Supplemental Materials and Methods

Southern blot hybridization, PCR, and pulsed-field gels

Genomic DNAs (gDNAs) were purified from 0.5 ml of overnight cultures of selected strains using the Qiagen DNeasy kit (cat# 69506) per manufacturer's instructions. Approximately 0.5 μg of gDNA was digested with NEB restriction enzymes and resolved on a 0.8% agarose gel for 2 hr at 80V. Genomic DNA fragments were transferred from the gel via capillary action to an S&S Nytran SPC membrane (0.45 μm Nylon) using alkaline transfer (1). The membrane was baked for 2hr at 80C under vacuum. Hybridization was performed at 45C overnight using Roche's Dig Easy Hybridization solution. Washing, blocking and detection were accomplished according to manufacturer's directions using the Roche Wash and Block Buffer Set along with the DIG DNA Labeling and Detection Kit. Probe for *telA* was produced via PCR of *A. tumefaciens* C58 genomic DNA using the primers 5'CTAGCCATCTGCAACATGAAGA3' and 5'AGCGACGTTTCGAGGTCGTT3'. Probe for *atu2522* was produced via PCR of *A. tumefaciens* C58 genomic DNA using the primers 5'ATGATGAAACGCAATCTGATAGG3' and 5'GTCGGAATGCAGTTTTACCA3'.

PCR amplification of putative gene transfer junctions and novel genes on chromosome II (2) was conducted using the primers 5'-AGATGTCGGATTACCTCGTCAAG-3' and 5'-AGGCAGCAAACATCACTATCTCC-3' to amplify the region adjacent to Atu3706; primers 5'-GTCGAAGAGATTAAGCCTGTTTCC-3' and 5'-GGTTACGAGACATGGTGGACATAC-3'to amplify a region within Atu3765; and primers 5'-CTTTACGGTCTCCTCTATGATCAGC-3' and 5'-CACTTGGCCTGTCTTTGATGGTTG-3' to amplify a region within Atu3669.

Pulsed-field gel electrophoresis was conducted as previously described (3).

Genome comparisons and analysis of sequence differences

All primary sequence data generated by Goodner *et al.* (4) were transferred from Monsanto Company to the University of Washington genome center, along with the original DNA sample used for construction of the genomic library. Comparison of the sequences identified 37 differences. Each variant locus was amplified by PCR using DNA from the two sequenced strains and seven additional C58 culture lines (Table S4). The amplified DNA was sequenced using BigDye termination reactions (Applied Biosystems, Inc.) according to manufacturer's directions.

Protein-coding gene prediction

Gene predictions in the revised C58 annotation are the result of a careful evaluation of the annotation of the two original projects (4, 5) complemented by automated annotation provided by the Comprehensive Microbial Resource (CMR) (6) and EasyGene predictions (7). Orthologous groupings were computed among three sequenced *Agrobacterium* strains and other sequenced members of the Rhizobiaceae (2, 4, 5, 8-13); synteny among these genomes provided additional evidence for the presence or absence of genes. The tool OrthoMCL (14) was used for ortholog computation. Gene predictions meeting the following criteria were evaluated by a team member and deleted or modified as necessary: Genes having low similarity (BLASTP e-value higher than 1e-15; (15)) to known genes, genes that overlapped other predicted genes by more than 30 base pairs, and genes with predicted start sites more than 10 codons different from the closest predicted ortholog or paralog. Finally, all regions greater than 500 bp that lacked genes were scanned using BLASTX (15) against GenBank's non-redundant sequence repository and against the genomes of *A. vitis* S4 and *A. radiobacter* K84. As a result of this process, a number of new

gene predictions have been added and some original predictions were deleted. Translation starts for the accepted gene predictions were also examined and revised when appropriate, taking into account a multiple alignment of close orthologs and the presence of potential ribosomal binding sites.

The gene identifiers (also known as locus tags) for the revised C58 annotation that refer to genes that have been kept from the original annotations are the same as those defined by Wood *et al.* (5) (format AtuXXXX), which have been widely adopted in the literature. Newly-predicted genes were given the locus tag pattern Atu8XXX, as were a number of genes that were initially predicted only by Goodner *et al.* (4) or through subsequent reanalysis by Chen et al. (23).

Functional assignment for protein-coding genes

Since the original annotations from 2001, many genes have been experimentally characterized, both in *Agrobacterium* and in related species. The annotation was updated to reflect these new analyses. Many product descriptions have also been improved using TIGRFams, a database of well characterized protein families (16).

RNA gene annotation

In the original annotation, ribosomal RNAs, tRNAs, a tmRNA, and an RNAse had been identified. Using data available from the RFAM database (17), we have located 24 additional RNAs with special features, including 14 mRNAs with riboswitches and 10 small, non-coding RNAs. We have also added to the annotation of the linear chromosome a tRNA interrupted by a self-splicing group I intron, which had been noted by Reinhold-Hurek and Shub in the early 1990's (18) but had not been included in either of the original annotations. The locus tags of the two tRNA halves are Atu8111.1 and Atu8111.2. The intron has been added to the database as a "genome feature".

Putative small RNA (sRNA) genes were derived from the work of Wilms *et al*. (19) and given numbers in the *atu9xxx* format. We added only those genes for which transcripts were detected by RNA sequencing experiments (Table S4 in the Wilms *et al*. manuscript).

Pseudogene annotation

The original UW annotation reported 11 pseudogenes (5). The corrected C58 sequence showed that two of these putative pseudogenes are actually predicted functional genes; Atu1168, a bifunctional riboflavin deaminase-reductase, and Atu3304, a cellulose synthesis gene. Other pseudogenes or gene fragments have been added or correctly labeled, for a current total of 28 pseudogenes or gene fragments, many of which are transposon fragments. Fifteen of the 28 pseudogenes or gene fragments are located on the plasmid pAtC58. Some of the newly identified pseudogenes were found by the tool GenVar (20).

Abbreviations

C58, *Agrobacterium tumefaciens* C58; S4, *Agrobacterium vitis* S4; K84, *Agrobacterium radiobacter* K84; ATCC, American Type Culture Collection; PCR, Polymerase Chain Reaction; Indel, insertion or deletion mutation; NCBI, National Center for Biotechnology Information; bp, base-pairs; Kbp, kilobase-pairs; Mbp, megabase-pairs.

