Genome sequence of the plant pathogen *Ralstonia solanacearum*

M. Salanoubat*†, S. Genin†‡, F. Artiguenave*§, J. Gouzy‡, S. Mangenot*, M. Arlat‡, A. Billault‖, P. Brottier*, J. C. Camus‡, L. Cattolico*, M. Chandler¶, N. Choisne#, C. Claudel-Renard☆, S. Cunnac‡, N. Demange*, C. Gaspin, M. Lavie‡, A. Moisan**, C. Robert*, W. Saurin*§, T. Schiex**, P. Siguier¶, P. Thébault‡, M. Whalen‡§, P. Wincker*, M. Levy*, J. Weissenbach* & C. A. Boucher‡**

* *Genoscope and CNRS UMR-8030, 2 rue Gaston Crémieux, CP5706, 91057 Evry Cedex, France*
‡ *Laboratoire de Biologie Moléculaire des Interactions Plantes-Microorganismes INRA-CNRS, BP27, 31326 Castanet-Tolosan Cedex, France*
‖ *Fondation Jean Dausset-CEPH, 27 rue Juliette Dodu, 75010 Paris, France*
¶ *LMGM CNRS, 118 Route de Narbonne, F 31062 Toulouse Cedex, France*
# *Genoscope and INRA URGV, 2 rue Gaston Crémieux, CP5706, 91057 Evry Cedex, France*
☆ *Laboratoire de Genetique Cellulaire, INRA, BP27, 31326 Castanet-Tolosan Cedex, France*
** *Unité de Biométrie et Intelligence Artificielle INRA, BP27, F31326 Castanet-Tolosan Cedex, France*
† *These authors contributed equally to this work*

Ralstonia solanacearum **is a devastating, soil-borne plant pathogen with a global distribution and an unusually wide host range. It is a model system for the dissection of molecular determinants governing pathogenicity. We present here the complete genome sequence and its analysis of strain GMI1000. The 5.8-megabase (Mb) genome is organized into two replicons: a 3.7-Mb chromosome and a 2.1-Mb megaplasmid. Both replicons have a mosaic structure providing evidence for the acquisition of genes through horizontal gene transfer. Regions containing genetically mobile elements associated with the percentage of G+C bias may have an important function in genome evolution. The genome encodes many proteins potentially associated with a role in pathogenicity. In particular, many putative attachment factors were identified. The complete repertoire of type III secreted effector proteins can be studied. Over 40 candidates were identified. Comparison with other genomes suggests that bacterial plant pathogens and animal pathogens harbour distinct arrays of specialized type III-dependent effectors.**

Studies addressing the molecular determinants of bacterial pathogenicity towards plants have concentrated on a limited number of bacterial species, representing the taxonomic diversity of principal Gram-negative plant pathogens that produce the most common diseases. Among these, *R. solanacearum*[1,2] has unique and significant features. It is a soil-borne pathogen that naturally infects roots. It exhibits a strong and tissue-specific tropism within the host, specifically invading, and highly multiplying in, the xylem vessels. In addition, with over 200 host species belonging to more than 50 botanical families[3], this bacterium has an unusually wide host range, and offers a unique opportunity for the analysis of a strong and generalized array of virulence factors. *Ralstonia solanacearum* has been studied intensively both biochemically and genetically, and has long been recognized as a model system for the analysis of pathogenicity[4]. It is well adapted to life in soil in the absence of host plants[5], thereby providing a good system to investigate functions governing adaptation to such an ecological niche. *Ralstonia solanacearum* is a β-proteobacterium and thus belongs to a group of bacteria whose genomic organization is still poorly characterized. The only other representative of this group for which the complete genome sequence has been determined and annotated is *Neisseria meningitidis*[6,7].

To date, the genome of only one plant pathogenic bacterium, *Xylella fastidiosa*, has been sequenced completely[8]. Here we report the complete nucleotide sequence of the genome of *R. solanacearum* strain GMI1000, a race 1 strain isolated from tomato. Notably, GMI1000 is pathogenic on the model plant *Arabidopsis thaliana*, the genome of which has been entirely sequenced[9], thereby facilitating studies of host response. We provide an integrative analysis of the predicted functions encoded in this organism with special emphasis on pathogenicity. Our analysis of the genome sequence provides clues to the evolution of pathogenicity functions. The genome sequence of *R. solanacearum* is a first step towards an exhaustive functional analysis of pathogenicity determinants in this pathogen.

A bipartite genome structure

After sequencing by the whole-genome random sequencing method, we assembled the *R. solanacearum* genome into two circular molecules: a large replicon of 3,716,413 bp and a smaller 2,094,509-bp replicon, yielding a total genome size of 5,810,922 bp. The two molecules have an almost identical G+C content (67.04% and 66.86% for the large and small replicon, respectively). These two molecules correspond to the two bands that are visualized on pulsed-field gel electrophoresis (see Supplementary Information Fig. 1). The genome encodes a total of 5,129 predicted proteins, and the general features are shown in Table 1.

Analysis of the sequence establishes that the smaller replicon corresponds to the previously described megaplasmid, as it encodes *hrp* genes that were previously localized on this replicon[10]. Such a bipartite genome structure seems to be a characteristic of *R. solanacearum*, as a megaplasmid has been detected in most of the

Table 1 General features of the *R. solanacearum* strain GMI1000 genome

Genome feature	Chromosome	Megaplasmid	Genome
Length (bp)	3,716,413	2,094,509	5,810,922
G+C ratio	67.04%	66.86%	66.97%
Protein-coding regions	87.8%	86.5%	87.3%
tRNAs	55	3	58
Ribosomal RNA operons	3	1	4
Protein-coding genes	3,448	1,681	5,129
Average length of protein-coding genes (bp)	946	1,077	989
Genes with functional assignment	1,609	652	2,261
Orphan genes*	12.6%	18.7%	14.6%
Regulatory genes*	7.2%	9.6%	8.0%
Insertion sequences and phage sequences†	3.4%	2.5%	3.1%

* Percentage of total protein-coding genes.
† Percentage of replicon size.

