The Genome of Agrobacterium tumefaciens C58: Reconciliation of sequence data, updated annotation, and distribution of linear chromosome in genus Agrobacterium.

Running Headline: Agrobacterium tumefaciens C58 genome update

Steven Slater, João C. Setubal, Brad Goodner, Kathryn Houniel, Jian Sun, Rajinder Kaul, Barry S. Goldman, Stephen K. Farrand, Nalvo Almeida Jr., Thomas Burr, Eugene Nester, David M. Rhoads, Ryosuke Kado, Trucian Ostheimer, Nicole Pride, Allison Sabo, Erin Henry, Erin Telepak, Lindsey Crome, Alana Harkleroad, Louis Oliphant, Phil Pratt-Szegila, Roy Welch, and Derek Wood

Address: Great Lakes Bioenergy Research Center and Department of Bacteriology, The University of Wisconsin – Madison, 1550 Linden Dr., Madison, WI 53706; Previous address: The Biodesign Institute, Arizona State University Tempe, AZ, 85287; Institute of Chemistry, University of São Paulo, São Paulo, SP 05508, Brazil; Virginia Bioinformatics Institute and Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA 24060; Department of Biology, Hiram College, Hiram, OH 44234; Department of Biology, Seattle Pacific University, Seattle, WA 98119; University of Washington Genome Center, Seattle, WA 98195-2145; Monsanto Company, 800 North Lindbergh Boulevard, St. Louis, MO 63167; Department of Microbiology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801; Faculty of Computing, Federal University of Mato Grosso do Sul, Campo Grande, Brazil; College of Agriculture and Life Sciences, Cornell University, and...
Address correspondence to Brad Goodner – goodnerbw@hiram.edu
ABSTRACT

Two groups independently sequenced the *Agrobacterium tumefaciens* C58 genome in 2001. We report here consolidation of these sequences, updated annotation, and additional analysis of the evolutionary history of the linear chromosome, which is apparently limited to the biovar I group of *Agrobacterium*. 
Agrobacterium tumefaciens C58 has an unusual genome structure consisting of one circular chromosome (Chromosome I), one linear chromosome (Chromosome II), and two plasmids (1-5). Previous studies showed that the linear chromosome is derived from a plasmid (4, 5). Isolates of Agrobacterium spp. have traditionally been subdivided into three different groups, called biovars, based on differences in physiology and host-range. Biovar I can be further subdivided into genomovars, with C58 belonging to genomovar 8 (6-10).

C58 was originally isolated in 1958 by Robert Dickey from a cherry gall in upstate New York (11). Lead authors of this manuscript independently sequenced the genomes of two isolates of A. tumefaciens C58 in 2001 (4, 5). Wood et al. (5) sequenced a C58 strain stored in frozen glycerol in the laboratory of Eugene Nester at the University of Washington (hereafter designated C58UW). Goodner et al. (4) sequenced the ATCC33970 isolate obtained from the American Type Culture Collection (ATCC) in 1999. This strain, also originating from the Nester lab via John Kleyn, was deposited in 1981 and subcultured three times by ATCC and once by researchers at the Monsanto Company prior to sequencing. The number of passages separating these strains from each other or the original strain isolated by Dickey is unknown.

A comparison of the two independent genome sequences identified 52 differences including two insertion/deletions (indels; Table S1). All disparate loci were resequenced following PCR amplification (See Supplemental Materials and Methods). Twenty-two of these apparent differences were base-calling errors and 30 were true differences. Of the 30 true differences, sixteen were single base changes residing in the 16s rRNA and tRNA-Ile region near 58.3 kbp on Chromosome I, apparently resulting from
recombination between rRNA loci. C58UW also contains two deletions relative to ATCC33970. The first is a 90 bp in-frame deletion within a putative two-component response regulator gene (atu5121). The second is a 111 base-pair symmetrical intergenic deletion on the circular chromosome that removes part of a short repeat sequence called CIR2 (12, 13).

The latter result prompted a broader search for short repeated palindromic sequences within the C58 genome, resulting in the identification of three classes of repeats (Figure 1). Two of these sequences, AgroCIR1 and AgroCIR2, were previously identified in a search for conserved motifs containing a binding site (GANTC) for the essential methylase CcrM (12, 13). The third element is herein designated AgroKE3 and bears no resemblance to the CIR repeats. A full KE3 repeat consists of 29 bp inverted repeats bracketing a variable region containing 49-76 bp (Figure 1). Like AgroCIR1 and AgroCIR2, KE3 elements are preferentially found on Chromosome 1, consistent with the evolution of these repeats on the ancestral chromosome during the radiation of the Rhizobiaceae prior to the origins of Chromosome II. Table S2 summarizes the distribution of KE3 repeats in several closely related, fully sequenced relatives of A. tumefaciens C58, including the recently sequenced Biovar I strain Agrobacterium H13-3 (14). Table S3 lists locations where these sequence repeats overlap a predicted open reading frame in the C58 genome. The biological function of the KE3 repeats has not yet been determined.

All true variant loci between C58UW and ATCC33970 were compared to the same loci in other A. tumefaciens C58 culture lines obtained from laboratories in the US and Europe (Tables S1 and S4). In 12 of 14 cases, including both indels, the
ATCC33970 sequence was identical to each of the C58 comparison strains, while in two cases all reference strains matched C58UW. While the cause of the additional variation in C58UW is unclear, it may be that the strain was passaged more frequently, or that one of the acquired variations resulted in a higher mutation rate.

The telomeres of the C58 linear chromosome are covalently-closed hairpin loops (4). This unusual structure meant that neither of our original studies was able to provide complete sequence for its ends; similarly, the telomeres have not yet been sequenced for H13-3 (14). Recently, however, the C58 telomere sequences, along with a biochemical characterization of the protelomerase enzyme that maintains them (TelA, encoded by \textit{atu2523}) have been published (15). The updated GenBank submission has been modified to include these data (see below).