Supplemental Figure Legend

Figure S1.

Southern blot hybridization for detection of *telA*. DNA samples from the strains listed in Table S5 were digested with restriction endonucleases and hybridized at low stringency. Only Biovar 1 strains gave a detectable signal, and the restriction endonuclease digestion patterns fell into several recognizable patterns, suggesting that the strains within each pattern group are closely related. Representative blots for each group are shown. The left lane in each panel contains DNA molecular weight marker set III (Roche Applied Bioscience) labeled with digoxigenin. The size of each marker band, in base pairs, is shown to the left of Panel A. Each panel is from a different gel, so each requires reference to the panel's specific marker lane. Note that the patterns shown in panels A and F may be RFLP variants, as may panels B and G. Strains falling into a particular pattern group include: Panel A, *A. tumefaciens* strains C58 and MUM-1; Panel B, *A. tumefaciens* strains A6, Ach5, ANT4, ATCC 15955, IIBV7, Bo542, and NCPPB 396; Panel C, *A. tumefaciens* strains B8-806 and KU12; Panel D., *A. tumefaciens* FACH; Panel E, *A. rhizogenes* DC-AR2; Panel F, *A. tumefaciens* T37; Panel G, *A. tumefaciens* R10; Panel H, *A. tumefaciens* Chry5; Panel I, *A. rhizogenes* K599; Panel J, *A. tumefaciens* FA-1. Samples of DNA in panels A – I were digested (left to right) using BclI, EcoRI, HindIII, and SacII; Panel J uses identical endonucleases, except that the SacII digestion is missing.

Supplemental Tables

Table S1. Positions and types of sequence discrepancies between the *A. tumefaciens* C58 genome sequences produced by Goodner *et al.* (4) and Wood *et al.* (5). The locus discussed in the text containing 16 sequence discrepancies downstream of the 16s rRNA gene at 58.3 kbp of Chromosome I is not included in the table, but is provided as a footnote[a].

Replicon[b]	Position C58UW[c]	Position ATCC[c]	Position this study[c]	Nucleotide C58UW[d]	Nucleotide ATCC[d]	Consensus Nucleotide[e]	Difference Type[f]	ATCC33970 sequence[g]	Mutant Strain[h]	Mutant ORF[i]	Coding Change[j]
Chr. I	80930	80729	80930	G	A	G	P	tcagggAtgacat	ATCC33970	Atu0079	P→S
Chr. I	156690	156490	156690	N	C	N	F	ccccccCgcctgt	ATCC33970	Atu0148	F[k]
Chr. I	258152	257952	258152	A	C	C	P	cacgttCttttcc	C58UW	Atu0260	K→N
Chr. I	326648	326448	326648	A	T	T	P	atggacTgcatca	C58UW	Atu0332	R→P
Chr. I	326687	326487	326687	G	C	C	P	ttttgaCgcagca	C58UW	Atu0332	Q→L
Chr. I	503054	502855	503055	N	C	C	F	ggcggcCgcacca	C58UW	Atu8131	F[l]
Chr. I	763961	763761	763962	C	N	C	E	gcgcctCaacaat	N.A. (E)	N.A.	NONE

Chr. I	822392	822191	822393	C	N	C	E	tcccccCttgagg	N.A. (E)	N.A.	NONE
Chr. I	893288	893068-893178	893270-893380	Seq. Absent	Seq. Present	Seq. Present	I	See Note[m]	C58UW	OUT	NONE
Chr. I	936905	936795	936997	G	N	N	E	cccaat*gggcga	N.A. (E)	N.A.	NONE
Chr. I	1162733	1162622	1162824	C	N	N	E	aacccc*aacctc	N.A. (E)	N.A.	NONE
Chr. I	1169341	1169229	1169431	C	N	N	E	aggccc*ggggag	N.A. (E)	N.A.	NONE
Chr. I	1220123	1220010	1220212	G	N	N	F	gggggg*cttttg	C58UW	OUT	NONE
Chr. I	1310128	1310015	1310217	G	A	A	P	accatcAccgtca	C58UW	Atu1319	T→A
Chr. I	1446784	1446670	1446873	A	N	A	E	tcgaaaAgcagtg	N.A. (E)	N.A.	NONE
Chr. I	1617958	1617844	1618047	G	C	C	P	cgcggcCgatgaa	C58UW	Atu1632	SILENT
Chr. I	1624250	1624137	1624339	N	A	N	E	taaaat*aaataa	N.A. (E)	N.A.	NONE
Chr. I	1661908	1661794	1661997	A	N	A	E	tttgaaAcagctt	N.A. (E)	N.A.	NONE
Chr. I	1906754	1906641	1906844	C	N	C	E	ccttggCcgaacg	N.A. (E)	N.A.	NONE
Chr. I	2090637	2090524	2090726	A	T	A	E	ttttccAgatccc	N.A. (E)	N.A.	NONE
Chr. I	2236399	2236285	2236640	C	N	N	E	cccccc*tctgcc	N.A. (E)	N.A.	NONE

Chr. I	2359302	2359188	2359390	T	C	T	E	gagatgTccggca	N.A. (E)	N.A.	NONE
Chr. I	2380180	2380066	2380268	T	C	C	E	tctcccCccttga	N.A. (E)	N.A.	NONE
Chr. I	2401441	2401328	2401426	N	C	N	E	gatgcc*cggcag	N.A. (E)	N.A.	NONE
Chr. I	2604773	2604661	2604862	N	G	N	E	cgacga*gcttac	N.A. (E)	N.A.	NONE
Chr. I	2626194	2626083	2626283	N	G	N	E	gctggc*gcgcgg	N.A. (E)	N.A.	NONE
Chr. I	2670041	2669930	2670130	G	T	G	E	ggtgacGccttcc	N.A. (E)	N.A.	NONE
Chr. I	2761079	2760968	2761168	G	A	A	P	tacgagAcggaag	C58UW	Atu2759	T→A
Chr. I	2825552	2825441	2825641	A	G	G	P	acgcccGaaatct	C58UW	Atu2822	E→K
Chr. II	173	2074583	181	T	G	T	E	aatcagTtttgtc	N.A. (E)	N.A.	NONE
Chr. II	173787	1900969	173795	G	A	A	E	tgcgggAccgatg	N.A. (E)	N.A.	NONE
Chr. II	1507966	566791	1507974	G	N	N	F	tttgcc*ggcatg	C58UW	Atu4370	F[n]
Chr. II	1628558	446200	1628566	T	N	N	E	ccaacg*ttgaaa	N.A. (E)	N.A.	NONE
Chr. II	1760142	314615	1760150	N	T	T	E	ttgcgcTtttttt	N.A. (E)	N.A.	NONE
pATC58	128201	128503-128592	128196-128285	Seq. Absent	Seq. Present	Seq. Present	I	See Note[o]	C58UW	Atu5121	I[p]