§ Present addresses: Genomining, 93–95 rue Henri Rochefort, 91000 Evry, France (F.A., W.S.); Department of Biology, San Francisco State University, 1600 Holloway Avenue, San Francisco, California 94132, USA (M.W.).

strains belonging to this species[11]. As no derivative of strain GMI1000 in which the megaplasmid is deleted has been obtained, the status of this replicon as a plasmid is still an open question. The distribution of genes between the two replicons provides some insights into this question.

The origin of replication of the large replicon, as identified by G+C skew analysis[12], has features that are characteristic of a chromosomal origin of replication in bacteria (see Supplementary Information Fig. 2). It is flanked by the *rnpA* gene on one side, and the *dnaA*, *dnaN*, *gyrAB* genes on the other. It also harbours a single consensus DnaA-binding box (TTATCCACAAA), the first nucleotide of which was arbitrarily chosen as the origin for the nucleotide numbering of the large replicon. Furthermore, the large replicon encodes a complete set of essential housekeeping genes including genes required for: (1) DNA replication, DNA repair and cell division; (2) transcription; and (3) translation. This last set includes all the ribosomal proteins, 3 complete ribosomal DNA loci, and 55 identified transfer RNAs, allowing recognition of all possible codons. Finally, all the essential genes required for purine and pyrimidine biosynthesis are located on the large replicon. Therefore, this replicon encodes all of the basic mechanisms required for the survival of the bacterium, and is referred to here as the 'chromosome'. Accordingly, the smaller replicon may be a dispensable genetic element. The putative origin of replication of this replicon, as predicted by G+C skew analysis, has characteristics of plasmid-borne *ori* loci. It is flanked by the *repA* gene and by at least 14 repetitions of a conserved motif (consensus G/CCGTACCCG/ATTTCTGCG) that may be RepA-binding boxes. Therefore, the smaller replicon appears to be a megaplasmid. The first nucleotide of the most upstream RepA-binding box located in the intergenic region preceding *repA* has been arbitrarily chosen as the origin for nucleotide numbering of the megaplasmid.

The megaplasmid carries several metabolically essential genes that are also present on the chromosome. These include a complete copy of a rDNA locus with 2 tRNA genes, a gene coding for the α-subunit of DNA polymerase III, and a gene for the protein elongation factor G. Moreover, several enzymes controlling primary metabolism, including amino acid and cofactor biosynthesis, are encoded on the megaplasmid with no counterpart on the chromosome. As a consequence, we predict that a megaplasmid-deleted derivative of strain GMI1000 will be auxotrophic for several metabolites, a status similar to that reported for certain internal deletion mutants of the megaplasmid[13].

Analysis of the genes present on the megaplasmid suggests that this replicon has a significant function in overall fitness and adaptation of the bacterium to various environmental conditions. As mentioned previously, the megaplasmid carries all of the *hrp* genes that are required to cause disease on plants, a trait that allows the bacterium to colonize a rather exclusive ecological niche. The megaplasmid also encodes the constituents of the flagellum and most of the genes governing exopolysaccharide synthesis. In addition, this replicon carries 315 out of the 748 genes of unknown function, a proportion significantly biased in favour of the megaplasmic ($P = 4 \times 10^{-7}$). On the other hand, a significant bias in favour of the chromosome is observed for the *R. solanacearum* genes that are shared with other bacteria (see Supplementary Information Fig. 3).

Mosaic structure of the genome

The gene prediction software FrameD (http://www.toulouse.inra.fr/FrameD.html) using the probabilistic model constructed on previously characterized *R. solanacearum* genes, led to a clear-cut prediction of genes in more than 90% of the genome. Using this matrix, a significant portion of the genome (7%) was predicted as non-coding (in regions spanning over 1 kb) although in some instances, BLASTX analysis revealed significant similarities in these regions with known proteins. On the basis of these homologies, an alternative matrix was constructed and used for gene prediction in such regions that we designated alternative codon-usage regions (ACURs). When analysed for base composition, most ACURs, but not all, differ significantly from the average 67% G+C content found for the entire genome, with variations ranging from 50% to 70% G+C content. In addition, codon usage in these regions differs significantly from codon usage in the rest of the genome (Supplementary Information Table 1 and Fig. 4). Furthermore, ACURs were often associated with mobile genetic elements. In 44 out of 93 ACURs, a prophage, insertion sequence or part of an insertion sequence occurred either encoded directly within the ACUR or within the 1-kb flanking region. The strongly biased distribution of genetically mobile elements with ACURs ($P = 1.7 \times 10^{-5}$) suggests that ACURs may have been acquired through horizontal gene transfer, consistent with the propensity of *R. solanacearum* to take up and recombine exogenous DNA through natural transformation[14].

