends in Microbiology* 15:538-545.
586 10.1016/j.tim.2007.10.009
- 587 Barash I, Panijel M, Gurel F, Chalupowicz L, and Manulis S. 2005. Transformation of *Pantoea*
588 *agglomerans* into a tumorigenic pathogen. In: Sorvari S, and Toldi O, editors.
589 Proceedings of the 1st International Conference on Plant-Microbe Interactions:
590 Endophytes and Biocontrol Agents. Lapland, Finland: Saariselka. p 10-19.
- 591 Bardaji L, Perez-Martinez I, Rodriguez-Moreno L, Rodriguez-Palenzuela P, Sundin GW, Ramos
592 C, and Murillo J. 2011. Sequence and role in virulence of the three plasmid complement
593 of the model tumor-inducing bacterium *Pseudomonas savastanoi* pv. *savastanoi* NCPPB
594 3335. *PLoS One* 6:e25705. 10.1371/journal.pone.0025705
- 595 Binns AN, and Costantino P. 1998. The *Agrobacterium* oncogenes. In: Spaink HP, Kondorosi A,
596 and Hooykaas PJJ, eds. *The Rhizobiaceae*. Netherlands: Springer, 251-166.
- 597 Bomhoff G, Klapwijk PM, Kester HC, Schilperoort RA, Hernalsteens JP, and Schell J. 1976.
598 Octopine and nopaline synthesis and breakdown genetically controlled by a plasmid of
599 *Agrobacterium tumefaciens*. *Molecular & General Genetics* 145:177-181.
- 600 Bouzar H, and Jones JB. 2001. *Agrobacterium larrymoorei* sp. nov., a pathogen isolated from
601 aerial tumours of *Ficus benjamina*. *International Journal of Systematic and Evolutionary*
602 *Microbiology* 51:1023-1026. 10.1099/00207713-51-3-1023
- 603 Broothaerts W, Mitchell HJ, Weir B, Kaines S, Smith LMA, Yang W, Mayer JE, Roa-Rodríguez
604 C, and Jefferson RA. 2005. Gene transfer to plants by diverse species of bacteria. *Nature*
605 433:629-633. 10.1038/nature03309

- 606 Burr T, Katz B, Abawi G, and Crosier D. 1991. Comparison of tumorigenic strains of *Erwinia*
607 *herbicola* isolated from table beet with *E. h. gypsophilae*. *Plant Disease* 75:855-858.
- 608 Castillo JA, and Greenberg JT. 2007. Evolutionary dynamics of *Ralstonia solanacearum*.
609 *Applied and Environmental Microbiology* 73:1225-1238. 10.1128/AEM.01253-06
- 610 Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in
611 phylogenetic analysis. *Molecular Biology and Evolution* 17:540-552.
- 612 Chan JZ-M, Halachev MR, Loman NJ, Constantinidou C, and Pallen MJ. 2012. Defining
613 bacterial species in the genomic era: insights from the genus *Acinetobacter*. *BMC*
614 *Microbiology* 12:302. 10.1186/1471-2180-12-302
- 615 Chang J-M, Di Tommaso P, and Notredame C. 2014. TCS: A new multiple sequence alignment
616 reliability measure to estimate alignment accuracy and improve phylogenetic tree
617 reconstruction. *Molecular Biology and Evolution* 31:1625-1637. 10.1093/molbev/msu117
- 618 Chilton MD, Drummond MH, Merio DJ, Sciaky D, Montoya AL, Gordon MP, and Nester EW.
619 1977. Stable incorporation of plasmid DNA into higher plant cells: the molecular basis of
620 crown gall tumorigenesis. *Cell* 11:263-271.
- 621 Clark E, Manulis S, Ophir Y, Barash I, and Y G. 1993. Cloning and characterization of *iaaM* and
622 *iaaH* from *Erwinia herbicola* pathovar *gypsophilae*. *Phytopathology* 83:234-240.
623 10.1094/Phyto-83-234
- 624 Cooksey DA. 1986. Galls of *Gypsophila paniculata* caused by *Erwinia herbicola*. *Plant Disease*
625 70:464. 10.1094/PD-70-464
- 626 Creason AL, Davis EW, Putnam ML, Vandeputte OM, and Chang JH. 2014a. Use of whole
627 genome sequences to develop a molecular phylogenetic framework for *Rhodococcus*
628 *fascians* and the *Rhodococcus* genus. *Frontiers in Plant Science* 5:406.
629 10.3389/fpls.2014.00406
- 630 Creason AL, Vandeputte OM, Savory EA, Davis EW, Putnam ML, Hu E, Swader-Hines D, Mol
631 A, Baucher M, Prinsen E, Zdanowska M, Givan SA, Jaziri ME, Loper JE, Mahmud T,
632 and Chang JH. 2014b. Analysis of genome sequences from plant pathogenic
633 *Rhodococcus* reveals genetic novelties in virulence loci. *PLoS One* 9:e101996.
634 10.1371/journal.pone.0101996
- 635 Crespi M, Messens E, Caplan AB, van Montagu M, and Desomer J. 1992. Fasciation induction
636 by the phytopathogen *Rhodococcus fascians* depends upon a linear plasmid encoding a
637 cytokinin synthase gene. *The EMBO Journal* 11:795-804.
- 638 Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, and Harris
639 SR. 2014. Rapid phylogenetic analysis of large samples of recombinant bacterial whole
640 genome sequences using Gubbins. *Nucleic Acids Research*:gku1196-.
641 10.1093/nar/gku1196
- 642 Darling A, Mau B, Blattner F, and Perna N. 2004. Mauve: multiple alignment of conserved
643 genomic sequence with rearrangements. *Genome Res* 14:1394-1403. 10.1101/gr.2289704
- 644 Delétoile A, Decré D, Courant S, Passet V, Audo J, Grimont P, Arlet G, and Brisse S. 2009.
645 Phylogeny and identification of *Pantoea* species and typing of *Pantoea agglomerans*
646 strains by multilocus gene sequencing. *Journal of Clinical Microbiology* 47:300-310.
647 10.1128/JCM.01916-08
- 648 DeYoung RM, Copeman RJ, and Hunt RS. 1998. Two strains in the genus *Erwinia* cause galls
649 on Douglas-fir in southwestern British Columbia. *Canadian Journal of Plant Pathology*
650 20:194-200. 10.1080/07060669809500427