pATC58	400626	401022	400715	G	N	N	E	acgggg*ccggcg	N.A. (E)	N.A.	NONE
pATC58	464502	464899	464591	N	A	N	E	caatga*atcagg	N.A. (E)	N.A.	NONE

[a] The sequence around the differing region of the rRNA loci are shown below. Variants are shown in bold:

C58UW - GTAAGACG**A**TC**GCAC**C**G**A-TCTTCG**GA**T**CAA**C**GC**G**GT**ATGAA

ATCC33970 -GTAAGACG**C**TC**CGGA**C**AA**G**TCTTCG**AC**T**TGC**C**TT**G**AG**ATGAA

[b]The replicon on which the designated sequence discrepancy resides. Chr. I, Chromosome I; Chr. II, Chromosome II. The two genome projects had no discrepancies in the sequence of pTiC58.

[c]Position of the sequence discrepancy in the original C58UW sequence files, the original ATCC33970 sequence files or the current ATCC33970 sequence files respectively. Renumbering of the current ATCC33970 sequence files followed the addition of telomere sequences and corrections noted in the text.

[d]Nucleotide at the sequence discrepancy in the designated sequence. N denotes that no nucleotide was called at the analogous position in the designated sequence.

[e]Consensus nucleotide at the variant locus based on comparison of all C58 strains shown in table S4. These data are used to determine which of the two sequenced strains has acquired a mutation at that locus. The correct sequence is noted in the case of a base calling error by one project team or the other.

[f]The types of sequence discrepancies identified. E, Base-calling error by one project team or the other; I, Large Indel; F, Single-base frameshift; P, single-base point mutation.

[g]The sequence of ATCC33970 surrounding the point of the sequence discrepancy. Capital letters indicate the location of the consensus base, according to sequencing of reference strains (see text). An asterisk (*) indicates the approximate location of the erroneous base call; note that it is not possible to identify which base in a string of identical bases was called incorrectly.

[h]To determine which of the two strains harbors a mutation at a given locus, both sequences were compared to the consensus sequence generated from the strains listed in Table S4. Note that at each variant position either C58UW or ATCC33970 matched all other C58 strains tested. N.A. indicates that neither strain is mutated at that position since the difference in the sequence derives from an initial base-calling error.

[i]The ORF designation indicates the locus ID of the open reading frame in which this variation is located. N.A. indicates that neither strain is a variant. OUT indicates that the variation is outside a predicted open reading frame, although some may be located in regulatory regions.

[j]Denotes the specific amino acid change seen in the mutant with respect to the consensus matched sequence. Note in all but the first two rows the mutations are found in C58UW. Single amino acid codes are shown separated by an arrow (→) indicating the amino acid found in the wild type strain and mutant respectively as defined in the mutant strain column. NONE indicates that the mutation was outside of an open reading frame. SILENT indicates no change in the encoded amino acid. F indicates a frameshift.

[k]The frameshift mutation in C58UW alters amino acids starting at position 198 to read as follows: HDGKACLGCGKRQAASDRARFTLCRR.

[l]Atu8131 is the wild-type version of a mutant gene designated Atu0512 by Wood et al. (5). The two genes differ in their N-terminal region, upstream of the frameshift mutation.

[m]The sequence of nucleotides deleted at this location in strain C58UW is:

5’ccccctcaaggggggagatcgatctgcggcaaggtttcgcccatctcaacctttgaggatgaagcgacgaaagggcctcctgccgatctccccccttgaggggggagatgcc3’.

[n]The frameshift mutation in C58UW alters amino acids starting at position 390 to read as follows: 389-DMPIGTLSGGNQQKIFISRWLATSPKLLLLDDPTKGIDLGAKADLFALMRQQADAGATILLYSSEDAEILEYADRILVFNGGRISAELTGADMTSVNMTRAAYGDAA.

[o]The sequence of nucleotides deleted at this location in strain C58UW is:

5'agcagatcgtccattgcaaggcagatcagtaccgtgcggcggtgtggctgactattcgcttcatccggtaaaaacatgtcgaggctcggc 3'.

[p]The deletion in C58UW removed the following amino acids from Atu5121 of ATCC33970 starting at position 34: PSLDMFLPDEANSQPHRRTVLICLAMDDLL.

Table S2. Distribution of AgroCir1, AgroCir2, and KE3 repetitive elements in sequenced *Agrobacterium* strains and close relatives.

	AgroCIR1				AgroCIR2				KE3			
	Full Elements		Half Elements		Full Elements		Half Elements		Full Elements		Half Elements	
Strain	Intergenic	Intragenic	Intergenic	Intragenic	Intergenic	Intragenic	Intergenic	Intragenic	Intergenic	Intragenic	Intergenic	Intragenic
A. tumefaciens C58												
ChrI	22	1	12	0	23	4	0	1	15[a]	3	18	1
ChrII	10	0	3	0	8	0	0	0	5[a]	0	9	0
A. tumefaciens H13-3												
ChrI	6	0	4	0	9	1	3	1	3[a]	0	15	2
ChrII	4	0	19	2	2	0	2	0	1[a]	0	8	0
A. vitis S4												
ChrI	0	0	0	0	1	0	7	1	0	0	4	0
ChrII	0	0	0	0	0	0	3	0	0	0	3	0
A. radiobacter K84												
Chr	0	0	0	0	0	0	0	1	2[b]	0	2	0
2.65 Mbp	0	0	0	0	0	0	0	0	0	0	0	0
Rhizobium etli												