In addition to ACURs, strain GMI1000 harbours several other elements that may have a function in genetic instability and rapid evolution of the genome (Fig. 1). Throughout the genome, there are at least 118 copies of complete or truncated insertion sequences representing 17 distinct elements belonging to 7 families. Three of these insertion sequences have been identified previously in *R. solanacearum* (ISRso1, IS*1421* and IS*1021*) (http://www-IS.biotoul.fr) and we have called the remaining 14 ISRso5–ISRso18. There is a preponderance of IS*3* (22 copies) and IS*5* (27 copies) family members. In addition, there are two, presumably non-autonomous, derivatives of ISRso1, composed uniquely of the terminal sequences separated by an identical 117-bp DNA segment. This is similar to the RUP elements observed in *Streptococcus pneumoniae*[15]. At least 4 possibly defective prophages were found on the chromosome together with a conjugative transposon located between positions 2,781,738–2,825,808 (Fig. 1). This transposon is related to the 55-kb transposon Tn*4371* from *Ralstonia metallidurans*[16]. The *R. solanacearum* conjugative transposon includes a set of *tra* and *trb* genes for conjugation as well as an integrase (RS00926); however, the genes coding for biphenyl resistance in Tn*4371* are replaced by a group of genes with undefined functions. There are several loci encoding Rhs and Vgr-related elements[17,18], found to be recombinational hotspots in *Escherichia coli*. Of note, ACURs and mobile genetic elements are not distributed evenly on the genome but are often clustered on both replicons (Figs 1 and 2).

The *R. solanacearum* sequence has a mosaic structure containing numerous elements signalling the potential for evolution. Genomic rearrangements have already been reported to occur naturally in this bacterium[13,19] and are further exemplified by the almost perfect tandem duplication of a 31-kb stretch of DNA on the megaplasmid (positions 1,648,630–1,710,832). The present genome sequence may represent a single snapshot of a structure that is variable from isolate to isolate and within derivatives from the same isolate.

Candidate genes responsible for pathogenesis

Apart from the virulence genes already described in *R. solanacearum*, a series of new genes putatively involved in pathogenicity were identified (Table 2; see Supplementary Information Table 2 for a complete list). These include genes coding for additional hydrolytic enzymes involved in the degradation of plant cell walls, and genes required for the production of the plant hormones auxin and ethylene or for the degradation of the plant signalling molecules ethylene and salicylic acid—a mediator of the plant systemic acquired resistance. Ten genes were predicted to code for proteins involved in resistance to oxidative stress. These 10 genes may be involved in the detoxification of the active oxygen species produced by infected plants, molecules reportedly representing a first line of defence against pathogen invasion[20]. Genes involved in the production of toxins or antibiotics were also identified: these include six

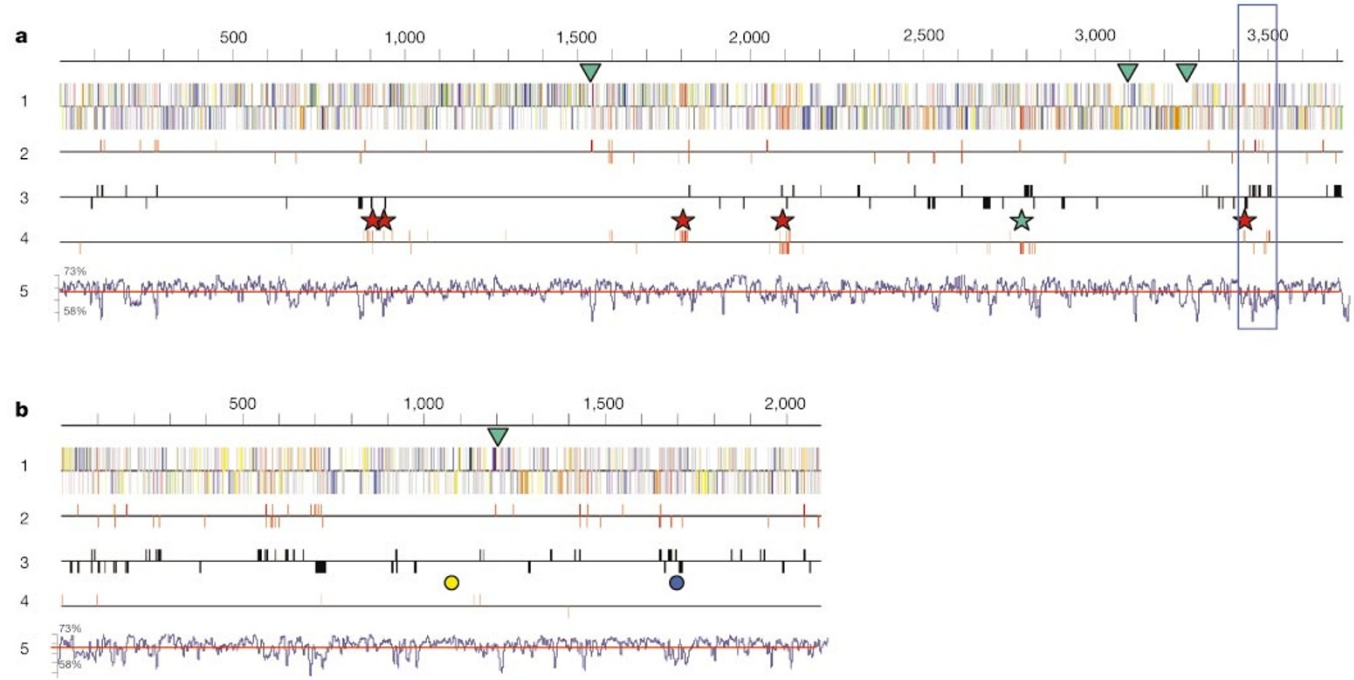

Figure 1 General organization of the *R. solanacearum* strain GMI1000 genome. The two circular replicons are represented linearly starting from base number 1 (**a**, chromosome; **b**, megaplasmid). Kilobases are indicated along the top. For each replicon the distribution of protein-coding genes (line 1), insertion sequences (line 2), ORFs from other genetically mobile elements (line 4) and ACURs (line 3) are represented. The percentage of G+C variation (from the average; red line) along the length of the genome using a 2,500-bp window is shown (line 5). Colours on line 1 correspond to the main classes of genes (the colour code is available at http://sequence.toulouse.inra.fr/R.solanacearum.html). Green triangles point towards the positions of rDNA loci; red stars locate the major bacteriophage remnants; the green star corresponds to the position of the conjugative transposon. The yellow circle indicates the position of the *hrp* locus and the blue circle indicates the tandem duplication of a 31-kb region. The blue box in **a** delineates the region of the genome enlarged in Fig. 2.

