JARQ 39 (4), 275 – 287 (2005) http://www.jircas.affrc.go.jp

Genome Sequence of *Xanthomonas oryzae* pv. *oryzae* Suggests Contribution of Large Numbers of Effector Genes and Insertion Sequences to Its Race Diversity

Hirokazu OCHIAI, Yasuhiro INOUE[1], Masaru TAKEYA, Aeni SASAKI and Hisatoshi KAKU*

Genetic Diversity Department, National Institute of Agrobiological Sciences
(Tsukuba, Ibaraki 305–8602, Japan)

Abstract

The plant-pathogenic prokaryote *Xanthomonas oryzae* pv. *oryzae* (*Xoo*) causes bacterial blight, one of the most important diseases of rice. The bacterium is a model organism for the analysis of plant-pathogen interaction, because more than 30 races differing in virulence and 25 resistance genes in rice have been reported to date. We present here the complete genome sequence of *Xoo* strain MAFF 311018. The size of the genome was 4,940,217 bp, in a single circular chromosome. The genome structure of *Xoo* MAFF 311018 was characterized by large numbers of effector (avr) genes of the *avrBs3*/*pth* family and insertion sequences (ISs). RFLP analysis of diverse strains using IS*Xo1* as a probe suggests that the prevalence of mobile elements in this species, which can bring about genome inversions and rearrangement, may have played a major role in generating the high degree of genetic diversity and race differentiation characteristic of this pathogen. The *Xoo* MAFF 311018 sequence was also highly similar to those of *X. axonopodis* pv. *citri* and *X. campestris* pv. *campestris* with the exception of the large number of effectors and IS elements, and numerous inversions and rearrangements.

Discipline: Biotechnology
Additional key words: *avr*, comparative genomics, *hrp*, pathogenicity

Introduction

All known species of the genus *Xanthomonas*, a member of the gamma subdivision of the Proteobacteria, are plant-associated and most are plant pathogens. Among them, *Xanthomonas oryzae* pv. *oryzae* (*Xoo*), is a pathogen of the staple crop plant rice (*Oryza sativa*). *Xoo* causes bacterial blight of rice, alongside rice blast caused by the fungus *Magnaporthe grisea*. In addition to the importance as a pathogen, the bacterium is known to be an ideal model for studying plant–pathogen interactions, race differentiation and evolution of plant pathogens. Therefore, the rice–*Xoo* interaction has been studied at the molecular level with special reference to the *hrp* genes, encoding the type III secretion system, and the *avr* genes, encoding Avr proteins[6,20]. The diversity of races in *Xoo* is remarkable: more than 30 races of different virulence have been reported worldwide[25]. Because of the variability of virulence, the breeding of resistant cultivars always confronts difficulties with the durability of resistance. Race differentiation is associated with diversity of host resistance genes, and the specificity is controlled in a 'gene-for-gene' manner[14]. More than 25 rice resistance genes for bacterial blight have been identified, mostly in Japan and at the International Rice Research Institute[13]. Recently, genome structure has been studied in certain plant-pathogenic bacteria as well as in many industrial and human-pathogenic prokaryotes, and comparative genomics of these organisms has revealed differences—key to understanding unique characteristics[4,31,33,36]. Genome sequencing has been completed for *Xanthomonas axonopodis* pv. *citri* (*Xac*) and *X. campestris* pv. *campestris* (*Xcc*)[9]. Plant resistance genes against these two *Xanthomonas* pathogens have not been reported, however, and race differentiation is not clear. Nor is

H. Ochiai and Y. Inoue contributed equally to this work.

Present address:
[1] Department of Plant Pathology, National Agricultural Research Center (Tsukuba, Ibaraki 305–8666, Japan)
*Corresponding author: +81–29–838–7470; e-mail hkaku@affrc.go.jp

Received 27 June 2005; accepted 13 July 2005.

extensive race differentiation clear in the other plant pathogenic bacteria, *Agrobacterium tumefaciens*[36], *Pseudomonas syringae* pv. *tomato*[4], *Ralstonia solanacearum*[31], and *Xylella fastidiosa*[33], whose genome sequencing has been completed. Therefore, *Xoo* is the plant pathogenic bacterium in which genome sequencing has revealed very extensive race differentiation. In addition, the whole genome sequence of its native host has also been completed[32], and analysis of the host–parasite interaction on the basis of the two genomes can be expected to be useful.

We report here the complete genome sequence of strain MAFF 311018 (T7174) of an apparently highly evolved plant-pathogenic bacterium *Xoo*; this sequence provides clues to the endless race between rice and *Xoo*.

Materials and methods

1. DNA sequencing and assembly

The bacterial strain sequenced was MAFF 311018 (T7174), a representative Japanese race 1 strain registered in the MAFF Genebank (National Institute of Agrobiological Sciences, Tsukuba, Japan). The nucleotide sequence was determined by the whole-genome shotgun strategy. The accumulated sequence data were assembled with the GenomeGambler version 1.51 program (Xanagen Inc., Tokyo, Japan)[30], which includes the Phred/Phrap/Consed package (Philip Green, University of Washington, Seattle, USA). In addition to the above sequences, both end sequences of BAC clones[27], with an average size of 107 kb, facilitated the gap-closure process as well as confirmation of the orientation and integrity of the entire genome. The final gaps in sequences were filled by the primer walking method.

2. Gene prediction and annotation

Protein-coding genes were predicted with GeneHacker[37], GenomeGambler version 1.51 and the Glimmer program[11]. We searched for predicted proteins in the non-redundant protein database with the BLASTP program[2]. Regions of the genome without ORFs were reevaluated with the BLASTX program. RNA species were identified by using the BLASTN and tRNAscan-SE programs[22]. Finally, each putatively identified gene was analysed with the XanaGenome program (Xanagen Inc., Tokyo, Japan) for functional annotation. Insertion sequences (ISs) were classified by BLAST analysis with the ISFinder database (www-is.biotoul.fr/).

3. Comparative genomics among *Xanthomonas* strains

We compared the genome assemblies for *X. oryzae* pv. *oryzae* strains MAFF 311018, *X. axonopodis* pv. *citri* and *X. campestris* pv. *campestris* by using the MUMmer program[19] with default values (minimum match length: 20 bp), and we compared the translated ORF sets of *Xoo*, *Xac* and *Xcc* by using the BLASTP program. Shared genes were defined using an e-value cutoff of e-20.