	Chr	11	1	13	1	8	2	1	2	6[c]	0	9	0
	P42c	0	0	0	0	0	0	0	0	0	0	1	0
	P42e	0	0	1	0	0	0	1	0	0	0	0	0
	P42f	0	0	0	0	0	0	0	0	1[a]	0	0	0
Rhizobium leguminosarum													
	Chr	3	1	0	0	2	2	1	0	7[a]	1	6	0
	PRL9	0	0	0	0	0	0	0	0	1[a]	0	0	0
	pRL10	0	0	1	0	0	0	0	0	0	0	0	0
	PRL12	0	0	0	0	0	0	0	0	1[a]	0	0	0
Sinorhizobium meliloti													
	Chr	15	0	47	1	1	0	5	2	4[b]	0	0	0
	pSymB	1	1	15	2	1	0	3	1	0	0	1	0
Mesorhizobium loti													
	Chr	0	0	0	0	6	2	7	2	0	0	6	0

[a]The full KE3 elements have internal spacers of <50 bp.

[b]The full KE3 elements have internal spacers of >100 bp.

[c]Five of the KE3 elements have internal spacers of <50 bp, and one has an internal spacer of >100 bp.

Table S3. Repeat elements that overlap open reading frames in the *A. tumefaciens* C58 Genome. None of these intragenic repetitive elements are found in orthologous genes in the *A. tumefaciens* H13-3 genome.

B.

Repeat Type	Gene ID	Encoded Protein
AgroCir1	Atu0976	CycL cytochrome C-type biogenesis transmembrane protein
AgroCir2	Atu1168	RibD bifunctional riboflavin deaminase/reductase
AgroCir2	Atu1660	hypothetical protein
AgroCir2	Atu2001	UvrB excinuclease ABC subunit B
AgroCir2	Atu2192	hypothetical protein
AgroCir2	Atu8020	hypothetical protein
KE3	Atu0805	hypothetical protein
KE3	Atu1026	hypothetical protein
KE3	Atu1427	hypothetical protein
KE3	Atu1896	hypothetical protein

Table S4. Strains used as sequence references when comparing the two sequenced versions of *A. tumefaciens* C58 (4, 5). All strains except the "original isolate" probably derive from the same C58 culture sent from E. Nester to J. Schell in 1971.

Strain Designation	Notes on Strain
C58 (original isolate)	This is from the original Cornell University strain collection. Originally collected in 1958 by Kenneth Parker from a Mazzard rootstock of a one-year old tart cherry tree (*Prunus cerasus*) near Sodus, NY (21, 22). The "C" in the strain name designates Cherry, the "58" designates the year collected). The culture was passed an unspecified

	number of times for storage before being lyophilized in 1979. The Cornell stock number is 1512. The strain was discussed and distributed by Steven Beer to attendees of the Crown Gall Conference at Cornell University in 2003.
C58C1	A derivative of C58 cured of pTiC58 by the Jeff Schell group (23). Obtained by S. Farrand from Jacques Tempé. Probably identical to C58C9 (24) but may have been misnamed in early literature.
C58C1RS	Derivative of C58C1 resistant to rifampicin and streptomycin (24). Obtained by S. Farrand from Allen Kerr.

GV3101arg13	An arginine auxotroph of C58 strain GV3101. Received by S. Farrand from G.R.K. Sastry. Originated in the laboratory of Marc Van Montague.	178
GMI9017	A derivative of GMI9027 that is cured of pAtC58 (25). GMI9027 is a C58 derivative resistant to rifampicin, spectinomycin, and possibly streptomycin.	
UIA5	A derivative of GMI9017 that is cured of pTiC58. Produced in the laboratory of S. Farrand.	
K439	A derivative of C58C1ChlEry harboring pWI1000, a conjugation-constitutive mutant of pTiC58 (26). Obtained by S. Farrand from Allen Kerr.	

Table S5. Survey of *Agrobacterium* strains for *telA*, *acvB*, and the presence of a linear chromosome II. A (+) indicates that the gene (or linear chromosome) was detected, a (–) indicates that it was not detected, and a NT indicates that it was not tested for a particular strain.

Biovar	*Agrobacterium* Strain	*tel*A	*Atu2522*	Linear ChrII	Source	Host / Origin / Isolated By (if known)
1	*tumefaciens* C58UW	+	+	+	D. Wood	Cherry gall / New York / Kenneth Parker
1	*tumefaciens* C58 ATCC33970	+	+	+	American Type Culture Collection	Cherry gall / New York / Kenneth Parker
1	*tumefaciens* A6	+	+	+	D. Wood	Black raspberry gall, Iowa
1	*tumefaciens* Ach5	+	NT	NT	D. Wood	Robert Dickey, Cornell University
1	*tumefaciens* ATCC15955	+	+	+	S. Farrand<---E. Nester	Tomato gall California / J.E. DeVay
1	*tumefaciens* IIBV7	+	+	NT	S. Farrand<---M. Tate	Chrysanthemum gall / Germany
1	*tumefaciens* ANT4	+	+	NT	S. Farrand<---Y. Dessaux	Crown gall on Ficus / France
1	*tumefaciens* B6-806	+	+	+	S. Farrand<---E. Nester	Tomato gall / Iowa
1	*tumefaciens* Bo542	+	+	NT	S. Farrand<---E. Nester	Dahlia gall / Germany