haemolysin-like genes belonging to the RTX toxin family[21] and several peptide or polyketide synthase genes. In particular, the two largest open reading frames (ORFs) in the genome (RS05859 and RS05860, coding for 5,953 and 6,889 amino-acid products, respectively) are highly related to the syringomycin synthase gene, which is required for the production of a *Pseudomonas syringae* toxin[22].

An unusual characteristic of the *R. solanacearum* genome is that it contains a large number of genes coding for outer-membrane proteins or components of bacterial appendages (pili, fimbriae) implicated in the attachment of the bacterium to external surfaces. We found at least 35 genes, distributed in 5 gene clusters, involved in the biogenesis of a type IV pilus. Type IV pili are known from several other bacterial systems to be adhesion factors, and are responsible for movement of bacteria over epithelial surfaces without the use of flagella[23]. Furthermore, two other gene clusters are predicted to encode an unusual type of pilus structure that mediates a

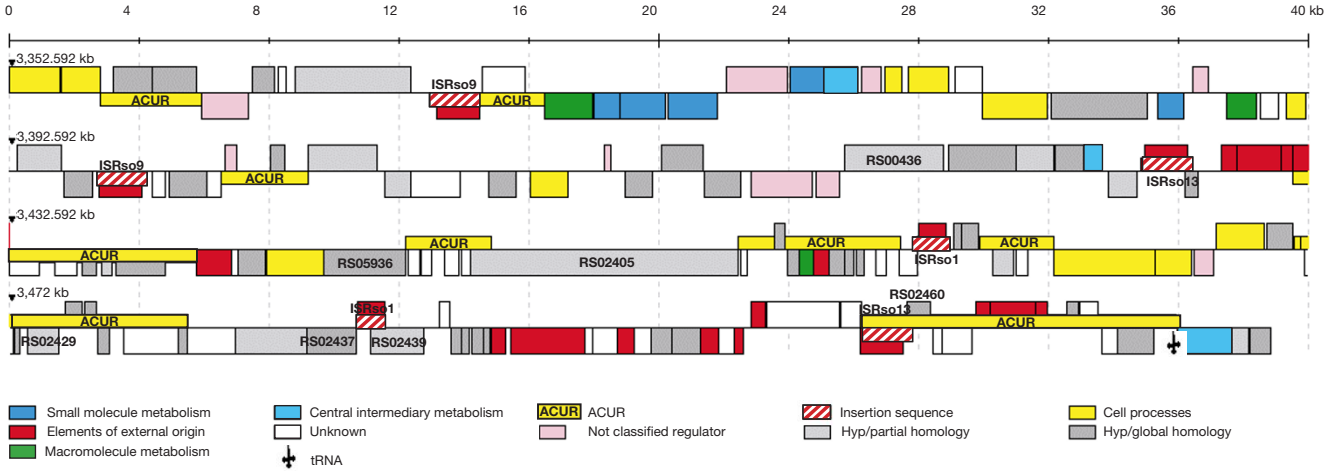

Figure 2 Enlarged view of part of the chromosome. This figure reveals the mosaic structure of the genome. ACURs (narrow yellow rectangles) alternate with genes from mobile genetic elements (large red rectangles with insertion sequences shown in striped white and red rectangles) and with genes that fit the standard Markov model for *R. solanacearum* genes. An Ala-tRNA gene is at the end of the last ACUR. This region contains genes encoding two candidate type-III secreted effectors (RS02429 and RS02460), three haemagglutinin-related proteins (RS00436, RS02405 and RS5936) and a Vgr-related protein (RS02437 and RS02439) inactivated by insertion of ISRso1. Colour codes for other genes are given at the bottom of the figure and correspond to the main classes defined in the nomenclature of ref. 46.

Table 2 Known and candidate genes responsible for pathogenesis in the genome of *R. solanacearum* strain GMI1000

Pathogenicity genes	No. of genes
Known	
Type III secretion system and secreted effectors	31
Global regulatory functions	11
Exopolysaccharide biosynthesis	18
Hydrolytic enzymes	4
Hormone production	1
Candidate	
Type III secretion-dependent effectors	51
Type III effectors based on homology	25
Type III effectors based on structural features	26
Adhesion/surface proteins	93
Haemagglutinin-related proteins	27
Type IV fimbrial biogenesis proteins	35
Other pili/fimbriae	24
Hydrolytic enzymes/host cell wall degradation	5
Toxins	13
Resistance to oxidative stress	10
Plant hormones and signalling molecules	7
Others	16

tight adherence to surfaces similar to those reported recently in *Caulobacter crescentus*[24] and *Actinobacillus actinomycetemcomitans*[25]. For all of these pili/fimbriae-coding systems, we found multiple copies of the pilin structural genes. This raises the possibility that these genes are expressed in different contexts or may have slightly different structural roles, thereby broadening the adaptative ability of this wide-host-range pathogen to interact with diverse environmental substrates, including host epidermal surfaces.

