The Genome of the Natural Genetic Engineer *Agrobacterium tumefaciens* C58

Derek W. Wood,[1] Joao C. Setubal,[2,4] Rajinder Kaul,[5] Dave E. Monks,[1] Joao P. Kitajima,[2,3] Vagner K. Okura,[2] Yang Zhou,[5] Lishan Chen,[1*] Gwendolyn E. Wood,[1] Nalvo F. Almeida Jr.,[6] Lisa Woo,[1] Yuching Chen,[1†] Ian T. Paulsen,[7] Jonathan A. Eisen,[7] Peter D. Karp,[8] Donald Bovee Sr.,[5] Peter Chapman,[5] James Clendenning,[5] Glenda Deatherage,[5] Will Gillet,[5] Charles Grant,[5] Tatyana Kutyavin,[5] Ruth Levy,[5] Meng-Jin Li,[5] Erin McClelland,[5] Anthony Palmieri,[5] Christopher Raymond,[5] Gregory Rouse,[5] Channakhone Saenphimmachak,[5] Zaining Wu,[5] Pedro Romero,[8] David Gordon,[9] Shiping Zhang,[10] Heayun Yoo,[10] Yumin Tao,[11] Phyllis Biddle,[10] Mark Jung,[10] William Krespan,[10] Michael Perry,[10] Bill Gordon-Kamm,[11] Li Liao,[10] Sun Kim,[10] Carol Hendrick,[11] Zuo-Yu Zhao,[11] Maureen Dolan,[10] Forrest Chumley,[10‡] Scott V. Tingey,[10] Jean-Francois Tomb,[10] Milton P. Gordon,[12] Maynard V. Olson,[5] Eugene W. Nester[1,13§]

The 5.67-megabase genome of the plant pathogen *Agrobacterium tumefaciens* C58 consists of a circular chromosome, a linear chromosome, and two plasmids. Extensive orthology and nucleotide colinearity between the genomes of *A. tumefaciens* and the plant symbiont *Sinorhizobium meliloti* suggest a recent evolutionary divergence. Their similarities include metabolic, transport, and regulatory systems that promote survival in the highly competitive rhizosphere; differences are apparent in their genome structure and virulence gene complement. Availability of the *A. tumefaciens* sequence will facilitate investigations into the molecular basis of pathogenesis and the evolutionary divergence of pathogenic and symbiotic lifestyles.

Agrobacterium tumefaciens is an α-proteobacterium of the family Rhizobiaceae and a member of the diverse *Agrobacterium* genus. A ubiquitous soil organism and etiological agent of the plant disease crown gall (*1*), *A. tumefaciens* infects more than 90 families of dicotyledonous plants, resulting in major agronomic losses (*2, 3*). The gall results from the transfer, integration, and expression of a discrete set of genes (T-DNA) located on the tumor-inducing (Ti) plasmid. Expression of these genes leads to biosynthesis of plant growth hormones as well as a bacterial nutrient source called opines (*4*). The processing and transfer of the T-DNA is mediated by the Ti plasmid virulence (*vir*) genes, and several virulence determinants initially characterized in *A. tumefaciens* have been found in plant symbionts and animal pathogens (*5–7*).

The genes within the T-DNA can be replaced by any DNA sequence, making *A. tumefaciens* an ideal vehicle for gene transfer and an essential tool for plant research and transgenic crop production. The research and commercial potential of *A. tumefaciens* has been broadened under laboratory conditions to include the transfer of T-DNA to recalcitrant plants, fungi (*8*), and human cells (*9*).

A. tumefaciens shares a similar habitat and close evolutionary relationship with the nitrogen-fixing symbionts of the Rhizobiaceae (*10*). Indeed, the introduction of a symbiotic plasmid from *Rhizobium phaseoli* into *A. tumefaciens* results in the weak but measurable formation of nitrogen-fixing root nodules (*11*), suggesting a shared genetic background. The recent publication of the genome sequences of two Rhizobiaceae, *Sinorhizobium meliloti* (*12*) and *Mesorhizobium loti* (*13*), allowed a genome-wide comparison with *A. tumefaciens*. We present the results of this comparison as well as a detailed analysis of the genome of *A. tumefaciens* strain C58 (*14, 15*).

General features of the genome. The 5.67-Mb genome of *A. tumefaciens* C58 (*16*) comprises four replicons (*17*): a circular chromosome, a linear chromosome, and the AtC58 and TiC58 plasmids (Table 1 and Fig. 1). The genome contains 5419 predicted protein-coding genes (*14*), of which we have assigned a putative function to 3475 (64.1%). The remaining 1944 genes (35.9%) include 1236 conserved hypothetical genes (22.8%) whose predicted products are similar to proteins of unknown function in other genomes, and 708 hypothetical genes (13.1%) with no significant matches in the sequence databases (Table 1). Our analysis assigns the *A. tumefaciens* genes to 501 paralogous families containing from 2 to 206 members (*14*). The two largest families are composed of genes belonging to the adenosine triphosphatase (ATPase) and membrane-spanning components of the ATP binding cassette (ABC) transport family.

The overall GC content of the *A. tumefaciens* genome is 58%. The TiC58 plasmid has two regions of distinctive GC content: the T-DNA (46%) and the *vir* region (54%) (Fig. 1). Low GC content was noted previously in the T-DNA of a related Ti plasmid (*18*). Reduced GC content (53%) is also seen within a 24-kb segment of pAtC58 (AT island, Fig. 1). This region includes 17 conserved hypothetical or hypothetical genes, an ATP-dependent DNA helicase, and an insertion sequence (IS) element. These genes are flanked by a phage integrase and a second IS element. The genes in these three regions have a distinct codon usage as compared to the rest of the genome, consistent with their recent evolutionary acquisition (*14*).