1	*tumefaciens* Chry5	+	+	+	S. Farrand<---S. Pueppke	Chrysanthemum gall / Florida
1	*tumefaciens* FA-1	+	+	+	S. Farrand<---P. Dion	Quebec / grape
1	*tumefaciens* FACH	+	+	+	S. Farrand<---P. Dion	Quebec / grape
1	*tumefaciens* KU12	+	+	NT	D. Wood	
1	*tumefaciens* MUM1	+	+	NT	S. Farrand<---M. Schroth	Chrysanthemum gall / California
1	*tumefaciens* NCPPB 396	+	+	NT	S. Farrand	
1	*tumefaciens* R10	+	+	+	S. Farrand<---J. Tempé	Vigne gall / Romania
1	*tumefaciens* T37	+	+	+	S. Farrand<---P. Dion	Walnut gall / California
1	*rhizogenes* K599	+	+	+	S. Farrand<---A. Kerr	Hairy root on cucumber / UK
1	*rhizogenes* DC-AR2	+	+	NT	S. Farrand<---N. Tanaka	Hairy root on melon / Japan
Intermediate[a]	*tumefaciens* NA164	-	NT	NT	S. Farrand<---P. Dion	Quebec
2	*tumefaciens* 28	-	-	-	S. Farrand<---P. Dion	Quebec
2	*tumefaciens* 180	-	-	-	S. Farrand<---P. Dion	Quebec
2	*tumefaciens* AB2/73	-	-	-	S. Farrand<---L. Moore	Lippia gall / Arizona
2	*tumefaciens* EU2	-	-	-	S. Farrand	Euonymous gall / Illinois

2	*tumefaciens* EU3-2 NK	-	-	NT	S. Farrand	Euonymous gall / Illinois
2	*tumefaciens* J84/95	-	-	NT	S. Farrand<---L . Moore	Apple gall / Californmia
2	*tumefaciens* J73	-	-	-	S. Farrand<---J.Thomson	Prunus gall / South Africa
2	*tumefaciens* K108	-	-	NT	S. Farrand<---A. Kerr	Almond gall / South Australia
2	*tumefaciens* NA147	-	-	-	S. Farrand<---P. Dion	Quebec
2	*tumefaciens* NA167	-	-	NT	S. Farrand<---P. Dion	Quebec
2	*tumefaciens* NA567	-	-	-	S. Farrand<---P. Dion	Quebec
2	*tumefaciens* Plum 3-5	-	-	NT	S. Farrand<---M. Schroth	Crown gall on Prunus / California
2	*tumefaciens* AG44	-	NT	NT	E. Nester	Frank White
2	*rhizogenes* 8196	-	-	-	S. Farrand<---J. Lippincott	Hairy root on apple / Riker
2	*rhizogenes* A4	-	-	-	S. Farrand<---J. Tempé	Hairy root on rose / California
2	*rhizogenes* TR7	-	NT	-	S. Farrand<---J. Lippincott	Hairy root on apple / Wisconsin
2	*radiobacter* K84	-	-	-	E. Nester	Soil around peach gall / South Australia
2	*radiobacter* K112	-	-	NT	S. Farrand<---A. Kerr	Soil around galled almond tree / South Australia

2	*radiobacter* F64/95	-	-	-	S. Farrand<---L . Moore	Apple gall / Washington
3	*vitis* S4	-	NT	-	T. Burr <--- E. Szegedi	Hungary / E. Szegedi
3	*vitis* F2/5	-	NT	-	T. Burr <--- J. Staphorst	South Africa / J. Staphorst
3	*vitis* CG49	-	-	NT	T. Burr	Grape / New York
3	*vitis* CG56	-	-	NT	T. Burr	Michigan
3	*vitis* CG60	-	-	NT	T. Burr	New York
3	*vitis* CG78	-	-	NT	T. Burr	New York
3	*vitis* CG81	-	-	NT	T. Burr	Michigan
3	*vitis* CG88	-	-	NT	T. Burr	New York
3	*vitis* CG98	-	-	NT	T. Burr	Virginia
3	*vitis* CG101	-	-	NT	T. Burr	Virginia
3	*vitis* CG102	-	-	NT	T. Burr	Virginia
3	*vitis* CG106	-	-	NT	T. Burr	Mississippi
3	*vitis* CG108	-	-	NT	T. Burr	New Mexico
3	*vitis* CG213	-	-	NT	T. Burr	Grape sap / New York

3	*vitis* CG228	-	-	NT	T. Burr	New York
3	*vitis* CG407	-	-	NT	T. Burr	Roots / New York
3	*vitis* CG412	-	-	NT	T. Burr	Soil / New York
3	*vitis* CG415	-	-	NT	T. Burr	Grape sap / New York
3	*vitis* CG424	-	-	NT	T. Burr	Soil / New York
3	*vitis* CG437	-	-	NT	T. Burr	Soil / Virginia
3	*vitis* CG448	-	-	NT	T. Burr	Cutting callus / New York
3	*vitis* CG475	-	-	NT	T. Burr	Grape root / New Mexico
3	*vitis* CG608	-	-	NT	T. Burr	Cutting callus / New York
3	*vitis* CG624	-	-	NT	T. Burr	Cutting callus / New York
3	*vitis* CG667	-	-	NT	T. Burr	Sap / New York
3	*vitis* CG714	-	-	NT	T. Burr	Cutting callus / New York
3	*vitis* CG964 (NW161)	-	-	NT	T. Burr <--- E. Bien	Germany
3	*vitis* CG968 (IIPV-BO-2147)	-	-	NT	T. Burr <--- C. Bazzi	Cutting callus / Italy
3	*vitis* CG988 (NW165)	-	-	NT	T. Burr <--- E. Bien	Germany

3	*vitis* CG1005 (K306)	-	-	NT	T. Burr <--- A. Kerr	Australia
3	*vitis* CG1055	-	-	NT	T. Burr	Grape gall / Mississippi
3	*vitis* CG1082 (TM4)	-	-	NT	T. Burr <--- E. Szegedi	Hungary
3	*vitis* CG1083 (Sz2)	-	-	NT	T. Burr <--- E. Szegedi	Hungary
3	*vitis* CG1106 (NW11)	-	-	NT	T. Burr <--- L. Otten	Germany
3	*vitis* CG1107 (NW113)	-	-	NT	T. Burr <--- L. Otten	Germany
3	*vitis* CG1115 (A,yolu2.1)	-	-	NT	T. Burr <--- S. Maden	Turkey
3	*vitis* CG1116 (Fid2.3)	-	-	NT	T. Burr <--- S. Maden	Turkey
3	*vitis* CG1126	-	-	NT	T. Burr <--- E. Szegedi	Hungary
3	*vitis* At2	-	NT	NT	E. Nester	
3	*vitis* At6	-	NT	NT	E. Nester	
3	*vitis* CG47	-	NT	NT	E. Nester <---T. Burr	Grape gall / New York
3	*vitis* AB4	-	NT	NT	E. Nester <--- E. Szegedi	
3	*vitis* SZ1	-	NT	NT	E. Nester <--- E. Szegedi	
3	*vitis* AG162	-	NT	NT	E. Nester	Grapevine tumor / USSR

