

Gall-ID: web-based tools for the rapid identification and characterization of gall-causing phytopathogenic bacteria

Edward W. Davis^{1,2*}, Alexandra J. Weisberg^{1*}, Javier F. Tabima¹, Melodie L. Putnam¹, Niklaus J. Grünwald^{1,2,3}, and Jeff H. Chang^{1,2,4}

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

Abstract

Understanding the genetic diversity of plant pathogens and the effect of agricultural practices on pathogen evolution are important for disease management. Advances in DNA sequencing technology have contributed to greater reliance on use of 16S rDNA, multilocus sequence analysis (MLSA), and whole genome sequences to genotype bacteria. Correct analysis and interpretation of sequencing data can be difficult. Therefore we have developed a set of web-based tools, termed Gall-ID, to facilitate the identification and characterization of phytopathogenic bacteria, with a focus on those that cause gall diseases. Users can compare 16S or MLSA gene sequences from an isolate against manually-curated databases, and generate a phylogenetic tree containing their isolate. Gall-ID also includes a tool for uploading and using whole genome sequencing reads to identify homologs of known virulence genes. Finally, Gall-ID provides downloadable software pipelines for core genome analysis (WGS Pipeline), calculation of average nucleotide identity (Auto ANI), and the generation of MLSA gene set databases (Auto MLSA).

Introduction to Gall-ID

- Web-based tools to easily identify a bacterial isolate using 16S or MLSA gene sequences
- Enables rapid identification without bioinformatics experience
- Downloadable software tools available for core genome and whole genome comparisons
- Available at:

<http://gall-id.cgrb.oregonstate.edu>



Instructions
Select an MLSA dataset, input corresponding FASTA format DNA sequences for one isolate, and select submit. Copy and paste your DNA sequences in FASTA format into the window below. Sequences should be named by gene name (ie ">gyrB"). Not all genes are required to be input, though accuracy will be improved with a complete set of MLSA gene sequences. Use the options in the Analysis section to generate a Distance Tree or Minimum Spanning Network after submitting your data. These trees can be exported as pdf files or in Newick format.

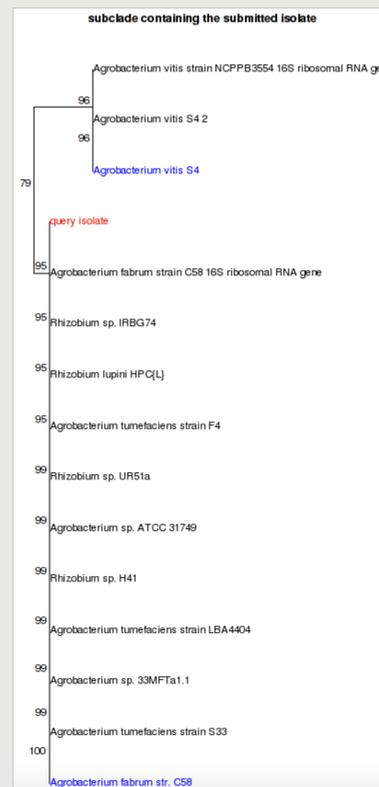
Select Dataset: Choose an MLSA or 16S gene set to compare your input sequences to:
Agrobacterium 16S

FASTA Input

```
>16S
AACGAACGCTGGCCGAGGCTTAACACATGCAAGTCAAGCCGACGAGGAGGAGTGGAGAGCGGTGATGACGCGTGGGAATCTACCCA
TCTCTGGGAAATAGCTCTGGGAACTGGAATTAATACCGCATACGCCCTACGGGGAAGATTTATCGGGATGATGAGCCGGCTGGTGGT
AGCTAGTGGTGGGTAAGGCTACCAAGGGGACGATCATAGCTGGTCTGAGAGAGATGATCAGCCACATTTGGGACTGAGACACGGCCAA
ACTCTACGGAGGCGAGTGGGGAATTTGGACAATGGGCGAAGCGCTGATCCAGCCATGCGCGCTGAGTATGAAAGCGCTAGGGTGT
AAAGCTCTTTCAGCGATAGAGATAATGACGGTACGTGCGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAG
AGCGTGTGGGAAATACGTGGGCTAAAGCCGACGTAGGCGATATTAAAGTCAAGGGGTAATCCGAGCTCACTGGGAGCTGCTCTT
ATACGTGGTATCTTGAGTATGGAAGAGTAAAGTGGAAATCCGAGTGTAGAGGTGAATCGTAGATTGGGGAACACCGATGGGGAAGCG
GCTTACTGTGCTCATTAGCAGCTGAGGTGCGAAGCGTGGGAGGAGCAACAGGATAGATACCCCTGGTGTCCAGCGGTAACAGTGAATGT
TAGCGTGGGCGAGTACTGTTGGTGGGCGAGCTAACGATTAACCAATTCGCGCTGGGAGTACGGTCCGCAAGATTAACCAAGTGAAT
TGACGGGGGCGCGCACAGCGGTTGAGCATGTGTTAATTCGAGAGAACGCGCAGACACTACAGCTCTTGACACTCGGGGTATGGCGATT
GGAGACGATGCTTCAAGTAAAGCTGGGCGCACAGCGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGT
CGAGCGAACCTCCGCTTAGTGGCAGCATTAAGTGGGCACTTAAGGGGACTGCCGGTATAGAGCGAGAGAGAGAGAGAGAGAGAGAG
CAAGCTCATGGCCCTACGGGCGGTGACACAGCTGCTCAATGTTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGT
AAAGCCATCTCAGTTCGATGCTGCAACTCGAGTGGATGAAGTGGATGCTAGTAGTAAAGTGGATGCTGCGGTGGTGAATCGTT
```