- 651 Farrand SK, Van Berkum PB, and Oger P. 2003. *Agrobacterium* is a definable genus of the
652 family *Rhizobiaceae*. *International Journal of Systematic and Evolutionary Microbiology*
653 53:1681-1687. 10.1099/ijs.0.02445-0
- 654 Gardan L, Bollet C, Abu Ghorrah M, Grimont F, and Grimont PAD. 1992. DNA relatedness
655 among the pathovar strains of *Pseudomonas syringae* subsp. *savastanoi* Janse (1982) and
656 proposal of *Pseudomonas savastanoi* sp. nov. *International Journal of Systematic*
657 *Bacteriology* 42:606-612. 10.1099/00207713-42-4-606
- 658 Gloyer WO. 1934. *Crown gall and hairy root of apples in nursery and orchard*. Geneva, NY:
659 New York Agricultural Experiment Station Bulletin 638.
- 660 Goodfellow M. 1984. Reclassification of *Corynebacterium fascians* (Tilford) Dowson in the
661 genus *Rhodococcus*, as *Rhodococcus fascians* comb. nov. *Systematic and Applied*
662 *Microbiology* 5:225-229. 10.1016/S0723-2020(84)80023-5
- 663 Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, and Tiedje JM. 2007.
664 DNA-DNA hybridization values and their relationship to whole-genome sequence
665 similarities. *International Journal of Systematic and Evolutionary Microbiology* 57:81-
666 91. 10.1099/ijs.0.64483-0
- 667 Heibl C. 2013. PHYLOCH: interfaces and graphic tools for phylogenetic data in R. Available at
668 <http://www.christopheheibl.de/Rpackages.html>.
- 669 Hildebrand EM. 1940. Cane gall of Brambles caused by *Phytomonas rubi* n.sp. *Journal of*
670 *Agricultural Research* 61:685--696 pp.
- 671 Hwang MSH, Morgan RL, Sarkar SF, Wang PW, and Guttman DS. 2005. Phylogenetic
672 characterization of virulence and resistance phenotypes of *Pseudomonas syringae*.
673 *Applied and Environmental Microbiology* 71:5182-5191. 10.1128/AEM.71.9.5182-
674 5191.2005
- 675 Iacobellis NS, Sisto A, Surico G, Evidente A, and DiMaio E. 1994. Pathogenicity of
676 *Pseudomonas syringae* subsp. *savastanoi* mutants defective in phytohormone production.
677 *Journal of Phytopathology* 140:238-248. 10.1111/j.1439-0434.1994.tb04813.x
- 678 Inouye M, Conway TC, Zobel J, and Holt KE. 2012. Short read sequence typing (SRST): multi-
679 locus sequence types from short reads. *BMC Genomics* 13:338. 10.1186/1471-2164-13-
680 338
- 681 Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ, Tomita T, Zobel J, and Holt KE.
682 2014. SRST2: Rapid genomic surveillance for public health and hospital microbiology
683 labs. *Genome Medicine* 6:90. 10.1186/s13073-014-0090-6
- 684 Jacques M-A, Durand K, Orgeur G, Balidas S, Fricot C, Bonneau S, Quillévéré A, Audusseau C,
685 Olivier V, Grimault V, and Mathis R. 2012. Phylogenetic analysis and polyphasic
686 characterization of *Clavibacter michiganensis* strains isolated from tomato seeds reveal
687 that nonpathogenic strains are distinct from *C. michiganensis* subsp. *michiganensis*.
688 *Applied and Environmental Microbiology* 78:8388-8402. 10.1128/AEM.02158-12
- 689 Janda JM, and Abbott SL. 2007. 16S rRNA gene sequencing for bacterial identification in the
690 diagnostic laboratory: pluses, perils, and pitfalls. *Journal of Clinical Microbiology*
691 45:2761-2764. 10.1128/JCM.01228-07
- 692 Joshi N, and Fass J. 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for
693 FastQ files. Available at <https://github.com/najoshi/sickle>.
- 694 Kado CI. 2014. Historical account on gaining insights on the mechanism of crown gall
695 tumorigenesis induced by *Agrobacterium tumefaciens*. *Frontiers in Microbiology* 5:340.
696 10.3389/fmicb.2014.00340