Another peculiarity concerning the abundance of adhesion/attachment functions in *R. solanacearum* is exemplified by a class of surface molecules encoded by long ORFs (9 frames with a coding potential greater than 2,500 amino acids). These translated products share homology with proteins that are adhesins in other bacterial pathogens, the filamentous haemagglutinin (FhaB) of *Bordetella pertussis* and the HMW1A/HMW2A adhesins of *Haemophilus influenzae*[26]. In total, there are 14 probable haemagglutinin-type proteins, and 13 additional ORFs coding for proteins containing variable internal repeats that are structurally related to filamentous haemagglutinins. *Ralstonia solanacearum* therefore has the greatest number of these haemagglutinin-related proteins of all the completed bacterial genomes.

We have also identified a family of related proteins presenting some degree of similarity to the *Agrobacterium tumefaciens* proteins AttM and AttZ, both of which are required for attachment to plant cells and for virulence[27]. The *R. solanacearum* genome is thus rich in attachment factors, perhaps functioning as determinants for a wide host range.

Effectors dependent on a type III secretion system

Ralstonia solanacearum possesses a cluster of *hrp* genes encoding a type III secretion system (TTSS) that is essential for pathogenicity. Bacterial TTSSs are conserved among both plant and animal pathogens and translocate effector proteins into the cytoplasm of the host cell[4,28]. Identification of translocated effectors and establishment of their mode of action after delivery into host cells are two chief challenges, the solution of which has high potential for the conception of new therapeutic strategies.

A principal outcome of the analysis of the *R. solanacearum* genome is the discovery of multiple genes related to the avirulence (*avr*) genes, described in several plant pathogenic bacteria. These *avr* genes encode effector proteins presumably injected into host cells through the *hrp*-encoded TTSS[28]. The 14 ORFs coding for products related to Avr determinants (8 with global homology and 6 with homology restricted to a specific domain) are distributed on both replicons. This finding is notable because to date *avr* genes have been found by functional assays only in phytopathogenic bacteria with a limited host range (often being confined to members of a single plant species or genus), the host range being molecularly governed by *avr* genes[29]. Furthermore, the presence of Avr-related proteins in *R. solanacearum* is surprising as no report of *avr*-dependent monogenic resistance of solanaceous crops towards bacterial wilt has been reported. It is probable that these Avr proteins must confer selective advantages, acting collectively as pathogenicity factors on a large set of host plants[30]. It is also possible that *R. solanacearum* possesses some general suppressor(s) of the plant defence response triggered by the recognition of Avr proteins by plant factors.

A second set of 9 proteins dispersed throughout the genome and potentially transiting through the TTSS was identified by homology to determinants in other plant pathogenic bacteria, located in the immediate regions flanking *hrp* loci. The *hrp* locus of *Pseudomonas syringae*, for example, is part of a pathogenicity island (PAI)[31] that harbours several Hrp-dependent effector proteins[32]. A PIP-box consensus motif (TTCGC-N15-TTCGC)[33], suggestive of *hrp*-dependent regulation, is present in the promoter sequence of 6 *avr* gene homologues (RS02644, RS04524, RS05218, RS05356, RS05373 and RS05468).

As the TTSS-dependent pathogenicity factors are thought to be injected directly into eukaryotic cells to exert their anti-host function, we looked for proteins exhibiting typical eukaryotic features and functions. As a result, we identified potential functional homologues of essential pathogenicity effectors found in bacterial pathogens of animals, such as protein kinases (RS01445 and RS05210) and a probable tyrosine phosphatase (RS03075). Another ORF product (RS04706) contains 3.5 internal repeats structurally related to the PPR motif[34]—a motif prevalent in plant organellar proteins, but never found to date in prokaryotic proteins. Three gene families encode proteins with ankyrin repeat domains[35], a pirin-like domain[36] and leucine-rich repeats (LRRs)[37]. Although these classes of proteins are not strictly restricted to eukaryotes, they appear to be implicated in eukaryotic signalling pathways, presumably through the properties of these specific domains to establish protein–protein interactions. Several LRR-containing proteins have been shown to transit through TTSSs in pathogens such as *Yersinia*, *Shigella* or *Salmonella* sp., and recently in *R. solanacearum*[28,38]. We found a total of 10 genes coding for 3 families of LRR-containing proteins scattered in the genome, drawing attention to the probable functional redundancy of these candidate effectors of pathogenicity. Considering the above observations, we estimate that the number of genes encoding potential TTSS-dependent effectors is 40 or higher. This number is greater than that (25) estimated for the bacterial agent of dysentery *Shigella flexneri*[39]. Genome-scale studies carried out on other plant-pathogenic bacteria will soon reveal whether this high number of effectors is correlated with the wide host range of *R. solanacearum*.