4. Phylogenetic and pathotypic analysis

Bacterial strains used in phylogenetic and pathotypic analysis were collected from Sri Lanka. They had been used for RFLP (restriction fragment length polymorphism) typing and pathogenic analysis in a previous study[26]. RFLP analysis of these Sri Lankan strains was performed using IS*Xo1* (AF225214) as a probe. The DNA banding patterns (haplotype) of each strain was coded in binary form by scoring the presence or absence of each band. Low intensity bands and fragments ≥ 7 kb in size were not considered. Dendrograms were constructed from the distance matrix data from the Dice similarity coefficient through the WinDist program[40] by unweight pair group method using arithmetic averages (UPGMA) using the NEIGHBOR and DRAWGRAM programs in the PHYLIP package[16]. A data matrix based on virulence to 11 near-isogenic lines and one cultivar containing a single gene for resistance, was generated from the virulence data of each race by scoring avirulence as 0 and virulence as 1. A similarity matrix was computed with the SIMQUAL program (NTSYS-pc, version 1.80; Exeter Biological Software) using simple matching coefficients of similarity. A phenogram was reproduced by UPGMA in the SAHN program.

Results and discussion

1. General features of the *Xoo* genome

The deduced genome of *X. oryzae* pv. *oryzae* (*Xoo*) strain MAFF 311018 was a circular chromosome of 4,940,217 bp, and the average G+C content was 63.7% (Fig. 1, Table 1). No plasmid was detected in the course of genome assembly. The genome size was slightly different from that previously indicated, a size of about 4.8 Mb by pulse-field gel electrophoresis[27]. Two copies of the *rrn* operon were identified on the genome in the order of 16S-tRNA^{Ala}-tRNA^{Ile}-23S-5S. A total of 53 tRNAs representing 43 tRNA species were found on the genome using the tRNA scan-SE program. A total of 4,372 ORFs were identified within the MAFF 311018 genome. Of the predicted genes, 2,799 (64%) were assigned putative functions, 1,383 (32%) had similarities to proteins of unknown function (conserved hypothetical proteins), and 190 (4%) had no significant similarity to any registered genes (Table 1). At least two possibly defective proph-

ages were found on the genome. *Xoo* MAFF 311018 had two gene clusters for a putative type I restriction modification system and one gene cluster for a type II restriction/modification system.

2. Insertion sequences

A total of 611 ISs of 25 types were found in the genome of *Xoo*. Their ratio in relation to the whole genome was approximately 10%. This percentage was remarkably high compared with that of other sequenced plant-pathogenic bacteria[4,9,31,33,36], which may be a characteristic feature of *Xoo*. ISs were located throughout the genome, and they were repeated tandemly in many loci. In some cases, multiple IS regions (IS islands) that encompassed about 30 kb were also present. We identified 386 of the ISs as full length, and the remaining 225

Table 1. General features of the *Xanthomonas oryzae* pv. *oryzae* genome

Length (bp)	4,940,217
G+C content (%)	63.72
Protein-coding region (% genome size)	84.12
Protein-coding genes	
With assigned function	2,799
Conserved hypothetical	1,383
Hypothetical	190
Total	4,372
Transfer RNA	53
Ribosomal RNA operons	2
Plasmid	0
Insertion sequence elements (IS)[a)]	386(225)

a): Number of full-length copies; number of incomplete copies is in parentheses.