- 697 Kamvar ZN, Tabima JF, and Grünwald NJ. 2014. Poppr: an R package for genetic analysis of
698 populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* 2:e281.
699 10.7717/peerj.281
- 700 Katoh K, and Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
701 Improvements in performance and usability. *Molecular Biology and Evolution* 30:772-
702 780. 10.1093/molbev/mst010
- 703 Katoh K, and Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment
704 program. *Briefings in Bioinformatics* 9:286-298. 10.1093/bib/bbn013
- 705 Kim H-S, Ma B, Perna NT, and Charkowski AO. 2009. Phylogeny and virulence of naturally
706 occurring type III secretion system-deficient *Pectobacterium* strains. *Applied and*
707 *Environmental Microbiology* 75:4539-4549. 10.1128/AEM.01336-08
- 708 Kim M, Oh H-S, Park S-C, and Chun J. 2014. Towards a taxonomic coherence between average
709 nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of
710 prokaryotes. *International Journal of Systematic and Evolutionary Microbiology* 64:346-
711 351. 10.1099/ijs.0.059774-0
- 712 Langmead B, and Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature*
713 *Methods* 9:357-359. 10.1038/nmeth.1923
- 714 Lassalle F, Campillo T, Vial L, Baude J, Costechareyre D, Chapulliot D, Shams M, Abrouk D,
715 Lavire C, Oger-Desfeux C, Hommais F, Guéguen L, Daubin V, Muller D, and Nesme X.
716 2011. Genomic species are ecological species as revealed by comparative genomics in
717 *Agrobacterium tumefaciens*. *Genome Biology and Evolution* 3:762-781.
718 10.1093/gbe/evr070
- 719 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R,
720 and Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and
721 SAMtools. *Bioinformatics* 25:2078-2079.
- 722 Lichter A, Barash I, Valinsky L, and Manulis S. 1995. The genes involved in cytokinin
723 biosynthesis in *Erwinia herbicola* pv. *gypsophylae*: characterization and role in gall
724 formation. *Journal of Bacteriology* 177:4457-4465.
- 725 Maes T, Vereecke D, Ritsema T, Cornelis K, Thu HN, Van Montagu M, Holsters M, and
726 Goethals K. 2001. The *att* locus of *Rhodococcus fascians* strain D188 is essential for full
727 virulence on tobacco through the production of an autoregulatory compound. *Molecular*
728 *microbiology* 42:13-28.
- 729 Mansfield J, Genin S, Magori S, Citovsky V, Sriariyanum M, Ronald P, Dow M, Verdier V,
730 Beer SV, Machado MA, Toth I, Salmond G, and Foster GD. 2012. Top 10 plant
731 pathogenic bacteria in molecular plant pathology. *Molecular Plant Pathology* 13:614-
732 629. 10.1111/j.1364-3703.2012.00804.x
- 733 Manulis S, and Barash I. 2003. The molecular basis for transformation of an epiphyte into a gall-
734 forming pathogen as exemplified by *Erwinia herbicola* pv. *gypsophylae*. *Plant-Microbe*
735 *Interactions* 6:19-52.
- 736 Manulis S, Haviv-Chesner A, Brandl MT, Lindow SE, and Barash I. 1998. Differential
737 involvement of indole-3-acetic acid biosynthetic pathways in pathogenicity and epiphytic
738 fitness of *Erwinia herbicola* pv. *gypsophylae*. *Molecular Plant-Microbe Interactions*
739 11:634-642. 10.1094/MPMI.1998.11.7.634
- 740 Marrero G, Schneider KL, Jenkins DM, and Alvarez AM. 2013. Phylogeny and classification of
741 *Dickeya* based on multilocus sequence analysis. *International Journal of Systematic and*
742 *Evolutionary Microbiology* 63:3524-3539. 10.1099/ijs.0.046490-0