Evolution of virulence

Although most of the candidate genes encoding TTSS-dependent effectors are not found near the *hrp* gene cluster and are distributed evenly on both replicons, approximately half of them appear to reside within clusters of ACURs. Five of these ACUR clusters have the typical features of PAIs such as the presence of DNA sequences indicative of gene mobility (insertion sequences, transposases) or recombination events (Rhs and Vgr elements), and, in some cases, the association of these regions with tRNA or prophage sequences[31] (Fig. 2). Moreover, 15 of the approximately 40 candidate TTSS-dependent effectors, identified on the basis of homologies or structural features, have a significantly different G+C content than the mean of the *R. solanacearum* genome (see Supplementary Information Table 2). This suggests that they may have been acquired through horizontal gene transfer. In addition, some observations suggest that ACURs could also contribute to the

evolution of new virulence genes. For example, there are three genes related to the host specificity determinant *pthG* from *Erwinia herbicola*[40] (RS04524, RS05166 and RS05218), located in three different ACURs. These gene products all have a common PthG amino terminus (54% identical residues over the 35 N-terminal amino acids) but with a variable carboxy terminus. Gene duplication and evolution may possibly account for the generation of new virulence specificities in the C termini while maintaining appropriate expression and targeting signals in the N termini. A similar mechanism of duplication followed by differential evolution of domains can also be observed with members of the haemagglutinin-related protein family (such as RS02101 and RS02405) located within or at the border of ACURs. In addition, the deletion of genes could also serve as a means of bacterial adaptation. Evolutionary models of interactions between plants and plant-pathogenic bacteria state that the emergence of a plant resistance gene that recognizes a virulence gene would abolish the value of the corresponding virulence gene[41]. Possible evidence for this hypothesis may be found in the occurrence of single or multiple frameshifts or insertion sequence insertions in putative virulence gene homologues (RS00660, RS03919 and RS02460). These deactivating mechanisms may be the result of an evolutionary elimination of genes that have become liabilities for the pathogen.

A close investigation of each of the 30-kb regions flanking the *hrp* locus did not reveal significant changes in the G+C content nor the presence of DNA mobility associated elements typically observed in ACURs. This indicates that, contrary to the *P. syringae hrp* gene cluster and flanking regions[32], the *R. solanacearum hrp* locus and flanking regions containing virulence genes is not, in the strictest sense, a PAI. Instead, our analysis suggests that this region is composed of a core group of ancestral pathogenicity genes that have been subjected to long coevolution with the *R. solanacearum* genome. Thus, the evolutionary status of the *hrp* region is in sharp contrast with the set of candidate effector-encoding genes that are scattered in the genome and often associated with ACURs. This leads us to speculate that *R. solanacearum*, along with an ancestral TTSS and associated effectors, acquired new effector genes through horizontal transfer, thereby remaining a successful pathogen.

Comparative genomics

A comparison of the *R. solanacearum* proteome with the proteome of bacteria that have been sequenced entirely reveals a close similarity of *R. solanacearum* with two other large-genome bacteria (Supplementary Information Figs 5 and 6), *P. aeruginosa*[42] and *Sinorhizobium melitoti*[43]. Like *R. solanacearum*, these two species interact with eukaryotic hosts and can be free living in soil. In particular, the proteome of *R. solanacearum* shares many common features with that of the opportunistic pathogen *P. aeruginosa*: (1) a diversity of nutritional pathways; (2) many membrane transport systems; (3) complex chemosensory systems; and (4) a large number of regulatory genes, with a high proportion of two-component regulatory system proteins (37 sensors, 55 response regulators and 6 sensor-response regulator hybrids in *R. solanacearum*). These features are consistent with the evolution of adaptative responses to changes in environment, permitting the bacteria to thrive in diverse ecological niches.

We used a similar comparative analysis on *R. solanacearum* candidate pathogenicity determinants (Supplementary Information Table 3). Only homologies restricted to degradative enzymes, resistance to oxidative stress, haemolysins, peptide synthases and certain attachment structures were found with the plant pathogen *X. fastidiosa*[8]. The same functions also seem to be widely conserved among other bacterial proteomes, contrarily to most of the predicted TTSS-dependent effectors. Several TTSS-dependent effectors appear to be conserved in the plant-pathogenic bacteria, as demonstrated by conservation of 13 known effector proteins. However, with two exceptions (AvrRxv/YopP-related proteins[4] and possibly RS02872), these effectors are not found in bacterial pathogens of animals. Significantly, most of these TTSS-dependent effectors are not found in the genome of *Ralstonia metallidurans*, a non-pathogenic bacterium that is taxonomically the closest related species to *R. solanacearum*. Our results imply that bacterial pathogens of plants and animals, despite extremely high conservation of the apparatus used to secrete effectors, harbour distinct arrays of specialized effectors. According to this hypothesis, TTSS-dependent effectors would have different cellular targets or effects on the respective hosts, resulting in the differential cellular processes observed during infection of animal and plant cells. □

Methods

Sequencing and assembly

We generated about 80,000 sequences from both ends of genomic clones ranging from 1.5 to 100 kb. The steps to obtain the final sequence (shotgun assembly, gap closure and polishing) are described in the Supplementary Information. The validity of the sequence was assessed by comparing the restriction enzyme pattern deduced from the sequence to the experimentally observed restriction pattern obtained by digestion of clone DNA of a minimal tilling path composed mainly of bacterial artificial chromosome (BAC) clones with 6-base recognition enzymes. A total of 99.25% of the sequences were validated with at least two different restriction enzymes, 0.49% with only one enzyme, and 0.26% of the sequence was not validated, potentially owing to local mis-assemblies (the largest region is less than 5.1 kb). The location of these regions can be found in the Supplementary Information.

Gene prediction and annotation

Sequence analysis and annotation were performed using iANT (integrated Annotation Tool)[44] as described for *S. meliloti*[45], except that the probabilistic Markov model for coding regions used by the gene prediction software FrameD was constructed on 77 *R. solanacearum* gene sequences obtained from public databanks. The alternative matrix was built using genes first identified in ACURs based on homology, as revealed by BLASTX analysis. Predicted ORFs were reviewed individually by gene annotators for start codon assignment. Output of Prosite search and BLASTP analysis on the corresponding products were also individually expertized to generate the proposed annotations. Proteins were classified according to Riley's rules[46]. The complete annotated genetic map, search tools (SRS, BLAST), annotation and process classification are available at http://sequence.toulouse.inra.fr/R.solanacearum.html.