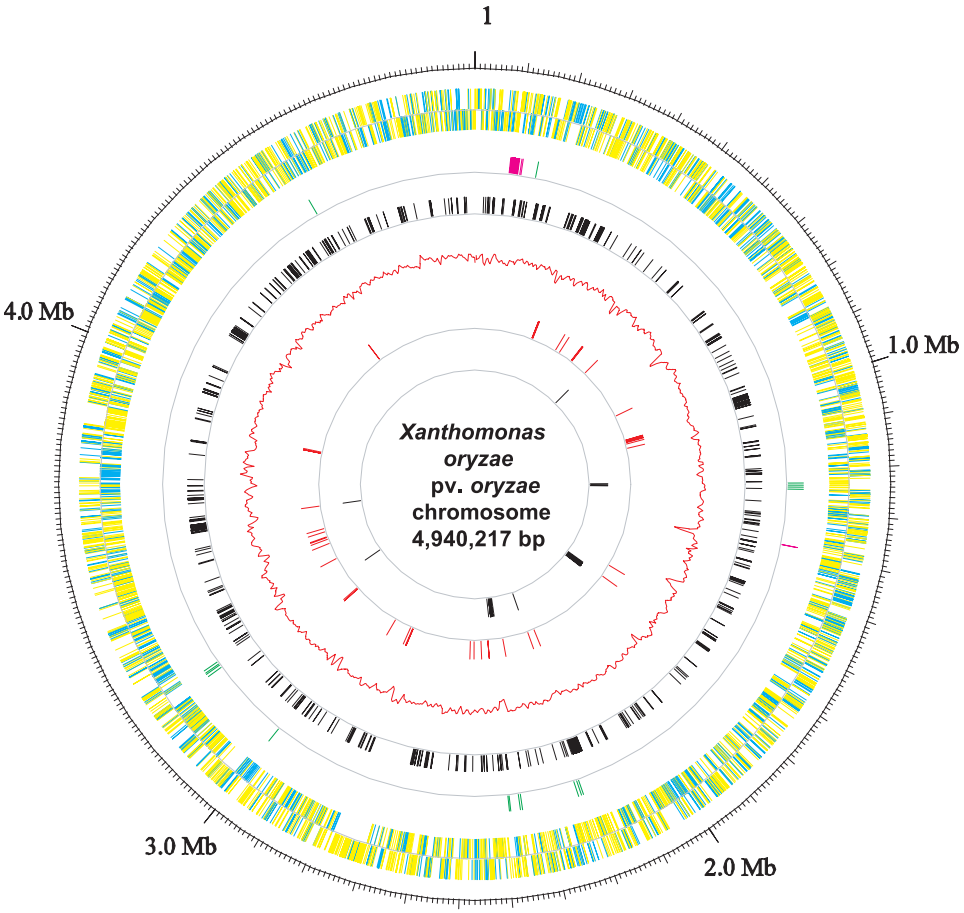

Fig. 1. Graphical representation of the chromosome of *Xanthomonas oryzae* pv. *oryzae* MAFF311018
Outermost circle indicates locations on the chromosome in base pairs (each unit is 100 kb). The second and the third circles show the positions of predicted genes in the clockwise and anticlockwise directions, respectively. Genes with assigned functions are depicted in yellow, and those whose functions could not be deduced are in blue. The fourth circle shows the locations of members of the *hrp* gene cluster (magenta) and *avr* genes (green). The fifth circle shows the positions of insertion sequences. The sixth circle indicates the 20-kb window-average of G+C content. The locations of tRNA and rRNA genes are shown in the seventh circle. The eighth circle shows the locations of prophage and phage-related genes.

Table 2. IS elements present in *X. oryzae* pv. *oryzae* MAFF 311018

Name	Family	Group	No. of full-length copies	No. of partial (truncated) copies	Structure
IS*Xo1*	IS*5*	IS*5*	42	2	orfA
IS*Xo2*	IS*NCY*		7	3	orfA
IS*Xo3*	IS*5*	IS*1031*	30	10	orfA / orfB
IS*Xo5*	IS*NCY*		47	21	orfA
IS*Xo7*	IS*630*		7	0	orfA
IS*Xo8*	ISI		36	34	orfA
IS*1112*	IS*30*		20	12	orfA
IS*1113*	IS*NCY*		8	2	orfA
IS*Xoo2*	IS*630*		11	15	orfA
IS*Xoo3*	IS*3*	IS*407*	16	26	orfA / orfB
IS*Xoo4*	IS*5*	IS*5*	11	41	orfA
IS*Xoo5*	IS*5*	IS*5*	10	2	orfA
IS*Xoo6*	IS*5*	IS*5*	10	2	orfA
IS*Xoo7*	IS*5*	IS*5*	2	0	orfA
IS*Xoo8*	IS*4*		11	8	orfA
IS*Xoo9*	IS*3*	IS*51*	1	1	orfA / orfB
IS*Xoo10*	IS*NCY*		1	0	orfA
IS*Xoo11*	unknown		46	0	orfA
IS*Xoo12*	unknown	IS*4*	42	0	orfA
IS*Xoo13*	IS*3L*		19	6	orfA
IS*Xoo14*	IS*5*		5	1	orfA
IS*Xoo15*	IS30		2	16	orfA
IS*Xoo16*	IS*630*		2	7	orfA
Others	IS*3*		0	8	
	unknown		0	8	

were truncated. On the basis of the ISFinder database, almost all of the ISs could be classified into seven families (Table 2). Members of the IS*5* and IS*NCY* families were highly represented, with 110 and 63 full-length copies, respectively. Novel elements included IS*Xoo11*, IS*Xoo12*, IS*Xoo13*, IS*Xoo14*, IS*Xoo15*, and IS*Xoo16*.

Rearrangement of genomes by IS has been reported in animal pathogens. Deng et al.[12] alluded to the important role played by repeat elements (namely IS elements) in explaining the unique genome rearrangement between two sequenced *Yersinia pestis* strains. Similar genome rearrangement and reductive evolution through gene loss were also reported in *Yersinia pestis* and *Y. pseudotuberculosis*[7]. Comparative analysis of the genomes of the two species which are drastically different in pathogenicity and transmission further supported the role played by IS elements in genome evolution. Similar findings were obtained in the two pathogenic *Burkholderia* species[24]. A highly evolved obligate parasite *Burkholderia mallei* genome harbors numerous IS elements that most likely have mediated extensive genome-wide insertion, deletion and inversion mutations relative to *B. pseudomallei*. Comparative alignment of genome sequences between *Burkholderia mallei* and its related species *B. pseudomallei* showed that numerous IS were observed in the synteny points of the genomes of the two species, which suggests recombination mediated by IS.

The relationship between pathotypic and genetic diversity of *Xoo* strains has been studied mainly based on analyses of differential interactions with resistance genes and DNA fingerprinting using insertion sequences[1,23,26]. The results suggested that some phylogenetic lineages based on insertion sequences were related to pathotypes (races). Using Sri Lankan strains of *Xoo*, we investigated whether there was some relationship between race and phylogenetic group. We observed an association between phylogenetic groups and race groups (Fig. 2 & Table 3). Thus, it is likely that the numerous ISs might be important in strain or race differentiation in *Xoo*, because