The genome contains 53 transfer RNAs (tRNAs) that represent all 20 amino acids (Table 1). These tRNAs are distributed unevenly between the circular and linear chromosomes. Transfer RNA species corresponding to the most frequently represented ala-

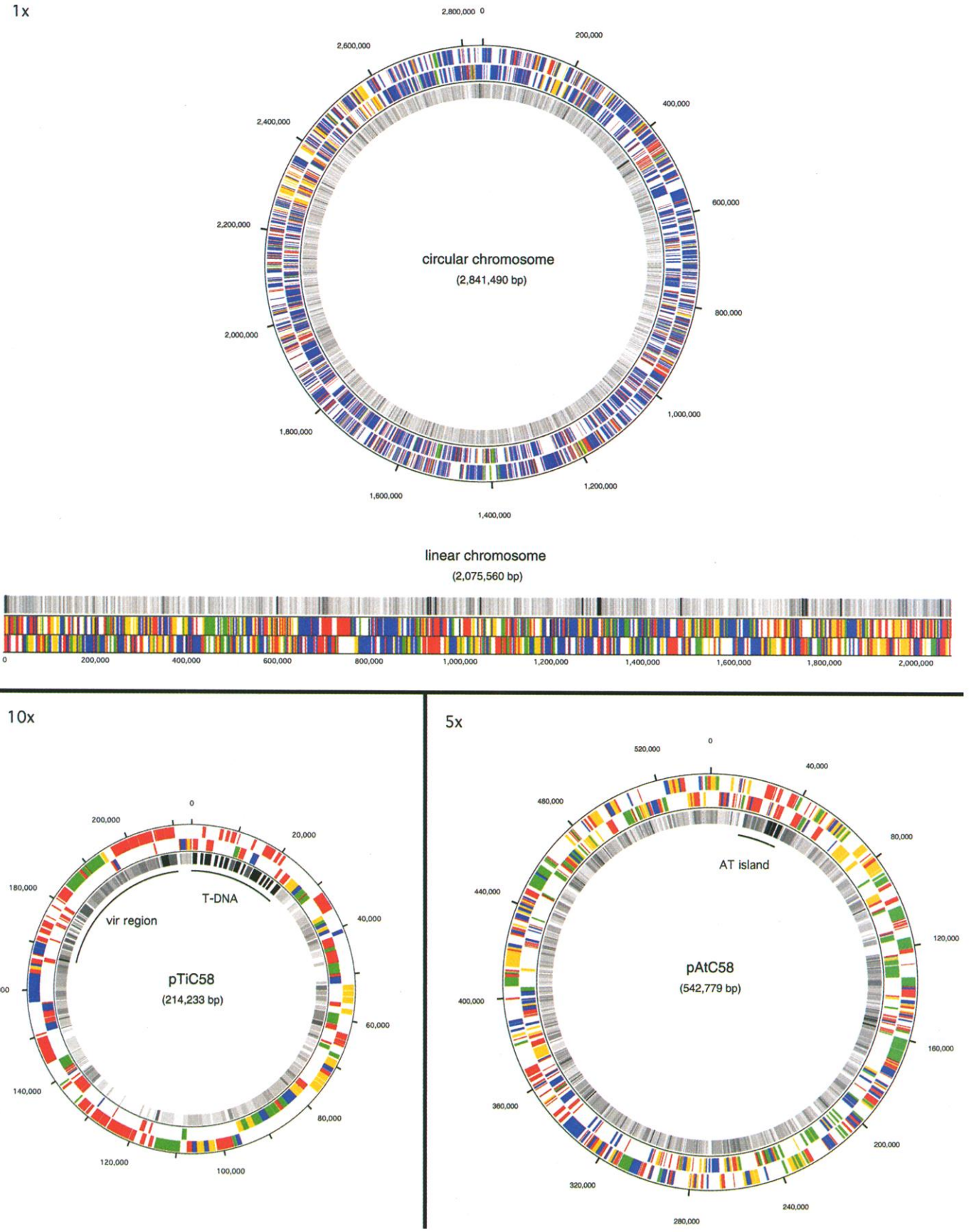
[1]Department of Microbiology, University of Washington, 1959 NE Pacific Street, Box 357242, Seattle, WA 98195, USA. [2]Bioinformatics Laboratory, Institute of Computing, [3]Center for Molecular Biology and Genetic Engineering (CBMEG), University of Campinas, CP 6176, Campinas SP 13083-970, Brazil. [4]Department of Genome Sciences, University of Washington, Box 357730, Seattle, WA 98195, USA. [5]Genome Center, University of Washington, Fluke Hall on Mason Road, Box 352145, Seattle, WA 98195, USA. [6]Department of Computing and Statistics, Federal University of Mato Grosso do Sul, CP 549, Campo Grande MS 79070-900, Brazil. [7]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. [8]Bioinformatics Research Group, SRI International, 333 Ravenswood Avenue, Menlo Park, CA 94025, USA. [9]Howard Hughes Medical Institute, University of Washington, Box 357730, Seattle, WA 98195, USA. [10]E. I. du Pont de Nemours Company, 1 Innovation Way, Newark, DE 19714, USA. [11]Pioneer Hi-Bred International Inc., 7300 NW 62nd Avenue, Post Office Box 1004, Johnston, IA 50131, USA. [12]Department of Biochemistry, University of Washington, 1959 NE Pacific Street, Seattle, WA 98195, USA. [13]Department of Botany, University of Washington, 1959 NE Pacific Street, Box 355325, Seattle, WA 98195, USA.

*Present address: Department of Pathology, University of Washington, Box 357470, Seattle, WA 98195, USA.

†Present address: Gene Function & Target Validation, Celltech R&D Inc., Bothell, WA 98021, USA.

‡Present address: Department of Plant Pathology, Kansas State University, 113 Waters Hall, Manhattan, KS 66506, USA.

§To whom correspondence should be addressed. E-mail: gnester@u.washington.edu


1x
2,800,000 0
200,000
400,000
600,000
800,000
1,000,000
1,200,000
1,400,000
1,600,000
1,800,000
2,000,000
2,200,000
2,400,000
2,600,000
circular chromosome
(2,841,490 bp)
linear chromosome
(2,075,560 bp)
0
200,000
400,000
600,000
800,000
1,000,000
1,200,000
1,400,000
1,600,000
1,800,000
2,000,000
10x
0
20,000
40,000
60,000
80,000
100,000
120,000
140,000
160,000
180,000
200,000
T-DNA
vir region
pTiC58
(214,233 bp)
5x
0
40,000
80,000
120,000
160,000
200,000
240,000
280,000
320,000
360,000
400,000
440,000
480,000
520,000
AT island
pAtC58
(542,779 bp)

nine, glutamine, and valine codons are found only on the linear replicon. The genome contains 25 predicted IS elements representing eight different families (*14*). The largest is the IS3 family comprising 10 IS elements. The IS elements are not equally distributed among the replicons but are located preferentially on the linear chromosome and pAtC58 (Table 1). The adjacent *virH1* and *virH2* genes of the Ti plasmid, encoding p450 mono-oxygenases (*19*), are flanked by IS elements, which suggests that they arrived in *A. tumefaciens* as part of a compound transposon. Twelve genes of probable phage origin were identified, most of which are on the circular chromosome (Table 1). Many of these genes cluster in two discrete regions and thus may represent prophage remnants. None of these clustered phage-related genes are shared with *S. meliloti*, which implies that they were lost from *S. meliloti* or entered the *A. tumefaciens* genome after these organisms evolutionarily diverged.