Statistical analysis

We performed statistical analysis using Fisher's exact test[47].

Received 10 July; accepted 30 November 2001.

1. Smith, E. F. A bacterial disease of tomato, pepper, eggplant and Irish potato (*Bacillus solanacearum* nov. sp.). *US Dep. Agric. Div. Vegetable Physiol. Pathol. Bull.* **12,** 1–28 (1896).
2. Yabuuchi, E., Kosako, Y., Yano, I., Hotta, H. & Nishiuchi, Y. Transfer of two *Burkholderia* and an *Alcaligenes* species to *Ralstonia* gen. nov.: proposal of *Ralstonia pickettii* (Ralston, Palleroni and Doudoroff 1973) comb. nov., *Ralstonia solanacearum* (Smith 1896) comb. nov. and *Ralstonia eutropha* (Davis 1969) comb. nov. *Microbiol. Immunol.* **39,** 897–904 (1995).
3. Hayward, A. C. in *Encyclopedia of Microbiology* Vol. 4 (ed. Lederberg, J.) 32–42 (Academic, San Diego, 2000).
4. Staskawicz, B. J., Mudgett, M. B., Dangl, J. L. & Galan, J. E. Common and contrasting themes of plant and animal diseases. *Science* **292,** 2285–2289 (2001).
5. Granada, G. A. & Sequeira, L. Survival of *Pseudomonas solanacearum* in soil, rhizosphere and plant roots. *Can. J. Microbiol.* **29,** 433–440 (1983).
6. Parkhill, J. *et al.* Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature* **404,** 502–505 (2000).
7. Tettelin, H. *et al.* Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* **287,** 1809–1815 (2000).
8. Simpson, A. J. *et al.* The genome sequence of the plant pathogen *Xylella fastidiosa*. *Nature* **406,** 151–157 (2000).
9. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815 (2000).
10. Boucher, C., Martinel, A., Barberis, P., Alloing, G. & Zischek, C. Virulence genes are carried by a megaplasmid of the plant pathogen *Pseudomonas solanacearum*. *Mol. Gen. Genet.* **205,** 270–275 (1986).
11. Rosenberg, C., Casse-Delbart, F., Dusha, I., David, M. & Boucher, C. Megaplasmids in the plant associated bacteria *Rhizobium meliloti* and *Pseudomonas solanacearum*. *J. Bacteriol.* **150,** 402–406 (1982).
12. Lobry, J. R. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* **13,** 660–665 (1996).
13. Boucher, C., Barberis, P. & Arlat, M. Acridine orange selects for deletion of *hrp* genes in all races of *Pseudomonas solanacearum*. *Mol. Plant-Microbe Interact.* **1,** 282–288 (1988).
14. Bertolla, F., Van Gijsegem, F., Nesme, X. & Simonet, P. Conditions for natural transformation of *Ralstonia solanacearum*. *Appl. Environ. Microbiol.* **63,** 4965–4968 (1997).
15. Oggioni, M. R. & Claverys, J. P. Repeated extragenic sequences in prokaryotic genomes: a proposal for the origin and dynamics of the RUP element in *Streptococcus pneumoniae*. *Microbiology* **145,** 2647–2653 (1999).