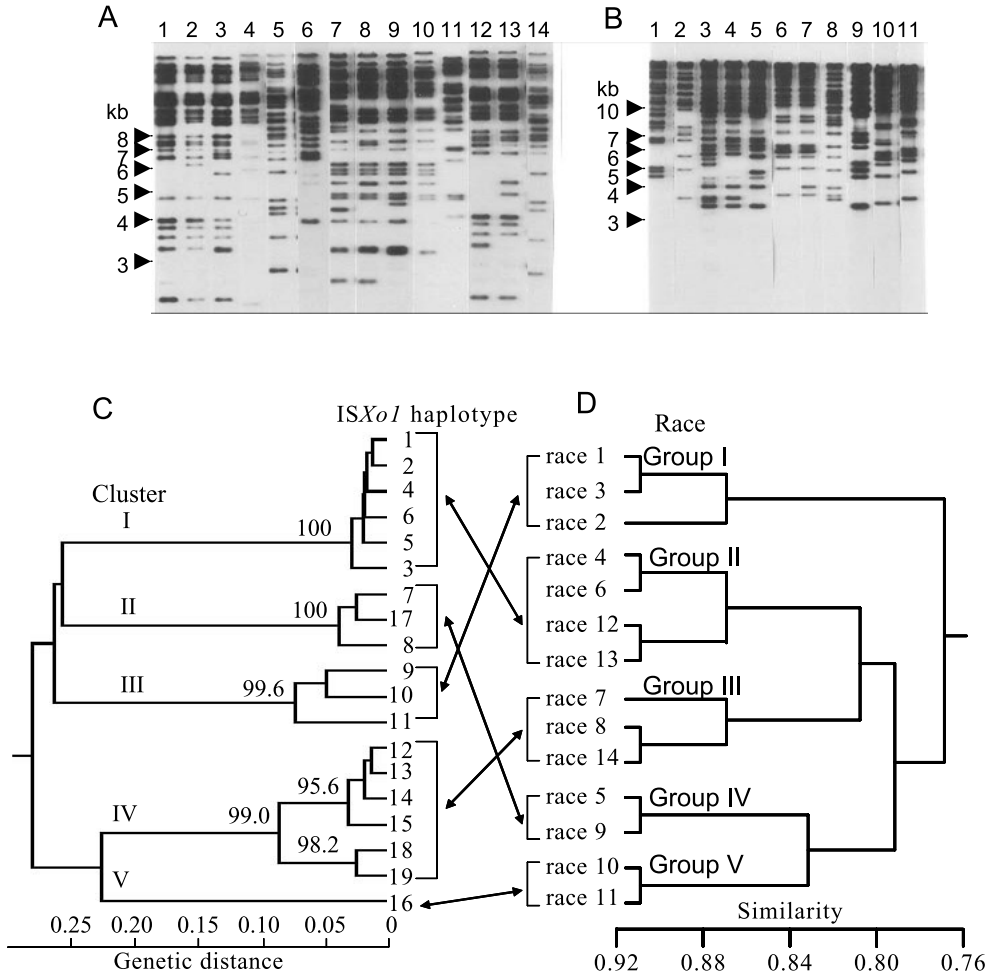

Fig. 2. Southern hybridization profiles and relationship between pathogenic race groups and phylogenetic groups based on RFLP types by IS*Xo1* insertion sequence

A: *Eco*RI- & B: *Bam*HI-digested DNA of Sri Lankan strains of *Xanthomonas oryzae* pv. *oryzae* by IS*Xo1* as a probe. Lane numbers refer to the restriction fragment length polymorphism types given in Table 3. Sizes (kilobases) are indicated on the left.

C: Dendrogram constructed with UPGMA of Sri Lankan strains of *X. oryzae* pv. *oryzae* derived from RFLP data of IS*Xo1* as a probe. D: Phenogram of Sri Lankan strains of *X. oryzae* pv. *oryzae* based on virulence to 11 near-isogenic lines and one cultivar containing a single gene for resistance. The IS*Xo1* type (haplotype) corresponds to those in Table 3. Race corresponds to those in Table 3. The numbers on the main branches of the left dendrogram indicate the percent bootstrap values for 1,000 replicates. Arrows indicate relative association between race group and phylogenetic group.

Table 3. RFLP and pathogenicity analysis of *Xanthomonas oryzae* pv. *oryzae* strains from Sri Lanka

Strain	RFLP types (IS*Xo1*)				Race[b)]	Strain	RFLP types (IS*Xo1*)				Race[b)]
	B	E	Haplotype	Cluster			B	E	Haplotype	Cluster	
SL9501	1	1	1	1	6	SL9571	4	6	10	3	3
SL9504	2	5	7	2	5	SL9574	9	5	17	2	9
SL9508	1	2	2	1	6	SL9575	4	6	10	3	3
SL9510	4	6	10	3	2	SL9577	6	10	14	4	7
SL9511	3	6	9	3	3	SL9585	2	5	7	2	5
SL9517	4	6	10	3	3	SL9587	10	7	18	4	8
SL9521	4	6	10	3	3	SL9592	8	11	16	5	5
SL9524	4	6	10	3	3	SL9595	8	11	16	5	10
SL9526	1	1	1	1	4	SL9599	2	14	8	2	5
SL9527	5	6	11	3	3	SL95103	2	5	7	2	9
SL9528	2	5	7	2	5	SL95111	1	13	6	1	14
SL9530	2	5	7	2	5	SL95113	6	8	12	4	8
SL9532	2	5	7	2	5	SL95115	4	6	10	3	3
SL9533	1	2	2	1	6	SL95119	1	3	2	1	12
SL9536	11	7	19	4	7	SL95120	8	11	16	5	11
SL9542	1	1	1	1	4	SL95122	1	3	3	1	12
SL9545	1	1	1	1	4	SL95123	1	4	4	1	6
SL9548	6	9	13	4	8	SL95127	1	12	5	1	12
SL9549	6	9	13	4	8	SL95130	1	1	1	1	4
SL9551	7	8	15	4	7	SL95133	8	11	16	5	11
SL9552	6	8	12	4	8	SL95136	1	1	1	1	6
SL9554	6	8	12	4	8	SL95138	8	11	16	5	9
SL9561	1	1	1	1	4	SL95139	6	8	12	4	8
SL9564	10	7	18	4	8	SL95143	1	2	2	1	13
SL9568	2	5	7	2	5	SL95144	9	5	17	2	9

a): The B and E columns indicate the RFLP patterns obtained with *Bam* HI and *Eco* RI, presented in Fig. 2 (A & B). The IS*Xo1* type (haplotype) was determined by the combination of RFLP patterns obtained from the two enzymes. Cluster types correspond to RFLP cluster in Fig. 2 (C). b): Races were determined in a previous study[26].

mobile elements can disrupt genes, enhance recombination, and introduce mutations.

3. Pathogenicity-related genes

(1) *hrp* gene cluster

Of the genes related to the pathogenicity of plant-pathogenic bacteria, *hrp* genes encoding a type III secretion system (TTSS) are the most important. The TTSS is conserved among both plant and animal pathogens and injects effector proteins into host cells[6]. The *hrp* gene cluster found in the *Xoo* genome is composed of 27 genes extending from *hpa2* to *hpaF*, and the structure was similar to those of other characterized *Xanthomonas hrp* gene clusters[15]. One exception was the *hpaB* (*hrpE2*)–*hrpF* region. In the *Xoo* genome, it is notable that three novel genes were found in the region between *hpaB* and *hrpF* (Fig. 3). One was located downstream of *hpaB* (*hrpE2*) and in the same direction as the *hrpE* operon. The other two genes were also located downstream of *hpaB* (*hrpE2*), but in the opposite direction. It was also interesting that four tandem transposase homologues were present between *hpaB* and *hrpF* (Fig. 3).