Phylogeny and whole-genome comparison. A comparison with all sequenced organisms reveals that the *A. tumefaciens* proteome is most similar to that of two rhizobial species, *S. meliloti* and *M. loti* (*14*). This result was obtained by cataloging top BLAST hits of predicted *A. tumefaciens* proteins and by classifying predicted proteins into clusters of orthologous groups (Fig. 2A) (*20*). Of the two rhizobial species, the *A. tumefaciens* proteome is most similar to that of *S. meliloti*. Phylogenetic analyses of broadly conserved proteins indicate that this similarity results from *A. tumefaciens* and *S. meliloti* sharing a recent common ancestor, and not from gene loss or branch rate variation (Fig. 2, B and C).

Sinorhizobium meliloti has a circular chromosome (3.65 Mb) and two plasmids (1.68 Mb and 1.35 Mb), with a total genome size 1.1 Mb larger than that of *A. tumefaciens* (*12*). The circular chromosomes of these organisms show extensive nucleotide colinearity and gene order conservation (Fig. 3) (*14*). Previously, such extensive colinearity has only been seen between members of the same genus. Chromosome-wide conservation of gene order is less pronounced between *S. meliloti* and *M. loti* (*14*). The comparison of the circular chromosomes of *A. tumefaciens* and *S. meliloti* also reveals major rearrangements near the putative replication origin and termini (Fig. 3, regions A and B). Similarly located rearrangements are commonly seen between closely related bacteria (*21*).

A comparison of the other replicons of *A. tumefaciens* with all replicons of *S. meliloti* reveals a mosaic pattern of ortholog distribution (Table 2 and Fig. 1). These orthologs are distributed across the *A. tumefaciens* elements as individual genes and small regions of gene order conservation. Two regions of the linear replicon exhibit extensive conservation of gene order with a segment of the *S. meliloti* chromosome (Fig. 3, region C). The first comprises 46 genes (44 kb) and the second contains 65 genes (89 kb). These regions are partially conserved in the *M. loti* chromosome. The large number of orthologs and the lack of extensive gene order conservation suggest that the smaller *A. tumefaciens* replicons underwent substantial rearrangement since the organisms diverged. This finding is consistent with differential evolutionary pressures acting on these elements. The nonorthologous genes, many of which are seen on the Ti plasmid, reflect lineage-specific gene loss or acquisition from other species. Taken together, these data support the recent evolutionary divergence of *A. tumefaciens* and *S. meliloti*.

Genus-specific genes. Comparison of the genomes of *A. tumefaciens*, *S. meliloti*, and *M. loti* identified genes in each organism that likely contribute to genus-specific biology (*14*). Of the 5419 predicted *A. tumefaciens* proteins, 853 (16%) are not found in these other organisms. Of these, 97 have an assigned function, whereas 756 are hypothetical or conserved hypothetical. The predicted products of these genes are diverse and include proteins involved in cellulose production, plasmid maintenance, cell growth, transcriptional regulation, and cell wall synthesis. Several additional proteins are predicted to catabolize plant cell wall materials, sugars, and exudates. These include polygalacturonases, a glycosidase, an endoglucanase, a myo-inositol catabolism protein, and a cell wall lysis–associated protein. Additional genes, predictably found on the Ti plasmid, include those encoding virulence, T-DNA, and conjugal transfer–associated proteins. With 756 open reading frames (ORFs) yet to characterize, much remains to be elucidated regarding the genetic distinction between *A. tumefaciens* and its Rhizobiaceae relatives.

Linear chromosome. Linear replicons, the predominant genetic element in eukaryotes, have been identified in only a few prokaryotes. These include members of the genera *Borrelia* and *Streptomyces* (*22, 23*). Although sequence analysis did not reveal distinct features associated with terminal secondary structures, Goodner *et al.* found that the termini of this replicon are covalently linked (*15*). This covalent linkage did not prevent nearly complete sequencing of the replicon termini as confirmed by Southern analysis (*14*). Proteins associated with the maintenance of linear ends in other systems, such as telomerases or the *Streptomyces tpg* proteins (*24*), are absent in *A. tumefaciens*. One notable feature of the replicon termini is the presence of IS elements near each end. The evolutionary origin of this replicon awaits investigation, as does the mechanism that *A. tumefaciens* uses to maintain it in a linear form.

There are 1882 protein-coding genes on the linear replicon, including those encoding ribosomal and DNA replication proteins, as

Fig. 1. (facing page) Schematic representation of the *A. tumefaciens* genome. Chromosomes are drawn to scale with plasmids represented at 5× or 10× magnification, as indicated. The outer two bands indicate opposing transcriptional orientations of predicted genes. Colors indicate orthology to proteins in the *S. meliloti* replicons: blue, chromosome; green, pSymA; gold, pSymB; red, nonorthologous. The inner circle depicts GC content for each coding region, with lower GC content indicated by darker shading. The *vir* and T-DNA regions of pTiC58 and the AT island of pAtC58 are indicated. Orthologs were identified by comparison of predicted proteins for each *A. tumefaciens* replicon with the genome of *S. meliloti*. Two proteins were considered orthologous if their BLASTP alignment covered at least 60% of each protein at an expect value of less than or equal to 10^{-5}. Proteins that did not match these criteria were considered nonorthologous (*14*).

Table 1. General features of the *A. tumefaciens* C58 genome.

Feature	Circular	Linear	pAtC58	pTiC58	Total
Size (bp)	2,841,490	2,075,560	542,779	214,233	5,674,062
G+C content (%)	59.4	59.3	57.3	56.7	58.13
Protein-coding genes					
Assigned function	1715	1286	333	141	3475 (64.1%)
Conserved hypothetical	710	353	128	45	1236 (22.8%)
Hypothetical	364	243	89	12	708 (13.1%)
Total	2789	1882	550	198	5419
Average ORF size (bp)	892	988	843	925	922
Coding (%)	87.9	89.9	85.4	85.5	88.3
Regulators (%)	7.7	10.4	11.8	5.1	9.0
ABC transport	47	80	20	6	153
RNA					
rRNA	2	2	0	0	4
tRNA	40	13	0	0	53
tmRNA	1	0	0	0	1
IS elements	2	10	10	2	24
Phage-related	10	1	1	0	12