Random Seed: Choose a Distance: Kimura 80

Select an organism, a 16S or MLSA gene dataset, and input your sequences



Gall-ID produces and displays a phylogenetic tree containing your isolate for easy identification

16S database

MLSA databases:

- *Agrobacterium*
- *Clavibacter*
- *Dickeya*
- *Pantoea agglomerans*
- *Pectobacterium*
- *Pseudomonas savastanoi*
- *Ralstonia*
- *Rhodococcus fascians*
- *Xanthomonas*
- *Xylella*

Identify the presence of known virulence genes in raw whole genome sequencing data: Gene coverage and read-mapping depth as well as the most closely related organism allele for each gene are reported.

Example output of the Vir-Search tool

Virulence Gene Search results

Submitted job name: test16a
Organism: Agrobacterium
Forward/Single reads file: A1_3258_trimmed_1.fasta
Reverse reads file:
Minimum % coverage of database gene: 90%
Maximum % divergence from database gene: 10%
Output file prefix: 141184284_6208

Result table

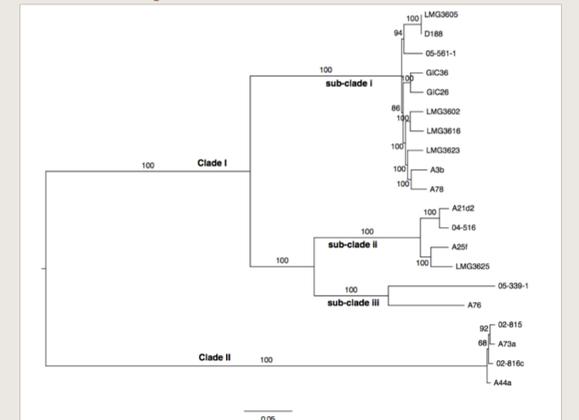
Agrobacterium tumefaciens CS8 genes used as reference for all virulence genes except for GALLS which came from Agrobacterium tumefaciens strain C589

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	-	-	-

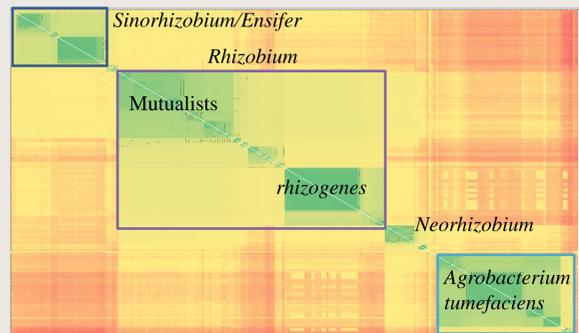
Found	Gene	Coverage (%)	Depth	Closest Allele
+	vrs0	100.0	28.21	Rhizobium_rubi_NBRC_13261
+	vrs4	100.0	30.795	Rhizobium_rubi_NBRC_13261
+	vrs5	100.0	23.916	Agrobacterium_arsenivivax_strain_KF9_330
+	vrs6	100.0	31.887	Agrobacterium_arsenivivax_strain_KF9_330
+	vrs7	100.0	28.272	Agrobacterium_arsenivivax_strain_KF9_330
+	vrs8	100.0	35.159	Rhizobium_rubi_NBRC_13261
+	vrs9	100.0	27.213	Rhizobium_rubi_NBRC_13261
+	vrs10	100.0	27.482	Rhizobium_rubi_NBRC_13261
+	hml1	100.0	25.191	Rhizobium_rubi_NBRC_13261
+	hml2	100.0	27.164	Rhizobium_rubi_NBRC_13261
+	hml3	100.0	24.153	Rhizobium_rubi_NBRC_13261
+	hml4	100.0	25.278	Rhizobium_rubi_NBRC_13261
+	hml5	100.0	27.164	Rhizobium_rubi_NBRC_13261
+	ocm	100.0	63.942	Rhizobium_ssp._Y1900
+	ocp	100.0	65.948	Rhizobium_spp._strain_Y1900
+	ocg	100.0	58.199	Agrobacterium_nitroreducens_K94

Simplified whole genome analysis: Use whole genome sequencing reads and reference genome sequences to generate a core genome phylogeny, or calculate pairwise average nucleotide identity (ANI) between genome assemblies, with minimal user input.

Phylogenetic tree of *R. fascians* genomes generated using the WGS Pipeline



Average nucleotide identity (ANI) heatmap of *A. tumefaciens* genomes generated using Auto ANI



Acknowledgements:

This work was supported by the National Institute of Food and Agriculture, U.S. Department of Agriculture award 2014-51181-22384 (JHC and NJG). Partial support also was provided by the USDA Agricultural Research Service Grant 5358-22000-039-00D (NJG), USDA National Institute of Food and Agriculture Grant 2011-68004-30154 (NJG), and the USDA ARS Floriculture Nursery Research Initiative (NJG). EWD is supported by a Provost's Distinguished Graduate Fellowship awarded by Oregon State University. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1314109 to EDW. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U. S. Department of Agriculture or National Science Foundation.



United States Department of Agriculture

National Institute of Food and Agriculture