16. Merlin, C., Springael, D., Mergeay, M. & Toussaint, A. Organisation of the *bph* gene cluster of transposon Tn4371, encoding enzymes for the degradation of biphenyl and 4-chlorobiphenyl compounds. *Mol. Gen. Genet.* **253,** 499–506 (1997).
17. Wang, Y. D., Zhao, S. & Hill, C. W. Rhs elements comprise three subfamilies which diverged prior to acquisition by *Escherichia coli*. *J. Bacteriol.* **180,** 4102–4110 (1998).
18. Wilderman, P. J., Vasil, A. I., Johnson, Z. & Vasil, M. L. Genetic and biochemical analyses of eukaryotic-like phospholipase D of *Pseudomonas aeruginosa* suggest horizontal acquisition and a role for persistence in a chronic pulmonary infection model. *Mol. Microbiol.* **39,** 291–303 (2001).
19. Brumbley, S. M., Carney, B. F. & Denny, T. P. Phenotype conversion in *Pseudomonas solanacearum* due to spontaneous inactivation of PhcA, a putative LysR transcriptional regulator. *J. Bacteriol.* **175,** 5477–5487 (1993).
20. Alvarez, M. E. *et al.* Reactive oxygen intermediates mediate a systemic signal network in the establishment of plant immunity. *Cell* **92,** 773–784 (1998).
21. Lally, E. T., Blake Hill, R., Kieba, I. R. & Korostoff, J. The interaction between RTX toxins and target cells. *Trends Microbiol.* **7,** 356–361 (1999).
22. Bender, C. L., Alarcon-Chaidez, F. & Gross, D. C. *Pseudomonas syringae* phytotoxins: mode of action, regulation, biosynthesis by peptide and polyketide synthases. *Microbiol. Mol. Biol. Rev.* **63,** 266–292 (1999).
23. Wall, D. & Kaiser, D. Type IV pili and cell motility. *Mol. Microbiol.* **32,** 1–10 (1999).
24. Skerker, J. M. & Shapiro, L. Identification and cell cycle control of a novel pilus system in *Caulobacter crescentus*. *EMBO J.* **19,** 3223–3234 (2000).
25. Kachlany, S. C. *et al. flp-1*, the first representative of a new pilin gene subfamily, is required for non-specific adherence of *Actinobacillus actinomycetemcomitans*. *Mol. Microbiol.* **40,** 542–554 (2001).
26. Jacob-Dubuisson, F., Locht, C. & Antoine, R. Two-partner secretion in Gram-negative bacteria: a thrifty, specific pathway for large virulence proteins. *Mol. Microbiol.* **40,** 306–313 (2001).
27. Matthysse, A. G., Yarnall, H., Boles, S. B. & McMahan, S. A region of the *Agrobacterium tumefaciens* chromosome containing genes required for virulence and attachment to host cells. *Biochim. Biophys. Acta.* **1490,** 208–212 (2000).
28. Cornelis, G. R. & Van Gijsegem, F. Assembly and function of type III secretory systems. *Annu. Rev. Microbiol.* **54,** 735–774 (2000).
29. Leach, J. E. & White, F. F. Bacterial avirulence genes. *Annu. Rev. Phytopathol.* **34,** 153–179 (1996).
30. Kjemtrup, S., Nimchuk, Z. & Dangl, J. L. Effector proteins of phytopathogenic bacteria: bifunctional signals in virulence and host recognition. *Curr. Opin. Microbiol.* **3,** 73–78 (2000).
31. Hacker, J. & Kaper, J. B. Pathogenicity islands and the evolution of microbes. *Annu. Rev. Microbiol.* **54,** 641–679 (2000).
32. Alfano, J. R. *et al.* The *Pseudomonas syringae* Hrp pathogenicity island has a tripartite mosaic structure composed of a cluster of type III secretion genes bounded by exchangeable effector and conserved effector loci that contribute to parasitic fitness and pathogenicity in plants. *Proc. Natl Acad. Sci. USA* **97,** 4856–4861 (2000).
33. Fenselau, S. & Bonas, U. Sequence and expression analysis of the *hrpB* pathogenicity operon of *Xanthomonas compestris* pv. *vesicatoria* which encodes eight proteins with similarity to components of the Hrp, Ysc, Spa, and Fli secretion systems. *Mol. Plant-Microbe Interact.* **8,** 845–854 (1995).
34. Small, I. D. & Peters, N. The PPR motif—a TPR-related motif prevalent in plant organellar proteins. *Trends Biochem. Sci.* **25,** 46–47 (2000).
35. Bork, P. Hundreds of ankyrin-like repeats in functionally diverse proteins: mobile modules that cross phyla horizontally? *Proteins* **17,** 363–374 (1993).
36. Wendler, W. M., Kremmer, E., Forster, R. & Winnacker, E. L. Identification of pirin, a novel highly conserved nuclear protein. *J. Biol. Chem.* **272,** 8482–8489 (1997).
37. Kajava, A. V. Structural diversity of leucine-rich repeat proteins. *J. Mol. Biol.* **277,** 519–527 (1998).
38. Guéneron, M., Timmers, A. C., Boucher, C. & Arlat, M. Two novel proteins, PopB, which has functional nuclear localization signals, and PopC, which has a large leucine-rich repeat domain, are secreted through the Hrp-secretion apparatus of *Ralstonia solanacearum*. *Mol. Microbiol.* **36,** 261–277 (2000).
39. Buchreiser, C. *et al.* The virulence plasmid pWR100 and the repertoire of proteins secreted by the type III secretion apparatus of *Shigella flexneri*. *Mol. Microbiol.* **38,** 760–771 (2000).
40. Ezra, D., Barash, I., Valinsky, L. & Manulis, S. The dual function in virulence and host range restriction of a gene isolated from the pPATH (Ehg) plasmid of *Erwinia herbicola* pv. *gypsophilae*. *Mol. Plant-Microbe Interact.* **13,** 683–692 (2000).
41. Alfano, J. R. & Collmer, A. Bacterial pathogens in plants: life against the wall. *Plant Cell* **8,** 1683–1698 (1996).
42. Stover, C. K. *et al.* Complete genome sequence of *Pseudomonas aeruginosa* PAOI, an opportunistic pathogen. *Nature* **406,** 959–964 (2000).
43. Galibert, F. *et al.* The composite genome of the legume symbiont *Sinorhizobium meliloti*. *Science* **293,** 668–672 (2001).
44. Thébault, P., Servant, F., Schiex, T., Kahn, D. & Gouzy, J. in *JOBIM Conf. Proc.* 361–365 (ENSA and LIRM Editor, Montpellier, 2000).
45. Capela, D. *et al.* Analysis of the chromosome sequence of the legume symbiont *Sinorhizobium meliloti* strain 1021. *Proc. Natl Acad. Sci. USA* **98,** 9877–9882 (2001).
46. Karp, P. D. *et al.* EcoCyc: Encyclopedia for *Escherichia coli* genes and metabolism. *Nucleic Acids Res.* **27,** 55–58 (1999).
47. Fischer, R. A. On the interpretation of chi-square from contingency tables, and the calculation of P. *J. R. Stat. Soc.* **85,** 87–94 (1922).

Supplementary Information accompanies the paper on *Nature*'s website (http://www.nature.com).

Acknowledgements

We thank N. Aiache and C. Cruaud for technical assistance. The Laboratoire de Biologie Moleculaire des Interactions Plantes Microorganismes is supported by INRA, CNRS and Toulouse Genopole. M. Whalen was supported by the National Institute for Health.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.A.B. (e-mail: boucher@toulouse.inra.fr). The sequences for the chromosome and megaplasmid are deposited in the EMBL database under accession numbers AL646052 and AL646053, respectively.