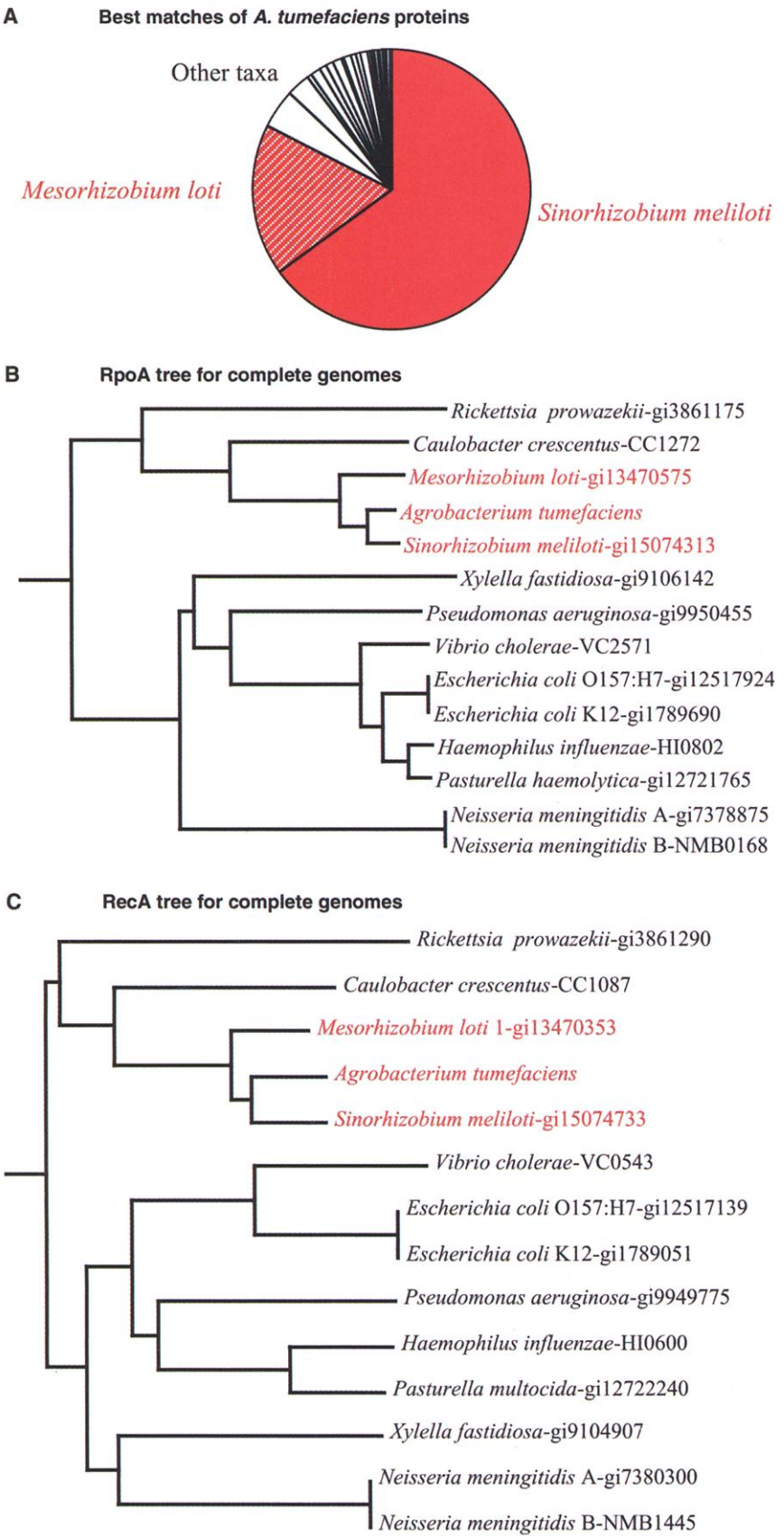


Fig. 2. Comparisons with fully sequenced genomes. (**A**) Distribution of best hits based on a comparison of predicted proteins of *A. tumefaciens* with proteins from all published genomes. (**B** and **C**) Phylogenetic trees generated using two broadly conserved proteins. The trees were generated using PAUP distance methods and a distance calculation based on PAM matrices (*14*).

well as 21 complete metabolic pathways. The presence of these genes confirms the chromosomal identity of this replicon. Additional features, however, resemble those traditionally associated with plasmids. For example, genes whose products are similar to the conjugative proteins TraA, MobC, and TraG are present, although an *oriT* is not apparent. Further, an intact and highly conserved *repABC* operon, the definitive element of the RepABC-type replicator family of circular plasmids, is located near the center of the linear chromosome. The presence of this operon, coupled with a colocalized GC-skew inversion, indicates a bidirectional plasmid-like mode of replication. If experimentally verified, this replication mechanism would prove unique among known linear replicons.

Plasmid replication and transfer. Replication of both pTiC58 and pAtC58 is mediated by RepABC-type systems commonly found in plasmids of the Rhizobiaceae. It is likely that the origin of replication for these plasmids is adjacent to the *repC* gene (*25*). In contrast to the pSymB plasmid of *S. meliloti* (*12*), both *A. tumefaciens* plasmids contain all necessary machinery for conjugation and do not contain essential genes. A new conjugal transfer system belonging to the Type IV secretion family (AvhB) (*26*) was identified on pAtC58. In contrast to the tight control of Ti plasmid conjugal transfer mediated by specific opines that activate quorum sensing (*27*), the conjugal transfer of pAtC58 appears to be constitutive.

Transport. Transporters constitute 15% of the *A. tumefaciens* genome, 87% of which are found on the chromosomes (*14*). These systems are predicted to confer broad capabilities for the transport of common nutrients found in the rhizosphere, including sugars, amino acids, and peptides. In addition, there are 11 LysE/RhtB amino acid efflux proteins, almost double the number seen in any bacterium outside of the Rhizobiaceae (*12, 28*). These transporters may function in the export of homoserine lactones or other signal molecules. There are also a large number of high-affinity tripartite ATP-independent periplasmic (TRAP) dicarboxylate transporters (*29*). Our analyses indicate that *A. tumefaciens* and the other sequenced members of the Rhizobiaceae have similar transport capabilities.