(2) *avr* genes

Avr proteins are one of the type III effectors that elicit disease resistance in hosts with corresponding resistance (*R*) genes and thereby function as determinants of race-cultivar specificity. Many have been shown to be dual-acting, contributing to virulence in the absence of a host *R* gene[8]. Some *avr* genes have been reported in many plant-pathogenic xanthomonads, as well as in many plant-pathogenic pseudomonads[20]. Strain MAFF 311018 exhibits a high degree of host-specificity at the rice cultivar level, and we therefore expected it to contain a number of avirulence genes. Although several types of *avr* gene have been characterized in xanthomonads[20], of these, *Xoo* possesses only *avrBs2* and members of the

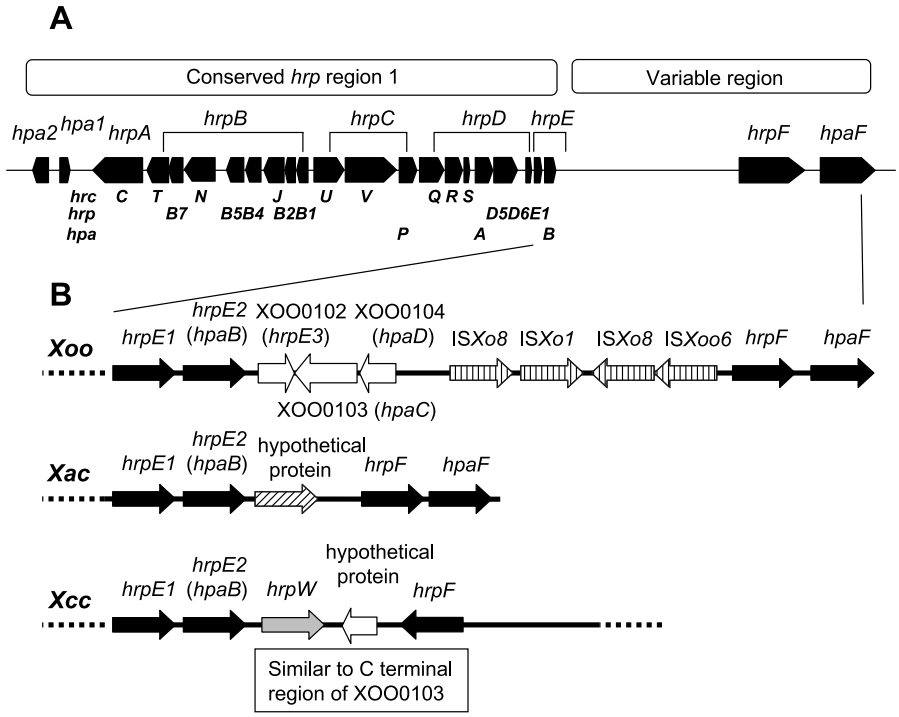

Fig. 3. Comparison of the genetic organization of the *hrp* gene clusters of *Xoo*, *Xcc* and *Xac*
A: Genetic organization of the common genes. B: The variable regions of the *hrp* gene cluster. Each gene is named above or below the arrows. The maps are illustrated based on the annotated nucleotide sequence data from the GenBank database. The following sequences were used: *X. axonopodis* pv. *citri* 306 genes (accession no. AE008923); *X. campestris* pv. *campestris* ATCC 33913 genes (accession no. AE008922).

Table 4. List of members of the *avr*/*pth* gene family of *Xoo* genome

Gene ID	Position	Strand	*avr* region	Length (bp)	No. of repeat units[a)]	Note
XOO1132	1,228,783–1,231,791	–	I	3,009	12.5	
XOO1134	1,232,781–1,236,191	–	I	3,411	16.5	
XOO1136	1,237,181–1,241,629	–	I	4,449	26.5	
XOO1138	1,242,619–1,246,335	–	I	3,717	19.5	
XOO1996	2,201,149–2,204,664	–	II	3,516	17.5	
XOO1998	2,205,654–2,209,376	–	II	3,723	19.5	
XOO2001	2,212,644–2,216,771	–	II	4,128	23.5	
XOO2127	2,352,186–2,356,040	–	III	3,855	20.5	
XOO2129	2,357,030–2,360,329	–	III	3,300	15.5	
XOO2158	2,383,115–2,386,828	+	IV	3,714	19.5	
XOO2160	2,387,818–2,392,653	+	IV	4,836	30.5	
XOO2667	2,999,333–3,001,924	–	V	2,592	17.5	Insertion of IS*Xoo9*
XOO2864	3,213,703–3,218,643	+	VI	4,941	31.5	
XOO2865	3,219,633–3,223,352	+	VI	3,720	19.5	
XOO2866	3,224,342–3,228,793	+	VI	4,452	26.5	
XOO2868	3,229,783–3,234,123	+	VI	4,341	25.5	Identical with *avrXa7*
XOO4014	4,525,416–4,529,345	+	VII	3,930	21.5	

a): Number of 102-bp (34-aa) repeat units in the central domain.

avrBs3/*pth* family, which are widely distributed in xanthomonads. Notably, *Xoo* contains multiple copies of members of the *avrBs3*/*pth* gene family. In *Xoo*, we found 17 copies on the genome, distributed and clustered in seven different genomic regions (Table 4). The structure of the *avrBs3/pth* gene family has been elucidated[20]. The proteins have the following functional domains or motifs: (1) a central region of near-perfect 34 amino acid (102 bp) repeats that vary in number; (2) leucine-rich repeats (LRRs); (3) nuclear localization signals (NLS); and (4) an acidic transcriptional activation domain (AAD). Although the nucleotide sequences were very similar among these multiple copies, they were distinct from each other. Their differences lay mainly in the number of repeats in the central domain. The number of repeats varied from 12.5 to 31.5 copies (Table 4). From two to four *avr* genes were in most *avr* regions, and transposable elements such as ISs and phage-related genes were located in neighbouring areas. These ISs might promote duplication and rearrangement of *avr* genes, and likely contributed to their proliferation in the *Xoo* genome. Maps of these *avr* gene loci are given in Fig. 4. The *avr* V region was an exceptional structure that was disrupted by a transposase gene (IS*Xoo9*) at the 3′-end. The gene direction of *avr* regions I, II and III was antisense strand and that of *avr* regions IV, VI and VII was sense strand in an opposing orientation (Fig. 4).