- 743 Matas IM, Lambertsen L, Rodríguez-Moreno L, and Ramos C. 2012. Identification of novel
744 virulence genes and metabolic pathways required for full fitness of *Pseudomonas*
745 *savastanoi* pv. *savastanoi* in olive (*Olea europaea*) knots. *New Phytologist* 196:1182-
746 1196. 10.1111/j.1469-8137.2012.04357.x
- 747 Montoya AL, Chilton MD, Gordon MP, Sciaky D, and Nester EW. 1977. Octopine and nopaline
748 metabolism in *Agrobacterium tumefaciens* and crown gall tumor cells: role of plasmid
749 genes. *Journal of Bacteriology* 129:101-107.
- 750 Mor H, Manulis S, Zuck M, Nizan R, Coplin DL, and Barash I. 2001. Genetic organization of
751 the *hrp* gene cluster and *dspAE/BF* operon in *Erwinia herbicola* pv. *gypsophylae*.
752 *Molecular Plant-Microbe Interactions* 14:431-436. 10.1094/MPMI.2001.14.3.431
- 753 Morris RO. 1986. Genes specifying auxin and cytokinin biosynthesis in phytopathogens. *Annual*
754 *Review of Plant Physiology* 37:509-538. 10.1146/annurev.pp.37.060186.002453
- 755 Ning Z, Cox AJ, and Mullikin JC. 2001. SSAHA: a fast search method for large DNA databases.
756 *Genome Research* 11:1725-1729. 10.1101/gr.194201
- 757 Nissan G, Manulis-Sasson S, Weinthal D, Mor H, Sessa G, and Barash I. 2006. The type III
758 effectors HsvG and HsvB of gall-forming *Pantoea agglomerans* determine host
759 specificity and function as transcriptional activators. *Molecular microbiology* 61:1118-
760 1131. 10.1111/j.1365-2958.2006.05301.x
- 761 Nizan R, Barash I, Valinsky L, Lichter A, and Manulis S. 1997. The presence of *hrp* genes on
762 the pathogenicity-associated plasmid of the tumorigenic bacterium *Erwinia herbicola* pv.
763 *gypsophylae*. *Molecular Plant-Microbe Interactions* 10:677-682.
764 10.1094/MPMI.1997.10.5.677
- 765 Nizan-Koren R, Manulis S, Mor H, Iraki NM, and Barash I. 2003. The regulatory cascade that
766 activates the Hrp regulon in *Erwinia herbicola* pv. *gypsophylae*. *Molecular Plant-*
767 *Microbe Interactions* 16:249-260. 10.1094/MPMI.2003.16.3.249
- 768 Opgenorth DC, Henderson M, and Clark E. 1994. First report of bacterial gall of *Wisteria sinensis*
769 caused by *Erwinia herbicola* pv. *milletiae* in California. *Plant Disease* 78:1217C.
770 10.1094/PD-78-1217C
- 771 Ophel K, and Kerr A. 1990. *Agrobacterium vitis* sp. nov. for Strains of *Agrobacterium* biovar 3
772 from Grapevines. *International Journal of Systematic Bacteriology* 40:236-241.
773 10.1099/00207713-40-3-236
- 774 Paradis E, Claude J, and Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R
775 language. *Bioinformatics* 20:289-290. 10.1093/bioinformatics/btg412
- 776 Parker JK, Havird JC, and De La Fuente L. 2012. Differentiation of *Xylella fastidiosa* strains via
777 multilocus sequence analysis of environmentally mediated genes (MLSA-E). *Applied and*
778 *Environmental Microbiology* 78:1385-1396. 10.1128/AEM.06679-11
- 779 Pearson T, Okinaka RT, Foster JT, and Keim P. 2009. Phylogenetic understanding of clonal
780 populations in an era of whole genome sequencing. *Infection, genetics and evolution :*
781 *journal of molecular epidemiology and evolutionary genetics in infectious diseases*
782 9:1010-1019. 10.1016/j.meegid.2009.05.014
- 783 Pérez-Yépez J, Armas-Capote N, Velázquez E, Pérez-Galdona R, Rivas R, and León-Barrios M.
784 2014. Evaluation of seven housekeeping genes for multilocus sequence analysis of the
785 genus *Mesorhizobium*: Resolving the taxonomic affiliation of the *Cicer canariense*
786 rhizobia. *Systematic and Applied Microbiology* 37:553-559.
787 10.1016/j.syapm.2014.10.003
- 788 Ponstingl H. 2013. SMALT. Available at <https://www.sanger.ac.uk/resources/software/smalt/>.