Like both *S. meliloti* and *M. loti*, *A. tumefaciens* has an abundance of ABC transporters, constituting 60% of its total transporter complement. There are 153 complete systems plus additional "orphan" subunits. The number of ABC transporters found in these organisms is greater than that found in any sequenced eukaryote and more than double the number found in any sequenced bacterium (*28, 30*). Predicted substrates of these ABC transport systems include sugars (53 systems), amino acids (29 systems), and pep-

tides (25 systems). Other organisms with large ABC transporter complements include photosynthetic bacteria such as *Synechocystis* PCC6803 and organisms that lack a tricarboxylic acid (TCA) cycle and an electron transfer chain, like the mycoplasmas and *Thermotoga maritima* (*28*). The generation of large ATP pools in these organisms, via photosynthesis or F-type ATPases, may explain their preference for ATP-driven transport. In contrast, the preference for ABC transporters in *A. tumefaciens* may reflect a need for high-affinity uptake systems for the acquisition of nutrients in the highly competitive soil and rhizosphere environments.

Regulation. Bacteria that inhabit diverse environments tend to have large complements of regulatory genes (*31*). Consistent with this, regulatory genes constitute a substantial proportion (9%) of the *A. tumefaciens* genome (Table 1) (*14*). This regulatory capacity likely facilitates survival of *A. tumefaciens* within the dynamic soil and rhizosphere environments. The genome encodes 11 extracellular sigma factors, proteins implicated in stress responses in other organisms (*32*). In addition, although several LuxR family motifs are evident, only one previously identified acyl-homoserine lactone synthase (*traI*) known to be involved in quorum sensing was detected (*4*). Several proteins are similar to eukaryotic signal transduction proteins rarely found in bacteria, including four regucalcin-like calcium-binding regulators and a serine-threonine kinase. As is true of other α-proteobacteria, no *rpoS* gene was identified. However, *A. tumefaciens* does have a homolog of the HF-1 protein known to regulate stationary phase and oxidative stress responses in *Escherichia coli* and *Brucella abortus* (*33*).

Our analysis identified numerous nucleotide cyclases in the plant symbionts *S. meliloti* and *M. loti* (25 and 12, respectively) and in the evolutionarily distinct human pathogen *Mycobacterium tuberculosis* (12). These cyclases are rarely found in other bacterial genomes. The nucleotide cyclases in *S. meliloti* have been noted previously and were postulated to function in signal transduction (*12*). Contrary to our expectation, there are only three nucleotide cyclases in *A. tumefaciens*. It is unclear why the nitrogen-fixing plant symbionts share similarly large numbers of nucleotide cyclases with a human pathogen, whereas few such genes are found in the evolutionarily related *A. tumefaciens*.

Attachment, cell surface, and secretion. The initial interaction of *Agrobacterium* with its plant hosts is mediated by several attachment-related genes (*34*). These include the *chvA*, *chvB*, *exoC*, and cellulose synthesis genes as well as the pAtC58-localized *att* region. Several additional genes encode proteins similar to adhesins in mammalian pathogens, including BfrA of *E. coli* and PsaA of *Streptococcus pneumoniae*.

Pili are extracellular appendages often required for bacterial association with their hosts. Although only the pilus encoded by the *virB* operon has been experimentally confirmed (*35*), the *trb* operon required for Ti plasmid conjugation likely produces a pilus (*36*). The *avhB* and *ctp* clusters, identified by our analyses, may also produce pili. Additional surface components include exopolysaccharides, lipopolysaccharides, and capsular polysaccharides, whose biosynthetic genes are primarily located on the linear chromosome. Such surface polysaccharides are commonly involved in invasion, growth, and survival of plant-associated bacteria.

Five protein secretion systems are found among Gram-negative bacteria (*37*), at least three of which are represented in *A. tumefaciens*. These include four potential type I secretion systems. Although components of the main terminal branch of type II secretion appear to be absent, the Sec system for protein secretion across the inner membrane is intact. There is, however, a type IV pilus biogenesis system with components similar to those of type II secretion systems. Similar to *S. meliloti* (*12*), no type III secretion system was identified. *Agrobacterium tumefaciens* encodes three type IV secretion systems: VirB, Trb (*27*), and AvhB. The genome also contains the twin arginine targeting system, Tat/Mtt (*38*).

Virulence. To date, most virulence determinants of *A. tumefaciens* have been found on the Ti plasmid. Other than the *virB* operon, these genes are not found in *S. meliloti*. The TiC58 plasmid contains a single T-DNA region, in contrast to the two found in a number of other strains (*18*), and the 25–base pair (bp) border regions that delineate the T-DNA are not present elsewhere in the genome.

The availability of the genome sequence has enabled the identification of genes whose products are similar to plant pathogen virulence proteins required for host cell wall degradation. These include pectinase (*kdgF*), ligninase (*ligE*), and xylanase as well as regulators of pectinase and cellulase production (*pecS/M*); *A. tumefaciens* may use such enzymes to breach the cell wall of its host before T-DNA transfer.

In addition, we have identified numerous orthologs of animal virulence genes. Examples include those involved in host survival, such as the *bacA* locus of *Brucella* (*39*) and two members of the widely conserved HtrA family of serine proteases implicated in response to oxidative stress in *Salmonella* and *Yersinia* (*40*). Interestingly, a *bacA* homolog is involved in *S. meliloti* symbiosis (*41*). Invasion-related homologs include the *ialA* and *ialB* genes of *Bartonella henselae* (*42*) as well as five hemolysin-like proteins with associated type I secretion systems. The highly conserved *mviN* gene, implicated in *Salmonella* virulence (*43*), is also present.

Metabolism. *Agrobacterium tumefaciens* grows on minimal medium and therefore possesses all pathways required for pro-