(3) HrpX regulons

Identification of the genes encoding TTSS effectors is a key step to understanding the function of the TTSS in plant pathogens. In xanthomonads, expression of the structural genes of the TTSS and some effector genes is mediated by the *hrpG* and *hrpX* gene products, both of which were present in *Xoo* as well as in *Xac* and *Xcc*. Several genes regulated in a HrpX-dependent manner possess the consensus nucleotide sequence

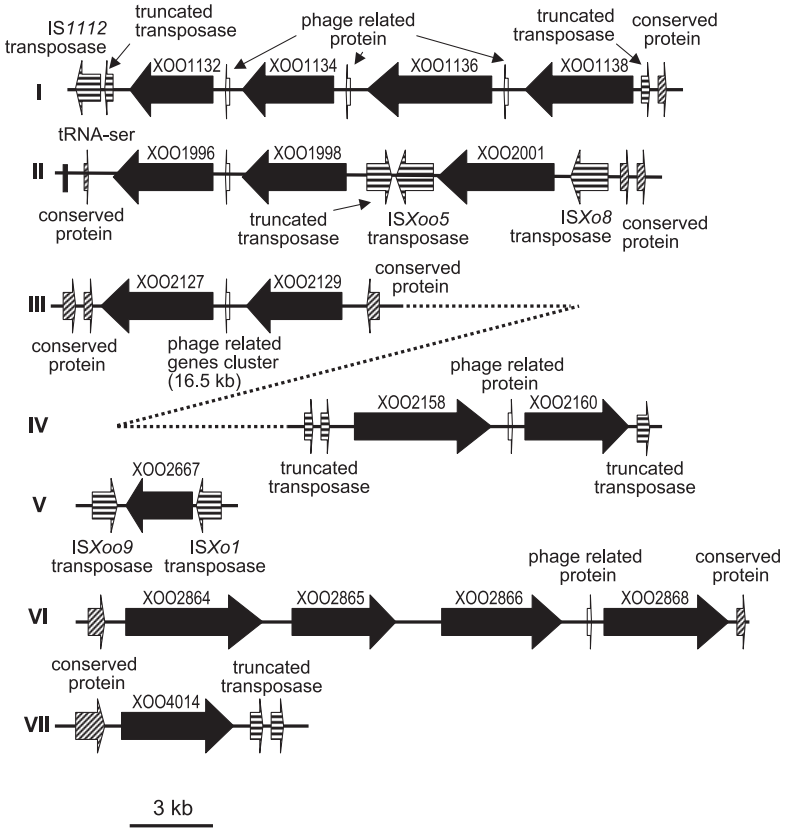

Fig. 4. Gene map of the *avr*/*pth* gene family of *Xanthomonas oryzae* pv. *oryzae*
Relative positions of each *avr* region are illustrated. Cluster coordinates are as follows: I, 1,228,783-1,246,335 bp; II, 2,201,149-2,216,771 bp; III, 2,352,186-2,360,329 bp; IV, 2,383,115-2,392,653 bp; V, 2,999,333-3,001,924 bp; VI, 3,213,703-3,234,123 bp; and VII, 4,525,416-4,529,345 bp. Information on each *avr* gene refers to Table 4.

Table 5. Proposed HrpX regulons

PIP position	Distance*	Gene ID	Gene product
***hrp* gene cluster**			
87,199	214	XOO0080	Hpa2
87,290	136	XOO0081	Hpa1
96,205	82	XOO0090	HrpB1
96,243	70	XOO0091	HrcU (HrpC1)
99,847	249	XOO0094	HrcQ (HrpD1)
Others			
107,769	97	XOO0104	Hypothetical protein
153,037	66	XOO0148	Avirulence protein (AvrBs2 homolog)
358,953	47	XOO0327	Hypothetical protein
482,032	147	XOO0440	Ribonucleoside-diphosphate reductase beta chain (NrdB)
535,147	291	XOO0490	Conserved hypothetical protein
1,293,989	250	XOO1183	RNA polymerase sigma-54 factor (RpoN)
1,630,648	65	XOO1488	Hypothetical protein
1,819,604	71	XOO1662	Leucin rich protein
1,823,472	171	XOO1669	Conserved hypothetical protein
1,833,094	192	XOO1674	Conserved hypothetical protein
1,881,304	118	XOO1706	TonB-dependent receptor (BtuB)
2,012,045	281	XOO1828	Homocysteine S-methyltransferase (MmuM)
2,017,385	147	XOO1832	Conserved hypothetical protein
2,028,028	94	XOO1842	Conserved hypothetical protein
2,129,487	296	XOO1925	Oxoglutarate dehydrogenase (OdhA)
2,151,054	278	XOO1950	Transcriptional regulator
2,431,967	141	XOO2193	6-phosphogluconolactonase
2,519,236	278	XOO2263	Conserved hypothetical protein
2,824,592	95	XOO2532	Peptidase
3,063,415	279	XOO2716	Beta-keto-adipate enol-lactone hydrolase (PcaD)
3,246,583	192	XOO2877	Conserved hypothetical protein
3,281,958	191	XOO2901	Conserved hypothetical protein
3,285,423	118	XOO2905	Conserved hypothetical protein
3,359,421	154	XOO2967	Conserved hypothetical protein
4,301,354	196	XOO3800	Conserved hypothetical protein
4,330,117	154	XOO3824	Anthranilate synthase component I (TrpE)
4,502,407	226	XOO3994	Transferase
4,596,231	297	XOO4073	Putative 5'-nucleotidase
4,607,453	78	XOO4083	2-keto-3-deoxy-D-gluconate transport system (KdgT)
4,662,134	64	XOO4134	Conserved hypothetical protein
4,713,113	39	XOO4168	Conserved hypothetical protein
4,931,705	115	XOO4365	Conserved hypothetical protein

* Distance (bp) upstream from predicted start codon.