- 789 Putnam ML, and Miller ML. 2007. *Rhodococcus fascians* in herbaceous perennials. *Plant*
790 *Disease* 91:1064-1076. 10.1094/PDIS-91-9-1064
- 791 Sachs T. 1975. Plant tumors resulting from unregulated hormone synthesis. *Journal of*
792 *Theoretical Biology* 55:445-453.
- 793 Sarkar SF, and Guttman DS. 2004. Evolution of the core genome of *Pseudomonas syringae*, a
794 highly clonal, endemic plant pathogen. *Applied and Environmental Microbiology*
795 70:1999-2012.
- 796 Schroth MN. 1988. Reduction in yield and vigor of grapevine caused by crown gall disease.
797 *Plant Disease* 72:241. 10.1094/PD-72-0241
- 798 Sisto A, Cipriani MG, and Morea M. 2004. Knot Formation caused by *Pseudomonas syringae*
799 subsp. *savastanoi* on olive plants is *hrp*-dependent. *Phytopathology* 94:484-489.
800 10.1094/PHYTO.2004.94.5.484
- 801 Slater S, Goldman B, Goodner B, Setubal J, Farrand S, Nester E, Burr T, Banta L, Dickerman A,
802 Paulsen I, Otten L, Suen G, Welch R, Almeida N, Arnold F, Burton O, Du Z, Ewing A,
803 Godsy E, Heisel S, Houmiel K, Jhaveri J, Lu J, Miller N, Norton S, Chen Q,
804 Phoolcharoen W, Ohlin V, Ondrusek D, Pride N, Stricklin S, Sun J, Wheeler C, Wilson
805 L, Zhu H, and Wood D. 2009. Genome sequences of three agrobacterium biovars help
806 elucidate the evolution of multichromosome genomes in bacteria. *J Bacteriol* 191:2501-
807 2511. 10.1128/JB.01779-08
- 808 Stackebrandt E, and Goebel BM. 1994. Taxonomic Note: A place for DNA-DNA reassociation
809 and 16S rRNA sequence analysis in the present species definition in bacteriology.
810 *International Journal of Systematic Bacteriology* 44:846-849. 10.1099/00207713-44-4-
811 846
- 812 Stajich JE. 2002. The Bioperl Toolkit: Perl modules for the life sciences. *Genome Research*
813 12:1611-1618. 10.1101/gr.361602
- 814 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
815 large phylogenies. *Bioinformatics* 30:1312-1313. 10.1093/bioinformatics/btu033
- 816 Stes E, Francis I, Pertry I, Dolzblasz A, Depuydt S, and Vereecke D. 2013. The leafy gall
817 syndrome induced by *Rhodococcus fascians*. *FEMS Microbiol Lett* 342:187-195.
818 10.1111/1574-6968.12119
- 819 Tabima JF, Everhart, SE, Larsen, MM, Weisberg, AJ, Kamvar, ZN, Tancos, MA, Smart, CD,
820 Chang, JH, and Grünwald, NJ. 2016. Microbe-ID: An open source toolbox for microbial
821 genotyping and species identification. *PeerJ Preprints* 10.7287/peerj.preprints.2005v1
- 822 Tancos MA, Lange HW, and Smart CD. 2015. Characterizing the genetic diversity of the
823 *Clavibacter michiganensis* subsp. *michiganensis* population in New York.
824 *Phytopathology* 105:169-179. 10.1094/PHYTO-06-14-0178-R
- 825 Temmerman W, Vereecke D, Dreesen R, Van Montagu M, Holsters M, and Goethals K. 2000.
826 Leafy gall formation is controlled by *fasR*, an AraC-type regulatory gene in *Rhodococcus*
827 *fascians*. *Journal of Bacteriology* 182:5832-5840.
- 828 Thompson DV, Melchers LS, Idler KB, Schilperoort RA, and Hooykaas PJ. 1988. Analysis of
829 the complete nucleotide sequence of the *Agrobacterium tumefaciens virB* operon. *Nucleic*
830 *Acids Research* 16:4621-4636.
- 831 Van Larebeke N, Engler G, Holsters M, Van den Elsacker S, Zaenen I, Schilperoort RA, and
832 Schell J. 1974. Large plasmid in *Agrobacterium tumefaciens* essential for crown gall-
833 inducing ability. *Nature* 252:169-170.