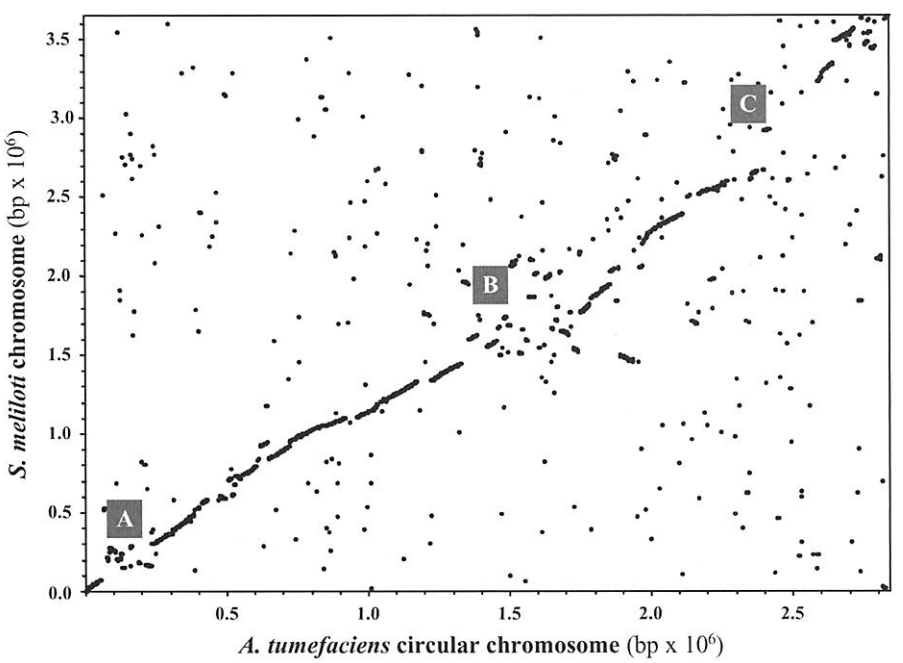


Fig. 3. Alignment of the proteomes of the *S. meliloti* chromosome and the *A. tumefaciens* circular chromosome. Each point in the figure is a bidirectional best hit. These hits were obtained by pairwise BLASTP searches of predicted *A. tumefaciens* proteins against those of *S. meliloti* with a maximum expect value of 10^{-4} (*14*). Putative origins (region A) and termini (region B) of replication are indicated, as well as a sizable region lacking colinearity (region C).

totrophic growth, an observation confirmed by our computational pathway analysis (*14*). These metabolic pathways are dispersed among the four replicons. Unlike their organization in *E. coli*, most genes of these pathways are not tightly clustered, which suggests that they are not present in operons. We identified pathways for the synthesis of all 20 amino acids as well as numerous enzyme cofactors. At least one nonribosomal protein synthesis system for the production of polyketides was identified. Encoded energy metabolism pathways include glycolysis, TCA cycle, and Entner-Doudoroff. *Agrobacterium tumefaciens* can catabolize 17 amino acids, including *S*-adenosylhomocysteine and 4-hydroxyproline. Pathways for use or degradation of plant metabolites typically found in the rhizosphere were also detected. These include sugars such as glucose, fructose, sucrose, ribose, xylose, xylulose, and lactose as well as compounds such as myo-inositol, hydantoin, urea, and glycerol. The capacity to metabolize glucuronate, galactonate, galactarate, gluconate, ribitol, glycogen, quinate, L-idonate, creatinine, stachydrine, ribosylnicotinamide, and 4-hydroxymandelate was also detected. Chemotaxis systems responding to many of these compounds are present in *A. tumefaciens* (*44*).

Agrobacterium tumefaciens encodes a variety of proteins that may protect against toxic compounds in the environment. Examples include four cytochrome p450s, two of which have been previously identified. One of these has been shown to modify ferrulic acid, an inducer of the *vir* genes (*45*). These highly oxidative enzymes may also detoxify or modify plant-derived compounds, including phytoalexins (*46*) and protocatechuate, and xenobiotics such as 1,2-dichloroethane, cyanate, 1,4-dichlorobenzene, and octane. In addition, antibiotic resistance genes targeted against tetracycline (*47*) and chloramphenicol are present.

Many components of nitrogen metabolism are conserved between *A. tumefaciens* and the nitrogen-fixing symbionts *S. meliloti* and *M. loti*. Examples include components of the nitrogen regulation (Ntr) system such as *ntrBC*, *ntrXY*, *ntrA*, *glnE*, *glnD*, *glnB*, and *glnK*. *Agrobacterium tumefaciens* harbors seven glutamine synthetase (GS) genes, which encode GS types I, II, and III. The presence of multiple GS genes may relate to the observation that *A. tumefaciens* requires high concentrations of glutamate for optimal growth. Other members of the Rhizobiaceae also contain multiple GS genes. In addition, *A. tumefaciens* has a gene predicted to encode the large hexameric adenosine monophosphate (AMP)–dependent glutamate dehydrogenase (*48*), but a gene encoding the AMP-independent glutamate dehydrogenase was not identified. In contrast, *S. meliloti* and *M. loti* contain both genes. The *A. tumefaciens* linear chromosome carries denitrification genes, including a periplasmic dissimilatory nitrate reductase (*nap*), nitrite reductase (*nir*), and nitric oxide reductase (*nor*), but lacks nitrous oxide reductase (*nos*). In contrast, all of these genes are present in *S. meliloti*. Nitrate transport genes are also located on the linear chromosome. Although *A. tumefaciens* is considered an aerobe, the existence of these genes implies that it could use nitrate as an electron acceptor under anaerobic conditions. As expected, *A. tumefaciens* lacks the subunits of nitrogenase and its cofactors. Most of the *nod* genes are also absent, except for three genes similar to those involved in nod factor production, *nodL*, *nodX*, and *nodN*.

Conclusions. The combination of a linear and a circular chromosome is found in only a few members of the genus *Agrobacterium* (*49*). This observation represents a key evolutionary distinction between *A. tumefaciens* and *S. meliloti*. On the basis of 16*S* ribosomal DNA phylogenetic analyses, it has been proposed that the genus *Agrobacterium* be reclassified into the genus *Rhizobium* (*10*). Combining what has been elucidated regarding genome structure with the complete genome sequence should allow a more accurate definition of the taxonomic position of *A. tumefaciens* in the Rhizobiaceae.

One striking finding from our analysis is the extensive similarity of the circular chromosomes of *A. tumefaciens* and the plant symbiont *S. meliloti*, which supports the view that these bacteria originated from a recent common ancestor. Galibert *et al.* speculate that the *S. meliloti* chromosome was present in a progenitor that later acquired pSymA and pSymB (*12*). The mosaic structure of the *A. tumefaciens* linear chromosome and plasmids, predominantly composed of orthologs found on each of the *S. meliloti* replicons, suggests that these organisms diverged after acquisition of the pSymA and pSymB ancestral molecules by this progenitor.

Recent models of bacterial evolution suggest that the differential acquisition and loss of genes in organisms that inhabit the same environment allows divergence into symbiotic and pathogenic lifestyles (*50*). The acquisition of such elements is apparent in both *A. tumefaciens* and *S. meliloti*. The *nod* genes of *S. meliloti* (*12, 51*), as well as the *vir* genes and T-DNA of *A. tumefaciens*, display GC content and codon usage distinct from the rest of the genome, which suggests recent evolutionary acquisition. In the case of the T-DNA, reduced GC content may facilitate expression in the plant host, where lower GC content is common. Moreover, none of the T-DNA and few of the *vir* genes of *A. tumefaciens* have orthologs in *S. meliloti*, and most *nod* genes are not found in *A. tumefaciens*. Differential selection and maintenance of such horizontally acquired genes likely led to the divergence into pathogenic and symbiotic states. Thus, these organisms provide a rich model system for further investigations into the evolutionary divergence of pathogens and symbionts.