TTCGC…N15…TTCGC (PIP box)[6]. The PIP box is a useful tool for identifying candidate genes in the HrpX regulon, especially genes encoding secreted effector proteins. In *Xoo*, we detected 37 perfect or near-perfect copies of the PIP box (TTCGN…N15…TTCGN) in putative promoter regions of predicted genes (Table 5). Five of these predicted genes were located in the *hrp* gene cluster, and others were scattered throughout the genome.

4. Other features

Activation of virulence gene expression in *Xoo* is thought to be under complex control, and has not yet

been completely clarified. In general, it is thought that many pathogenicity-related genes, including the *hrp* gene cluster, are controlled by HrpG and HrpX, key regulatory proteins[35]. However, activation of *hrpG*, which is a transcriptional activator of the *ompR* family of two-component regulators, is still not understood. Interestingly, we found that *Xoo* possesses homologues of two sets of characteristic two-component regulators involved in microbial interactions with plants. One set is the *virA/virG* two-component system of *Agrobacterium tumefaciens*[34], which activates T-DNA transfer in response to monosaccharides and phenolic compounds. The other is the *nodV/nodW* two-component system of *Bradyrhizobium japonicum*[17], which responds to plant-derived flavonoids and provides an alternative pathway for activating genes involved in legume nodulation and symbiosis. The role of the homologues in *Xoo* remains unknown, but their similarity to the *vir* and *nod* gene regulators suggests that they may be involved in regulation of pathogenicity in response to plant or environmental signals.

Attachment to the tissue surface is the first step for pathogenic bacteria in establishing infection. Many Gram-negative bacteria have type IV fimbriae (also called pili) for adhesion. In *Xoo*, type IV fimbriae genes were distributed in several loci on the genome. In addition, a previous report showed that *Xoo* had a non-fimbrial adhesion protein designated as XadA, which was an outer membrane protein that plays a role in virulence[29]. We identified two *xadA* homologues at different loci in MAFF 311018.

In MAFF 311018 we identified gene homologues related to the production of toxins. One cluster was related to putative thermostable hemolysin-like genes of *Vibrio cholerae*[18]. A second was related to a putative colicin V secretion protein of *Xylella fastidiosa*[33]. A third was related to rhizobitoxine of *Bradyrhizobium elkanii*[41].

Several regions composed of specific genes, which were not present in other xanthomonads, were found on the MAFF 311018 genome. Many genes belonging to these regions are related mainly to *P. syringae*, *R. solanacearum*, *Mesorhizobium*, and *Bradyrhizobium* strains, but their functions are unknown (conserved hypothetical proteins). Their lengths range from approximately 15 kb to 90 kb.

5. Comparative genomics of three *Xanthomonas* species, *Xoo, Xac and Xcc*

First, we analysed the orthologous relationships among the proteins deduced from the genes, excluding the transposable elements in *Xoo*, *Xac* and *Xcc* (Fig. 5). Of the 3,697 proteins of *Xoo*, 2,771 (74.9%) were identi-

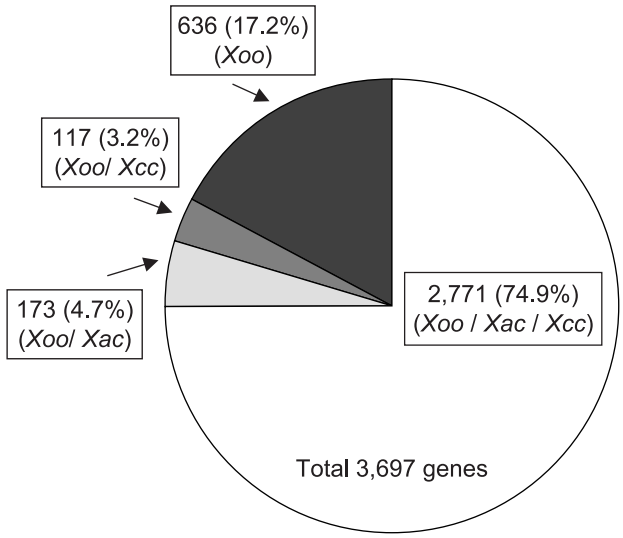

Fig. 5. Conservation of *Xoo* genes in *Xac* and *Xcc*

Genes were considered conserved if a BLASTP alignment of their predicted products yielded e-values less than or equal to 1×10^{-20}. Transposable elements were excluded from this analysis.
: Genes shared by all three species.
: Genes shared by *Xoo* and *Xac*.
: Genes shared by *Xoo* and *Xcc*.
: Genes unique to *Xoo*.

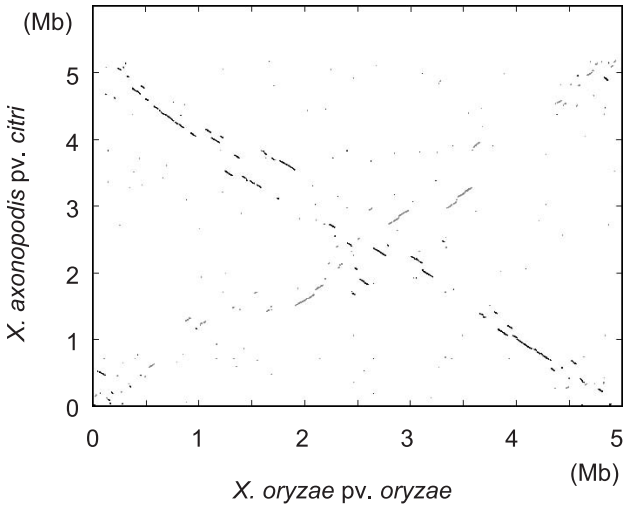

Fig. 6. MUMmer comparison analysis of the chromosomes between *X. oryzae* pv. *oryzae* MAFF 311018 and *X. axonopodis* pv. *citri* 306

Nucleotide sequence alignments were generated by using the software nucmer in the MUMmer package 3.0. The forward matches are displayed in the upward slope, and the reverse matches are displayed in the downward slope.

fied as orthologous with those of both *Xac* and *Xcc*; 173 proteins (4.7%) were shared with only *Xac*, and 117 proteins (3.2%) were shared with only *Xcc*. The other 636 putative proteins (17.2%) had no relationship (e-values $>1 \times 10^{-20}$) to proteins encoded in the genomes of *Xac* and *Xcc*. Of these 636 putative proteins, 190 were hypothetical genes whose amino acids sequences showed no significant similarity to those of any other proteins, and 446 were conserved proteins. Of the 446 conserved proteins, 325 had unknown functions, and 201 were proteins conserved in other bacteria (such as *Pseudomonas*, *Ralstonia* and *Rhizobium*).