- 834 Vasanthakumar A, and McManus PS. 2004. Indole-3-acetic acid-producing bacteria are
835 associated with cranberry stem gall. *Phytopathology* 94:1164-1171.
836 10.1094/PHYTO.2004.94.11.1164
- 837 Velázquez E, Palomo JL, Rivas R, Guerra H, Peix A, Trujillo ME, García-Benavides P, Mateos
838 PF, Wabiko H, and Martínez-Molina E. 2010. Analysis of core genes supports the
839 reclassification of strains *Agrobacterium radiobacter* K84 and *Agrobacterium*
840 *tumefaciens* AKE10 into the species *Rhizobium rhizogenes*. *Systematic and Applied*
841 *Microbiology* 33:247-251. 10.1016/j.syapm.2010.04.004
- 842 Vereecke D, Cornelis K, Temmerman W, Jaziri M, Van Montagu M, Holsters M, and Goethals
843 K. 2002. Chromosomal locus that affects pathogenicity of *Rhodococcus fascians*. *Journal*
844 *of Bacteriology* 184:1112-1120.
- 845 Ward JE, Akiyoshi DE, Regier D, Datta A, Gordon MP, and Nester EW. 1988. Characterization
846 of the *virB* operon from an *Agrobacterium tumefaciens* Ti plasmid. *Journal of Biological*
847 *Chemistry* 263:5804-5814.
- 848 Weinthal DM, Barash I, Panijel M, Valinsky L, Gaba V, and Manulis-Sasson S. 2007.
849 Distribution and replication of the pathogenicity plasmid pPATH in diverse populations
850 of the gall-forming bacterium *Pantoea agglomerans*. *Applied and Environmental*
851 *Microbiology* 73:7552-7561. 10.1128/AEM.01511-07
- 852 Wertz JE, Goldstone C, Gordon DM, and Riley MA. 2003. A molecular phylogeny of enteric
853 bacteria and implications for a bacterial species concept. *Journal of Evolutionary Biology*
854 16:1236-1248.
- 855 Young JM. 2003. Classification and nomenclature of *Agrobacterium* and *Rhizobium* - a reply to
856 Farrand et al. (2003). *International Journal of Systematic and Evolutionary Microbiology*
857 53:1689-1695. 10.1099/ijs.0.02762-0
- 858 Young JM, Kuykendall LD, Martínez-Romero E, Kerr A, and Sawada H. 2001. A revision of
859 *Rhizobium* Frank 1889, with an emended description of the genus, and the inclusion of all
860 species of *Agrobacterium* Conn 1942 and *Allorhizobium undicola* de Lajudie et al. 1998
861 as new combinations: *Rhizobium radiobacter*, *R. rhizogenes*, *R. rubi*, *R. undicola* and *R.*
862 *vitis*. *International Journal of Systematic and Evolutionary Microbiology* 51:89-103.
863 doi:10.1099/00207713-51-1-89
- 864 Young JM, Park D-C, Shearman HM, and Fargier E. 2008. A multilocus sequence analysis of the
865 genus *Xanthomonas*. *Systematic and Applied Microbiology* 31:366-377.
866 10.1016/j.syapm.2008.06.004
- 867 Zeigler DR. 2003. Gene sequences useful for predicting relatedness of whole genomes in
868 bacteria. *International Journal of Systematic and Evolutionary Microbiology* 53:1893-
869 1900.
- 870 Zerbino DR, and Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de
871 Bruijn graphs. *Genome Research* 18:821-829. 10.1101/gr.074492.107
- 872 Zhu J, Oger PM, Schrammeijer B, Hooykaas PJ, Farrand SK, and Winans SC. 2000. The bases
873 of crown gall tumorigenesis. *Journal of Bacteriology* 182:3885-3895.
874

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

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

Table 2 (on next page)

Statistics for the WGS Pipeline

1 **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

2

Table 3 (on next page)

Strain identity of 14 isolates associated with crown gall

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

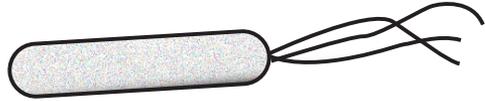
Isolate name	Host	Positive ID based on	# high quality read pairs	Clade (based on 16S rDNA)	# of virulence genes ID'ed
13-2099-1-2	Quaking Aspen	<i>virD2</i> PCR	1,244,074	<i>Agrobacterium</i>	63
13-626	Pear	<i>virD2</i> PCR	220,903	<i>Agrobacterium</i>	2 (<i>nocM</i> , <i>nocP</i>)
AC27/96	Pieris	Not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

2
3

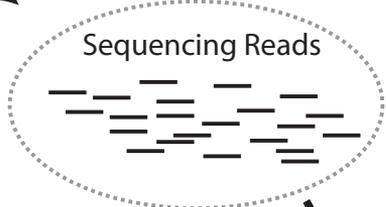
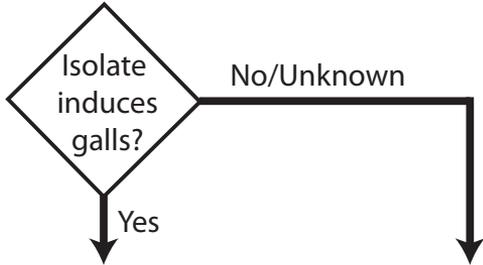
Figure 1(on next page)

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.



16S rRNA or MLSA



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.