As the central biological tool in the generation of transgenic plants for research and agriculture, *A. tumefaciens*, and the availability of its genome sequence, will continue to have an impact on plant biotechnology. Detailed studies, supplemented by this sequence, should lead to a directed refinement of plant transformation that increases both the host range and transformation efficiency of this versatile genetic tool. Genes likely to be targeted by such work include potential virulence factors that are shared between plant and animal pathogens. Examination of these genes in the genetically tractable *Agrobacterium* system may also serve to elucidate the molecular role they play in animal pathogens. It is our hope that this work will broaden the scientific foundation from which to address the worldwide debate over the production, use, and safety of genetically modified organisms.

Table 2. Number of orthologous genes of *A. tumefaciens* with respect to *S. meliloti*. The number of orthologous genes is shown in bold, with the percentage of each *A. tumefaciens* replicon they represent shown in square brackets. The remainder of the genes, which are not orthologs, are shown in the last row. Numbers of putative protein coding genes for each replicon are shown in parentheses.

		A. tumefaciens			
		Circular (2789)	Linear (1882)	pAtC58 (550)	pTiC58 (198)
S. meliloti	Chromosome (3341)	**1867** [67%]	**673** [36%]	**114** [21%]	**30** [15%]
	pSymA (1293)	**104** [4%]	**218** [12%]	**118** [21%]	**34** [17%]
	pSymB (1570)	**221** [8%]	**478** [25%]	**108** [20%]	**23** [12%]
	Nonorthologous	**597** [21%]	**513** [27%]	**210** [38%]	**111** [56%]

References and Notes

1. E. F. Smith, C. O. Townsend, *Science* **25**, 671 (1907).
2. M. DeCleene, J. DeLey, *Bot. Gaz.* **42**, 389 (1976).
3. L. Moore, in *Biology of the Rhizobiaceae*, K. L. Giles, A. G. Atherly, Eds. (Academic Press, New York, 1981), pp. 15–46.
4. J. Zupan, T. R. Muth, O. Draper, P. Zambryski, *Plant J.* **23**, 11 (2000).
5. P. J. Christie, *Mol. Microbiol.* **40**, 294 (2001).
6. N. Inon de Iannino, G. Briones, M. Tolmasky, R. A. Ugalde, *J. Bacteriol.* **180**, 4392 (1998).
7. A. Sola-Landa *et al.*, *Mol. Microbiol.* **29**, 125 (1998).
8. M. J. de Groot, P. Bundock, P. J. Hooykaas, A. G. Beijersbergen, *Nature Biotechnol.* **16**, 839 (1998).

9. T. Kunik *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 1871 (2001).
10. J. M. Young, L. D. Kuykendall, E. Martinez-Romero, A. Kerr, H. Sawada, *Int. J. Syst. Evol. Microbiol.* **51**, 89 (2001).
11. E. Martinez, R. Palacios, F. Sanchez, *J. Bacteriol.* **169**, 2828 (1987).
12. F. Galibert *et al.*, *Science* **293**, 668 (2001).
13. T. Kaneko *et al.*, *DNA Res.* **7**, 331 (2000).
14. The complete annotated sequence, detailed methods, and all supplementary data, which will be periodically updated, are available at our Web site (www.agrobacterium.org). Supplemental data are also available at *Science* Online (www.sciencemag.org/cgi/content/full/294/5550/2317/DC1). The sequence has been deposited at GenBank (accession numbers: circular chromosome, AE008688; linear chromosome, AE008689; pAtC58, AE008687; pTiC58, AE008690).
15. At a late phase of this project, we became aware of an independent effort by Goodner *et al.* (*52*) to sequence the C58 strain of *A. tumefaciens*. Comparison of the two independently finished sequences indicates excellent overall agreement. There are two small insertions in the other sequence, one in the circular chromosome and one in pAtC58, relative to our sequence. Single nucleotide discrepancies are far below current accuracy standards for finished genomic sequence (99.99%). Further work will determine which of the observed discrepancies reflect strain differences and which reflect sequencing errors.
16. R. H. Hamilton, M. Z. Fall, *Experientia* **27**, 229 (1971).
17. A. Allardet-Servent, S. Michaux-Charachon, E. Jumas-Bilak, L. Karayan, M. Ramuz, *J. Bacteriol.* **175**, 7869 (1993).
18. J. Zhu *et al.*, *J. Bacteriol.* **182**, 3885 (2000).
19. V. S. Kalogeraki, S. C. Winans, *J. Bacteriol.* **180**, 5660 (1998).
20. R. L. Tatusov *et al.*, *Nucleic Acids Res.* **29**, 22 (2001).
21. J. A. Eisen, J. F. Heidelberg, O. White, S. L. Salzberg, *Genome Biol.* **1**, Research 0011.1 (2000).
22. C. M. Fraser *et al.*, *Nature* **390**, 580 (1997).
23. J. N. Volff, J. Altenbuchner, *FEMS Microbiol. Lett.* **186**, 143 (2000).
24. K. Bao, S. N. Cohen, *Genes Dev.* **15**, 1518 (2001).
25. M. A. Ramirez-Romero, N. Soberon, A. Perez-Oseguera, J. Tellez-Sosa, M. A. Cevallos, *J. Bacteriol.* **182**, 3117 (2000).
26. L. Chen, unpublished data.
27. S. K. Farrand, in *The Rhizobiaceae*, H. P. Spaink, A. Kondorosi, P. J. J. Hooykaas, Eds. (Kluwer Academic, Dordrecht, Netherlands, 1998), pp. 199–233.
28. I. T. Paulsen, L. Nguyen, M. K. Sliwinski, R. Rabus, M. H. Saier Jr., *J. Mol. Biol.* **301**, 75 (2000).
29. J. A. Forward, M. C. Behrendt, N. R. Wyborn, R. Cross, D. J. Kelly, *J. Bacteriol.* **179**, 5482 (1997).
30. R. Sanchez-Fernandez, T. G. Davies, J. O. Coleman, P. A. Rea, *J. Biol. Chem.* **276**, 30231 (2001).
31. C. K. Stover *et al.*, *Nature* **406**, 959 (2000).
32. D. Missiakas, S. Raina, *Mol. Microbiol.* **28**, 1059 (1998).
33. G. T. Robertson, R. M. Roop Jr., *Mol. Microbiol.* **34**, 690 (1999).
34. A. G. Matthysse, G. W. Kijne, in *The Rhizobiaceae*, H. P. Spaink, A. Kondorosi, P. J. J. Hooykaas, Eds. (Kluwer Academic, Dordrecht, Netherlands, 1998), pp. 235–249.
35. K. J. Fullner, J. C. Lara, E. W. Nester, *Science* **273**, 1107 (1996).
36. P. L. Li, I. Hwang, H. Miyagi, H. True, S. K. Farrand, *J. Bacteriol.* **181**, 5033 (1999).
37. J. R. Harper, T. J. Silhavy, in *Principles of Bacterial Pathogenesis*, E. A. Groisman, Ed. (Academic Press, San Diego, CA, 2001), pp. 43–74.
38. J. H. Weiner *et al.*, *Cell* **93**, 93 (1998).
39. K. LeVier, R. W. Phillips, V. K. Grippe, R. M. Roop II, G. C. Walker, *Science* **287**, 2492 (2000).
40. M. J. Pallen, B. W. Wren, *Mol. Microbiol.* **26**, 209 (1997).
41. A. Ichige, G. C. Walker, *J. Bacteriol.* **179**, 209 (1997).
42. S. A. Coleman, M. F. Minnick, *Infect. Immun.* **69**, 4373 (2001).
43. M. Carsiotis, B. A. Stocker, D. L. Weinstein, A. D. O'Brien, *Infect. Immun.* **57**, 3276 (1989).
44. Y. Dessaux, A. Petit, S. K. Farrand, P. J. Murphy, in *The Rhizobiaceae*, H. P. Spaink, A. Kondorosi, P. J. J. Hooykaas, Eds. (Kluwer Academic, Dordrecht, Netherlands, 1998), pp. 173–197.
45. V. S. Kalogeraki, J. Zhu, A. Eberhard, E. L. Madsen, S. C. Winans, *Mol. Microbiol.* **34**, 512 (1999).
46. K. J. Miller, J. M. Wood, *Annu. Rev. Microbiol.* **50**, 101 (1996).
47. Z. Q. Luo, S. K. Farrand, *J. Bacteriol.* **181**, 618 (1999).
48. B. Minambres, E. R. Olivera, R. A. Jensen, J. M. Luengo, *J. Biol. Chem.* **275**, 39529 (2000).
49. E. Jumas-Bilak, S. Michaux-Charachon, G. Bourg, M. Ramuz, A. Allardet-Servent, *J. Bacteriol.* **180**, 2749 (1998).
50. H. Ochman, N. A. Moran, *Science* **292**, 1096 (2001).
51. J. A. Downie, J. P. Young, *Nature* **412**, 597 (2001).
52. B. Goodner *et al.*, *Science* **294**, 2323 (2001).