3	*vitis* K305	-	NT	NT	S. Farrand <--- A. Kerr	Australia / A. Kerr
3	*vitis* K308	-	NT	NT	S. Farrand <--- A. Kerr	Grape gall / South Australia
3	*vitis* K309	-	NT	NT	S. Farrand <--- A. Kerr	Grape gall / South Australia
3	*vitis* H6	-	NT	NT	S. Farrand <--- J. Thomson	Agrocin-producing nonpathogen / South Africa / R. Goodman
3	*vitis* S2/3	-	NT	NT	S. Farrand <--- R. Goodman	Hungary
3	*vitis* Hm1	-	NT	NT	S. Farrand <--- L. Otten	Grape gall / France
3	*vitis* 2608	-	NT	NT	S. Farrand <--- L. Otten	Grape gall / France
3	*vitis* 2641	-	NT	NT	S. Farrand <--- L. Otten	Grape gall / France
3	*vitis* Ag57	-	NT	NT	S. Farrand <--- R. Goodman	Grape / Greece

[a] This strain has an intermediate phenotype. It has the Biovar 1 characteristic of 3-ketolactose production, and the Biovar 2 characteristics of growth on erythritol as a sole carbon source and sensitivity to 2% NaCl.

Table S6. Biovar I-specific protein-coding genes. These genes were identified using the Phylogenetic Profiler function within the IMG database (img.jgi.doe.gov). Biovar I-specific genes are those conserved at >70% amino acid sequence identity within Biovar I genomes (C58, H13-3, & ATCC31749 [draft]) but not found within any other non-Biovar I member of the Rhizobiaceae at >30% amino acid sequence identity.

IMG Gene_oid Number	Locus Tag	Product Name
639295636	Atu0146	sulfite reductase (NADPH) alpha subunit (EC 1.8.1.2) (IMGterm)
639295657	Atu0167	transcriptional regulator, AraC family (IMGterm)
639295689	Atu0200	hypothetical protein
639295725	Atu0236	hypothetical protein
639295732	Atu0243	hypothetical protein
639295746	Atu0257	hypothetical protein
639295756	Atu0267	hypothetical protein
639295757	Atu0268	hypothetical protein
639295806	Atu0317	hypothetical protein
639295985	Atu0496	hypothetical protein

639295993	Atu0504	hypothetical protein
639296020	Atu0531	hypothetical protein
639296077	Atu0588	hypothetical protein
639296086	Atu0597	hypothetical protein
639296149	Atu0660	mannosylfructose-phosphate phosphatase (IMGterm)
639296375	Atu0886	hypothetical protein
639296381	Atu0892	hypothetical protein
639296413	Atu0924	hypothetical protein
639296432	Atu0943	hypothetical protein
639296472	Atu0983	hypothetical protein
639296497	Atu1008	hypothetical protein
639296507	Atu1018	hypothetical protein
639296535	Atu1046	ABC transporter, substrate binding protein
639296568	Atu1079	hypothetical protein
639296627	Atu1138	protease VII precursor

639296636	Atu1147	hypothetical protein
639296639	Atu1150	hypothetical protein
639296644	Atu1155	hypothetical protein
639296691	Atu1203	hypothetical protein
639296701	Atu1213	hypothetical protein
639296831	Atu1343	hypothetical protein
639296927	Atu1439	hypothetical protein
639296991	Atu1503	hypothetical protein
639297041	Atu1553	hypothetical protein
639297056	Atu1568	transcriptional regulator, TetR family (IMGterm)
639297150	Atu1662	hypothetical protein
639297204	Atu1716	hypothetical protein
639297215	Atu1727	hypothetical protein
639297225	Atu1737	hypothetical protein
639297259	Atu1771	hypothetical protein

639297265	Atu1777	hypothetical protein
639297269	Atu1781	hypothetical protein
639297293	Atu1805	hypothetical protein
639297351	Atu1863	permease
639297353	Atu1865	hypothetical protein
639297442	Atu1954	hypothetical protein
639297460	Atu1972	hypothetical protein
639297461	Atu1973	hypothetical protein
639297490	Atu2002	esterase/lipase
639297521	Atu2033	hypothetical protein
639297561	Atu2074	hypothetical protein
639297623	Atu2137	hypothetical protein
639297675	Atu2189	hypothetical protein
639297684	Atu2198	hypothetical protein
639297729	Atu2243	hypothetical protein

639297812	Atu2326	hypothetical protein
639297856	Atu2370	hypothetical protein
639297860	Atu2374	hypothetical protein
639297873	Atu2387	NTP pyrophosphohydrolase, MutT family
639297918	Atu2432	hypothetical protein
639297959	Atu2473	ABC transporter, substrate binding protein
639297961	Atu2475	ABC transporter, membrane spanning protein
639297982	Atu2496	hypothetical protein
639297988	Atu2502	heme oxygenase
639298009	Atu2523	hypothetical protein
639298012	Atu2526	hypothetical protein
639298019	Atu2533	hypothetical protein
639298193	Atu2707	rhizobiocin
639298283	Atu2797	cobalamin biosynthesis protein
639298292	Atu2806	hypothetical protein

639298345	Atu3024	carbohydrate ABC transporter substrate-binding protein, CUT1 family (TC 3.A.1.1.-) (IMGterm)
639298346	Atu3025	oligo alginate lyase
639298366	Atu3045	hypothetical protein
639298378	Atu3057	hypothetical protein
639298413	Atu3092	hypothetical protein
639298450	Atu3129	polygalacturonase-like protein
639298451	Atu3130	carbohydrate ABC transporter substrate-binding protein, CUT1 family (TC 3.A.1.1.-) (IMGterm)
639298459	Atu3138	hypothetical protein
639298467	Atu3146	proline dipeptidase
639298478	Atu3157	ABC transporter, membrane spanning protein [C4-dicarboxylate]
639298596	Atu3275	invasion associated locus B
639298608	Atu3287	thiamine biosynthesis associated protein
639298654	Atu3336	hypothetical protein