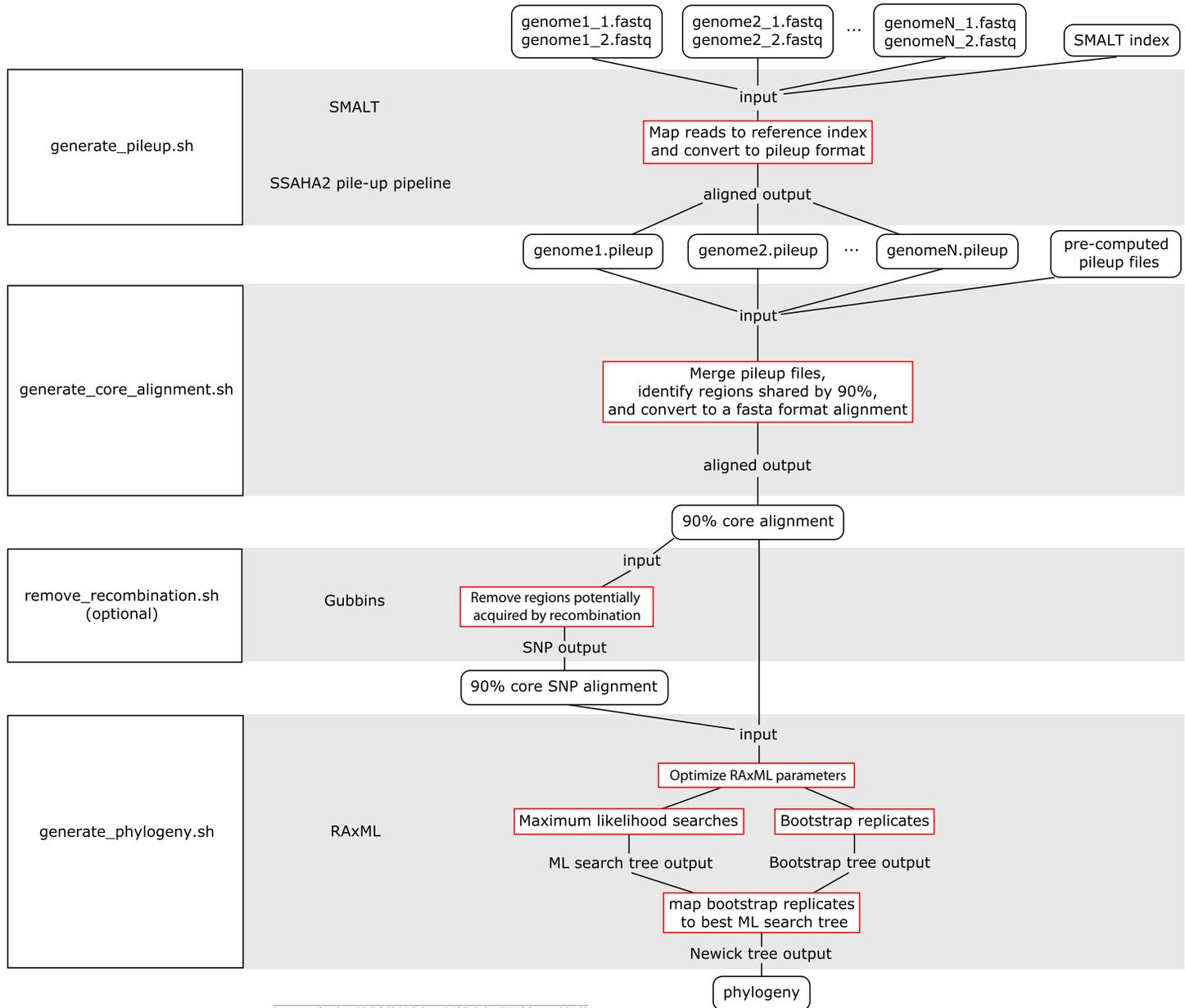


Figure 3(on next page)

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_LJ5258_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: aveisberg@gmail.com

Result table

SRST2 full gene output file: http://gall-ic.ogbr.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 K999

Oncogenes					Opine Synthesis/Transport/Catabolism				
Found	Gene	Coverage (%)	Depth	Closest Allele	Found	Gene	Coverage (%)	Depth	Closest Allele
+	tms2	100.0	31.699	Rhizobium_rubi_NBRC_13261	+	accA	100.0	2,4915	Rhizobium_rubi_NBRC_13261
+	tms1	100.0	27.473	Rhizobium_rubi_NBRC_13261	+	accB	100.0	21.128	Rhizobium_rubi_NBRC_13261