We hypothesize that linearization of Chromosome II was a seminal event in the divergence of Biovar I strains, such as \textit{A. tumefaciens} C58, from Biovar III strains, such as \textit{Agrobacterium vitis} S4 (16). The simplest model for its linearization involves a single crossover between the ancestral circular Chromosome II and a linear phage or plasmid, thereby incorporating both telomeres and \textit{telA} into the genome in one event. Surprisingly, however, the \textit{telA} gene is located on the circular Chromosome I (4, 16). Comparison of the C58 and S4 genomes shows significant synteny between a single region of S4 Chromosome I and three regions of the C58 Chromosome. Our analysis of these relationships suggests that multiple recombination events in the \textit{atu2521-atu2523} (\textit{telA}) region transferred \textit{telA} to Chromosome I and initiated two large DNA transfers to Chromosome II (Figure 2). The breakpoint for the translocation of genes \textit{atu3507} through \textit{atu8188} (0.558-0.575 Mbp on Chromosome I of C58) from the circular to the linear...
chromosome is adjacent to telA. A similar translocation breakpoint occurs on Chromosome I immediately upstream of atu2521 and extends through atu4172 (lysC) into the adjacent rRNA loci (1.311-1.292 Mbp on C1 of C58). These genomic reorganizations transferred a rRNA operon and several essential genes to Chromosome II, while placing telA on Chromosome I. Intriguingly, atu2521 and atu2522 are more similar to orthologs in Rhizobium and Sinorhizobium, respectively, than they are to their orthologs in S4 (avi3961 and avi3963), suggesting that atu2521, atu2522 and telA may have entered the C58 genome together, perhaps as part of a linear plasmid.

We surveyed a large number of strains that have historically been classified as Agrobacterium, including Biovar I (A. tumefaciens), Biovar II (A. radiobacter) and Biovar III (A. vitis), for the presence of a linear mega-size DNA molecule by PFGE, and for telA and its adjacent ORF atu2522 (acvB) by PCR or Southern Blot (Table S5, Figure S1). It is important to note in considering this analysis that strong evidence supports the reclassification of Biovar II strains as Rhizobium (6, 10, 16-19). Our survey data indicate that linear chromosomes are unique to Biovar I strains (Table S5; (14, 20)). Based on this comparison, we can now define the unique genomic content of Biovar I as containing a linear replicon accompanied by a telA gene, in addition to other diagnostic genes (Table S6).

We have added the recently published telomere sequences and consolidated our two earlier versions of the C58 genome sequence into a single version with updated annotation from our own work and that of others. ATCC33970 was chosen as the standard sequence because it is most similar to other reference strains analyzed (Table S1). Notations are included in this update indicating the variations found in the C58UW
strain. The gene identifiers (locus tags) referring to genes kept from the original annotations are the same as those defined by Wood et al. (format atuXXXX) (5). Newly predicted protein-coding 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 by analyses subsequent to the initial genome deposit (21). Newly predicted small RNA genes (22) were designated with the locus tag pattern atu9xxx. Details of the reannotation are provided under supplementary materials and methods. The GenBank Accession numbers for the consolidated sequences are: Chromosome I, AE007869; Chromosome II, AE007870; pTiC58, AE007871; pAtC58, AE007872. These sequence files replace the two original versions of the A. tumefaciens C58 sequence files submitted by our research groups (4, 5).

Acknowledgements
This work was supported by National Science Foundation Grants 0333297, 0603491 and 0736671, by a grant from the M. J. Murdock Charitable Trust Life Sciences program (2004262:JVZ), by a science education grant from the Howard Hughes Medical Institute (52005125) and by the Monsanto Fund. We thank Kelly Williams for pointing out the group I intron within the tRNA gene. We wish to thank the hundreds of students at Hiram College, Seattle Pacific University, the University of Arizona, and the Mesa (Arizona) Biotechnology Academy who have contributed to the annotation of this genome.
Figure 1. Potential secondary structures for genomic repeat elements. (a) The structure emphasizes the inverted repeats present at each end of the CIR2 full element. (b) An alternative secondary structure for a CIR2 full element that emphasizes the larger overall inverted repeat. The extent of the stem forming above position 44 depends on the sequence of the spacer region. The shaded polygon indicate where a CcrM binding site sits if one is present (it is only found in about 1/3 of the elements). (c) Potential secondary structure for CIR1 elements. The structure emphasizes the larger overall inverted repeat. The shaded indicate the usual location of a CcrM binding site, but quite often a second CcrM binding site exists directly opposite the first location in a full element (the example here does not contain such a second site). (d) Potential secondary structure for KE3 half elements and ends of KE3 full elements.

Figure 2. Genomic transfers between Chromosome I and Chromosome II associated with the atu2521-atu2523 (telA) region. The top diagram represents genes on the A. vitis S4 Chromosome I from avi4581 (tRNA-met) through avi3984 (a conserved hypothetical gene) (16). The lower diagrams represent the syntenic regions of the A. tumefaciens C58 genome, with lines connecting orthologous genes of C58 and S4. Note the breakpoints in synteny immediately upstream of avi3961 and downstream of avi3963 in the S4 genome, identifying breakpoints for large genomic transfers between Chromosome I and Chromosome II in the C58 genome. Protein-coding genes are shown in gray, tRNA genes in black, and rRNA genes in white. The genomic locations of each region correspond to the base-pair numbering in the sequence files available at NCBI.


Agrobacterium species by *recA* allele analysis: *Agrobacterium recA* diversity.

Microbial ecology 60:862-872.