639298683	Atu3365	hypothetical protein
639298702	Atu3384	hypothetical protein
639298787	Atu3469	phosphoglycerate mutase (EC 5.4.2.1) (IMGterm)
639298817	Atu3499	hypothetical protein
639298840	Atu3522	glucans biosynthesis protein
639298851	Atu3533	monosaccharide ABC transporter substrate-binding protein, CUT2 family (TC 3.A.1.2.-) (IMGterm)
639298866	Atu3548	glycosyltransferase
639298867	Atu3549	hypothetical protein
639298868	Atu3550	succinoglycan biosynthesis transport protein
639298871	Atu3553	hypothetical protein
639298873	Atu3555	hypothetical protein
639298876	Atu3558	glycosyltransferase
639298880	Atu3562	hypothetical protein
639298885	Atu3567	hypothetical protein

639298887	Atu3569	hypothetical protein
639298891	Atu3573	hypothetical protein
639298973	Atu3655	antioxidant enzyme
639299093	Atu3775	hypothetical protein
639299141	Atu3823	methionine synthase (B12-independent) (EC 2.1.1.14) (IMGterm)
639299142	Atu3824	hypothetical protein
639299180	Atu3862	hypothetical protein
639299232	Atu3914	cytoplasmic alpha-amylase
639299234	Atu3916	exogenous ferric siderophore receptor
639299313	Atu3995	transcriptional regulator, AraC family (IMGterm)
639299314	Atu3996	hypothetical protein
639299358	Atu4040	hypothetical protein
639299418	Atu4100	hypothetical protein
639299419	Atu4101	hypothetical protein
639299424	Atu4106	hypothetical protein

639299520	Atu4202	methyltransferase
639299664	Atu4346	hypothetical protein
639299696	Atu4378	hypothetical protein
639299898	Atu4580	hypothetical protein
639299906	Atu4588	sodium bile acid symporter family protein
639299907	Atu4589	probable transmembrane transport protein
639299932	Atu4614	glycosyltransferase
639299961	Atu4643	hypothetical protein
639299974	Atu4656	hypothetical protein
639299987	Atu4669	hypothetical protein
639300052	Atu4734	aconitase (EC 4.2.1.3) (IMGterm)
639300069	Atu4751	hypothetical protein
639300106	Atu4788	hypothetical protein
639300110	Atu4792	hypothetical protein
639300129	Atu4811	undecaprenyl-phosphate alpha N-acetylglucosaminyltransferase

639300130	Atu4812	hypothetical protein
639300131	Atu4813	hypothetical protein
639300133	Atu4815	glycosyltransferase
639300134	Atu4816	glycosyltransferase
639300135	Atu4817	glycosyltransferase
639300371	Atu5157	hypothetical protein
639300488	Atu5275	hypothetical protein
639300572	Atu5360	hypothetical protein
639300641	Atu5430	hypothetical protein
639300642	Atu5431	hypothetical protein
639300697	Atu5487	hypothetical protein

Supplemental References

1. **Sambrook J, Fritsch E, Maniatis T.** 1989. Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press, New York.
2. **Slater SC, Goldman BS, Goodner B, Setubal JC, Farrand SK, Nester EW, Burr TJ, Banta L, Dickerman AW, Paulsen I, Otten L, Suen G, Welch R, Almeida NF, Arnold F, Burton OT, Du Z, Ewing A, Godsy E, Heisel S, Houmiel KL, Jhaveri J, Lu J, Miller NM, Norton S, Chen Q, Phoolcharoen W, Ohlin V, Ondrusek D, Pride N, Stricklin SL, Sun J, Wheeler C, Wilson L, Zhu H, Wood DW.** 2009. Genome sequences of three *Agrobacterium* biovars help elucidate the evolution of multichromosome genomes in bacteria. J Bacteriol **191:**2501-2511.
3. **Goodner BW, Markelz BP, Flanagan MC, Crowell CB, Jr., Racette JL, Schilling BA, Halfon LM, Mellors JS, Grabowski G.** 1999. Combined genetic and physical map of the complex genome of *Agrobacterium tumefaciens*. J Bacteriol **181:**5160-5166.
4. **Goodner B, Hinkle G, Gattung S, Miller N, Blanchard M, Qurollo B, Goldman BS, Cao YW, Askenazi M, Halling C, Mullin L, Houmiel K, Gordon J, Vaudin M, Iartchouk O, Epp A, Liu F, Wollam C, Allinger M, Doughty D, Scott C, Lappas C, Markelz B, Flanagan C, Crowell C, Gurson J, Lomo C, Sear C, Strub G, Cielo C, Slater S.** 2001. Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58. Science **294:**2323-2328.
5. **Wood DW, Setubal JC, Kaul R, Monks DE, Kitajima JP, Okura VK, Zhou Y, Chen L, Wood GE, Almeida NF, Woo L, Chen YC, Paulsen IT, Eisen JA, Karp PD, Bovee D, Chapman P, Clendenning J, Deatherage G, Gillet W, Grant C, Kutyavin T, Levy R, Li MJ, McClelland E, Palmieri A, Raymond C, Rouse G, Saenphimmachak C, Wu ZN, Romero P, Gordon D, Zhang SP, Yoo HY, Tao YM, Biddle P, Jung M, Krespan W, Perry M, Gordon-Kamm B, Liao L, Kim S, Hendrick C, Zhao ZY, Dolan M, Chumley F, Tingey SV, Tomb JF, Gordon MP, Olson MV, Nester EW.** 2001. The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. Science **294:**2317-2323.
6. **Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O.** 2001. The Comprehensive Microbial Resource. Nucleic Acids Res **29:**123-125.
7. **Larsen TS, Krogh A.** 2003. EasyGene--a prokaryotic gene finder that ranks ORFs by statistical significance. BMC Bioinformatics **4:**21.
8. **Young JP, Crossman LC, Johnston AW, Thomson NR, Ghazoui ZF, Hull KH, Wexler M, Curson AR, Todd JD, Poole PS, Mauchline TH, East AK, Quail MA, Churcher C, Arrowsmith C, Cherevach I, Chillingworth T, Clarke K, Cronin A, Davis P, Fraser A, Hance Z, Hauser H, Jagels K, Moule S, Mungall K, Norbertczak H, Rabbinowitsch E, Sanders M, Simmonds M, Whitehead S, Parkhill J.** 2006. The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. Genome Biol **7:**R34.
9. **Galibert F, Finan TM, Long SR, Puhler A, Abola P, Ampe F, Barloy-Hubler F, Barnett MJ, Becker A, Boistard P, Bothe G, Boutry M, Bowser L, Buhrmester J, Cadieu E, Capela D, Chain P, Cowie A, Davis RW, Dreano S, Federspiel NA, Fisher RF, Gloux S, Godrie T, Goffeau A, Golding B, Gouzy J, Gurjal M, Hernandez-Lucas I, Hong A, Huizar L, Hyman RW, Jones T, Kahn D, Kahn ML, Kalman S, Keating DH, Kiss E, Komp C, Lelaure V, Masuy D, Palm C, Peck MC, Pohl TM, Portetelle D, Purnelle B, Ramsperger U, Surzycki R, Thebault P, Vandenbol M, Vorholter F-J, Weidner S, Wells DH, Wong K, Yeh K-C, Batut J.** 2001. The Composite Genome of the Legume Symbiont *Sinorhizobium meliloti*. Science **293:**668-672.