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
X	galls	-	-	-

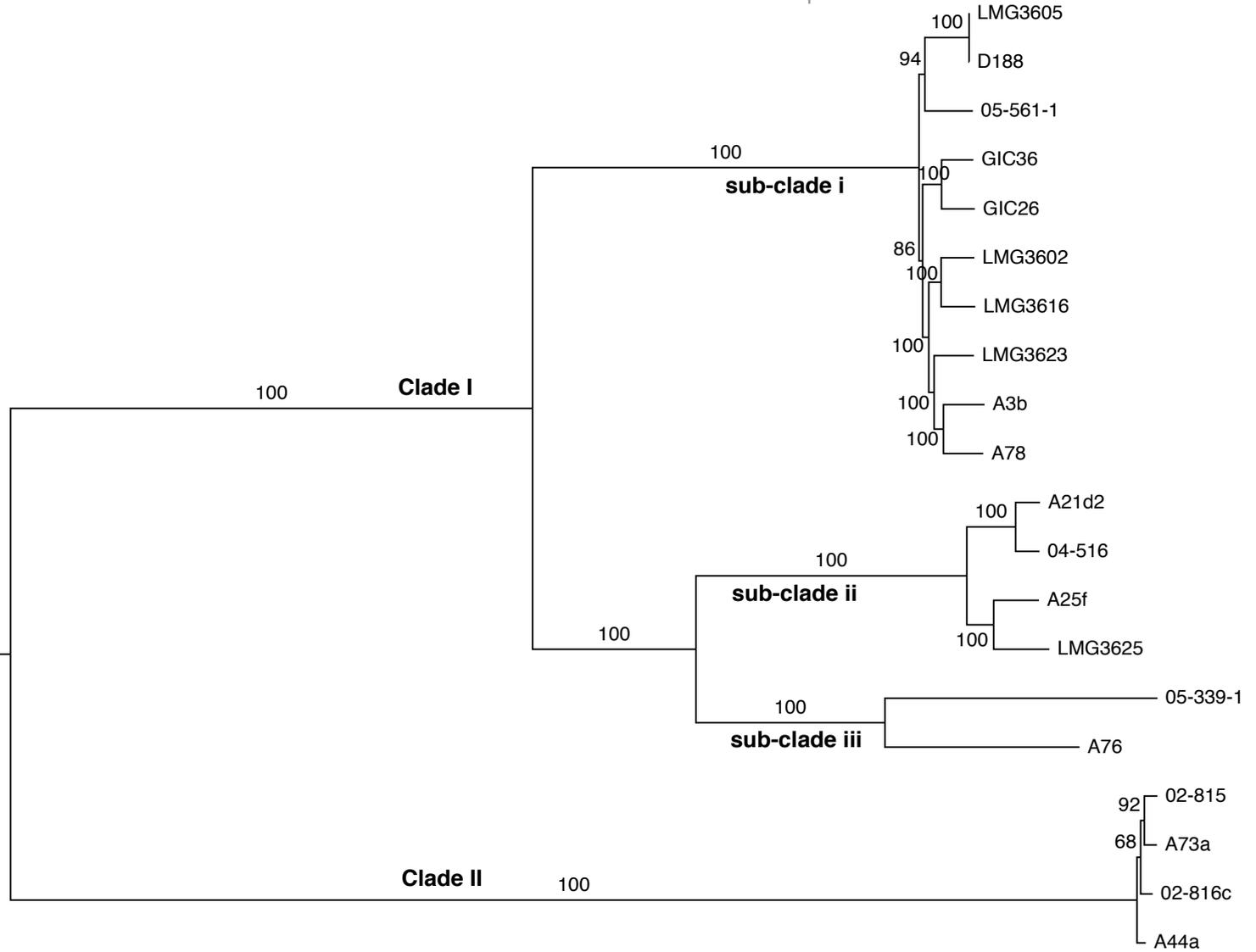
+	virB3	100.0	28.21	Rhizobium_rubi_NBRC_13261	+	nocT	100.0	26.191	Rhizobium_rubi_NBRC_13261
+	virB4	100.0	30.705	Rhizobium_rubi_NBRC_13261	+	nos	100.0	27.164	Rhizobium_rubi_NBRC_13261
+	virB5	100.0	23.976	Agrobacterium_arsenijevicii_strain_KFB_330	+	noxA	100.0	24.153	Rhizobium_rubi_NBRC_13261
+	virB6	100.0	31.687	Agrobacterium_arsenijevicii_strain_KFB_330	+	noxB	100.0	25.276	Rhizobium_rubi_NBRC_13261
+	virB7	100.0	28.272	Agrobacterium_arsenijevicii_strain_KFB_330	+	occJ	100.0	27.164	Rhizobium_rubi_NBRC_13261
+	virB8	100.0	35.109	Rhizobium_rubi_NBRC_13261	+	occM	100.0	63.642	Rhizobium_sp_YR060
+	virB9	100.0	27.213	Rhizobium_rubi_NBRC_13261	+	occP	100.0	66.848	Rhizobium_tropici_strain_YR530
+	virB10	100.0	27.482	Rhizobium_rubi_NBRC_13261	+	occQ	100.0	59.199	Agrobacterium_radiobacter_K84
					X	ocs	-	-	-

89 Agrobacterium_tumefaciens_P4
 Agrobacterium_tumefaciens_WRT31
 query_isolate
 84 Agrobacterium_tumefaciens_5A
 L Agrobacterium_arsenijevicii_strain_Ach5
 Agrobacterium_sp_10MFCol1.1
 Agrobacterium_sp_H19-3
 Agrobacterium_arsenijevicii_strain_KFB_330
 Rhizobium_nepotum_39/7
 Agrobacterium_fabrum_str_C58
 Agrobacterium_tumefaciens_strain_LBA4404
 Agrobacterium_tumefaciens_strain_F4
 Agrobacterium_sp_ATCC_31749
 Agrobacterium_sp_LC34
 Agrobacterium_tumefaciens_strain_SUL3
 Agrobacterium_sp_224MFTsu3.1
 Agrobacterium_sp_UNC420CL41Cvi
 Agrobacterium_tumefaciens_GW4
 Agrobacterium_tumefaciens_CONWGS0286
 Agrobacterium_tumefaciens_F2
 Agrobacterium_tumefaciens_strain_S2
 Agrobacterium_tumefaciens_str_Cherry_2E-2-2
 Agrobacterium_sp_33MFTat.1
 Rhizobium_sp_IRBG74
 Rhizobium_sp_UR51a
 Rhizobium_sp_LMB1
 Rhizobium_lupini_HPC(L)
 Rhizobium_sp_H41
 Agrobacterium_tumefaciens_strain_S33

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.



0.05