53. We thank P. Green for useful discussions, R. Gibson for providing the genome_plot program, and J. Staley for critical review of the manuscript. Supported by NSF grant 0092815 (M.V.O.), NIH Public Health Service grant GM32618 (E.W.N.), NIH doctoral fellowship GM19642 (D.W.W.), Sao Paulo State Research Foundation (FAPESP) sabbatical fellowship 1999/11876-5 (J.C.S.), a FUNDECT-MS grant (N.F.A.), and the E. I. du Pont de Nemours Company.

3 October 2001; accepted 9 November 2001

Genome Sequence of the Plant Pathogen and Biotechnology Agent *Agrobacterium tumefaciens* C58

Brad Goodner,[1,2] Gregory Hinkle,[3] Stacie Gattung,[4] Nancy Miller,[4] Mary Blanchard,[4] Barbara Qurollo,[3] Barry S. Goldman,[3,4] Yongwei Cao,[3] Manor Askenazi,[3] Conrad Halling,[3] Lori Mullin,[3] Kathryn Houmiel,[4] Jeffrey Gordon,[3] Mark Vaudin,[4] Oleg Iartchouk,[3] Andrew Epp,[3] Fang Liu,[3] Clifford Wollam,[4] Mike Allinger,[4] Dahlia Doughty,[2] Charlaine Scott,[2] Courtney Lappas,[2] Brian Markelz,[2] Casey Flanagan,[2] Chris Crowell,[2] Jordan Gurson,[2] Caroline Lomo,[2] Carolyn Sear,[2] Graham Strub,[2] Chris Cielo,[2] Steven Slater[3]*

Agrobacterium tumefaciens **is a plant pathogen capable of transferring a defined segment of DNA to a host plant, generating a gall tumor. Replacing the transferred tumor-inducing genes with exogenous DNA allows the introduction of any desired gene into the plant. Thus, *A. tumefaciens* has been critical for the development of modern plant genetics and agricultural biotechnology. Here we describe the genome of *A. tumefaciens* strain C58, which has an unusual structure consisting of one circular and one linear chromosome. We discuss genome architecture and evolution and additional genes potentially involved in virulence and metabolic parasitism of host plants.**

Agrobacterium tumefaciens is a plant pathogen with the unique ability to transfer a defined segment of DNA to eukaryotes, where it integrates into the eukaryotic genome. This ability to transfer and integrate DNA is used for random mutagenesis and has been adapted into a powerful tool for production of transgenic plants, including soybean, maize, and cotton (*1*, *2*). *A. tumefaciens* was identified early in the 20th century as the causal agent of crown gall disease in plants (*3*). Pathogenesis is initiated when *Agrobacterium* detects small molecules released by actively growing cells in a plant wound. These molecules induce a series of virulence (*vir*) genes whose encoded products export the single-stranded transferred DNA (T-DNA) to the plant cell, where it integrates into the genome at an essentially random location. Once integrated, T-DNA gene expression alters plant hormone levels, leading to cell proliferation typical of a gall tumor. The T-DNA also encodes enzymes for the synthesis of opines, a class of nutrient molecules used almost exclusively by *A. tumefaciens* (*4–7*).

A. tumefaciens strains fall into three biovars, which differ in their host range, metabolic characteristics, relationships with other genera in the family Rhizobiaceae, and potentially their chromosome structure (*4–13*). The taxonomy of the Rhizobiaceae family is not without controversy, but we expect that

[1]Department of Biology, Hiram College, Hiram, OH 44234, USA. [2]Department of Biology, University of Richmond, Richmond, VA 23173, USA. [3]Cereon Genomics, LLC, 45 Sidney Street, Cambridge, MA 02139, USA. [4]Monsanto Company, 800 North Lindbergh Boulevard, St. Louis, MO 63167, USA.

*To whom correspondence should be addressed. E-mail: steven.c.slater@cereon.com