10. **Barnett MJ, Fisher RF, Jones T, Komp C, Abola AP, Barloy-Hubler F, Bowser L, Capela D, Galibert F, Gouzy J, Gurjal M, Hong A, Huizar L, Hyman RW, Kahn D, Kahn ML, Kalman S, Keating DH, Palm C, Peck MC, Surzycki R, Wells DH, Yeh KC, Davis RW, Federspiel NA, Long SR.** 2001. Nucleotide sequence and predicted functions of the entire *Sinorhizobium meliloti* pSymA megaplasmid. Proc Natl Acad Sci U S A **98:**9883-9888.
11. **Gonzalez V, Santamaria RI, Bustos P, Hernandez-Gonzalez I, Medrano-Soto A, Moreno-Hagelsieb G, Janga SC, Ramirez MA, Jimenez-Jacinto V, Collado-Vides J, Davila G.** 2006. The partitioned *Rhizobium etli* genome: Genetic and metabolic redundancy in seven interacting replicons. Proceedings of the National Academy of Sciences of the United States of America **103:**3834-3839.
12. **Kaneko T, Nakamura Y, Sato S, Asamizu E, Kato T, Sasamoto S, Watanabe A, Idesawa K, Ishikawa A, Kawashima K, Kimura T, Kishida Y, Kiyokawa C, Kohara M, Matsumoto M, Matsuno A, Mochizuki Y, Nakayama S, Nakazaki N, Shimpo S, Sugimoto M, Takeuchi C, Yamada M, Tabata S.** 2000. Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*. DNA Res **7:**331-338.
13. **Capela D, Barloy-Hubler F, Gouzy J, Bothe G, Ampe F, Batut J, Boistard P, Becker A, Boutry M, Cadieu E, Dreano S, Gloux S, Godrie T, Goffeau A, Kahn D, Kiss E, Lelaure V, Masuy D, Pohl T, Portetelle D, Puhler A, Purnelle B, Ramsperger U, Renard C, Thebault P, Vandenbol M, Weidner S, Galibert F.** 2001. Analysis of the chromosome sequence of the legume symbiont *Sinorhizobium meliloti* strain 1021. Proc Natl Acad Sci U S A **98:**9877-9882.
14. **Li L, Stoeckert CJ, Jr., Roos DS.** 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res **13:**2178-2189.
15. **Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:**3389-3402.
16. **Haft DH, Selengut JD, White O.** 2003. The TIGRFAMs database of protein families. Nucleic Acids Res **31:**371-373.
17. http://www.sanger.ac.uk/Software/Rfam.
18. **Reinhold-Hurek B, Shub DA.** 1992. Self-splicing introns in tRNA genes of widely divergent bacteria. Nature **357:**173-176.
19. **Wilms I, Overloper A, Nowrousian M, Sharma CM, Narberhaus F.** 2012. Deep sequencing uncovers numerous small RNAs on all four replicons of the plant pathogen Agrobacterium tumefaciens. RNA biology **9:**446-457.
20. **Yu GX, Snyder EE, Boyle SM, Crasta OR, Czar M, Mane SP, Purkayastha A, Sobral B, Setubal JC.** 2007. A versatile computational pipeline for bacterial genome annotation improvement and comparative analysis, with *Brucella* as a use case. Nucleic Acids Res **35:**3953-3962.
21. **Hamilton RH, Fall MZ.** 1971. The loss of tumor-initiating ability in *Agrobacterium tumefaciens* by incubation at high temperature. Experientia **27:**229-230.
22. **Hamilton RH.** 2005. Commentary on the paper: Hamilton, R. H., and Fall, M.Z. 1971. The loss of tumor-inhabiting ability in *Agrobacterium tumefaciens* by incubation at high temperature., p. 45. *In* Nester E, Gordon MP, Kerr A (ed.), *Agrobacterium tumefaciens:* From Plant Pathology to Biotechnology. APS Press, The American Phytopathological Society, St. Paul, MN.
23. **Van Larebeke N, Engler G, Holsters M, Van den Elsacker S, Zaenen I, Schilperoort RA, Schell J.** 1974. Large plasmid in *Agrobacterium tumefaciens* essential for crown gall-inducing ability. Nature **252:**169-170.
24. **Tempé J, Petit A, Holsters M, Montagu MV, Schell J.** 1977. Thermosensitive step associated with transfer of the Ti plasmid during conjugation: Possible relation to transformation in crown gall. Proc Natl Acad Sci U S A **74:**2848-2849.
25. **Rosenberg C, Huguet T.** 1984. The pAtC58 plasmid of *Agrobacterium tumefaciens* is not essential for tumour induction. Mol. Gen. Genet. **196:**533-536

.
26. **Ellis JG, Murphy PJ, Kerr A.** 1982. Isolation and properties of transfer regulatory mutants of the nopaline Ti-plasmid pTiC58. Molecular and General Genetics **186:**275-281.