We performed whole genome nucleotide alignment to determine the synteny of *Xoo* relative to *Xcc* and *Xac*. The results showed partial synteny, but had numerous inversions, rearrangements and deletions (Fig. 6). As the *Xoo* genome harbours a greater number and variety of IS elements than *Xac* and *Xcc*, genome-wide inversion, rearrangement and deletion might be expected to have occurred to a greater extent in the *Xoo* genome. In fact, most of the synteny break-points between the *Xoo* and *Xac* genomes were bounded by IS elements in *Xoo*.

The *Xanthomonas hrp* gene cluster contains six operons (*hrpA* to *hrpF*, composed of 22 genes) and an additional two genes outside *hrpA*[15]. Comparison of the *hrp* cluster among *Xoo*, *Xac* and *Xcc* revealed that the gene orders were similar but the clusters were located in different regions on the genomes. One exception to their similarity was a region between *hpaB* (*hrpE2*) and *hrpF*, which varied in terms of its length and predicted genes (Fig. 3). *Xoo* possessed three genes and four transposase genes in this region, whereas *Xac* had only one gene and *Xcc* had two. Although *hrpW* was present within this region in *Xcc* and outside the *hrp* cluster in *Xac*, it was absent from the *Xoo* genome. The *hrp* core region (*hrpA* to *hpaB*), encoding mainly proteins of the TTSS, was highly conserved between *Xoo* and *Xac*, with more than 80% identity at the amino acid level with the exception of HpaA. In contrast, the identity value was relatively low between *Xoo* and *Xcc*. The genes encoding Hpa1, HrpD6, HrpE1, and HrpF proteins had relatively low identities among *Xoo*, *Xac* and *Xcc*. These differences may influence pathogenicity and interactions with host plants. In addition to the TTSS secretion system, other secretion systems such as type II and type IV are important for pathogenic bacteria to deliver degradative enzymes and toxins and to transfer T-DNA into the cell. Two type II secretion systems (xps cluster and xcs cluster) were identified in both *Xac* and *Xcc*, but the xcs cluster was absent in *Xoo*. Also, both *Xac* and *Xcc* had type IV secretion systems (*virB* operon), which have been well characterized in *A. tumefaciens*. Genes for this system are absent in *Xoo*. The roles which these elements play in xanthomonad pathogenicity, if any, and the reason why *Xoo* lacks these elements remain to be elucidated[5].

In xanthomonads, three *avr* gene families, *avrBs1*, *avrBs2* and *avrBs3*/*pth*, have been reported to date[20]. Furthermore, the *avrPphE*, *avrC*, *yopJ* , and *avrXca* families have been identified in *Xac* or *Xcc*[9]. There was a sharp contrast in the distribution of *avr* gene families among *Xoo*, *Xac* and *Xcc* (Table 6). The *avrBs2* gene was common among the three xanthomonads, whereas

Table 6. Quantitative comparison of insertion sequences, proposed HrpX regulons and putative *avr* genes among the three *Xanthomonas* species

	X. oryzae pv. *oryzae* MAFF 311018	*X. axonopodis* pv. *citri* 306	*X. campestris* pv. *campestris* ATCC 33913
Length (bp)	4,940,217	5,175,554	5,076,187
G+C content (%)	63.7	64.7	65.0
Total number of predicted genes	4,372	4,313	4,182
Plasmid	0	2	0
Insertion sequence elements (IS)	386(225)	87	109
Proposed HrpX regulons	37	20	17
Putative avirulence genes (family)			
avrBs1	0	0	2
avrBs2	1	1	1
avrBs3 / pth	16 (genome)	4 (plasmid)	0
avrPphE	0	3	1
avrXca	0	0	2
yopJ	0	0	1
avrC	0	0	1

avrBs1 was found only in *Xcc*. Four *avrBs3*/*pth*–like genes are found in *Xac*, on two plasmids, and none are in *Xcc*, in stark contrast to the 16 members of the *avrBs3*/*pth* family found in the genome of *Xoo* MAFF 311018. Previous studies have reported that *Xoo* has at least 30 races[25]. Race differentiations of *Xac* and *Xcc* have not yet been reported. To our knowledge, there are no reports of natural mechanisms of how new *avr* genes or race differentiation are generated. Mutational analyses of *avrBs3*/*pth* genes revealed that the order and type of repeats in the central domain were important for race-specificity or pathogen fitness[3,38,39]. In general, repetitive regions of DNA are known to be active sites for homologous recombination[28]. In *Xanthomonas axonopodis* pv. *malvacearum*, it appears that multiple members of this *avr* gene family have arisen from duplication and divergence by intragenic or intergenic recombination, and they may promote race-change mutations[10]. In *Xoo*, insertion sequences and phage-related genes were abundant in most of the regions neighbouring this *avr* gene family (Fig. 4). Duplication and rearrangement of *avr* genes in the *Xoo* genome might be promoted by these ISs, and these multiple copies of *avr* genes, which contained various types of repeats in the central domain, might be generated via subsequent unequal homologous recombination. Therefore, a role for the multitude of these genes and the numerous IS elements in generating the diversity of races in *Xoo* is a compelling conclusion.

Conclusions

Our complete genome sequencing of *Xoo* MAFF 311018 revealed very unique features of the genome structure of *Xoo* suggestive of a highly evolved plant pathogen. These unique features were a large number of *avr* genes of one family and a large number of insertion sequences (IS). Based on comparison among *Xanthomonas* strains, we propose that these numerous effector genes and mobile elements are involved in the high degree of race differentiation. Comparative genomic analysis among multiple strains will be necessary to address this possibility definitively. In addition, the genome sequence of Korean strain KACC10331 was recently published[21], and comparative analysis of genomes among multiple strains of *Xoo* became possible. Comparative analysis shows overall a high degree of similarity with KACC10331 but they have differences in effector and IS element content as well as gene alignment. Therefore, such comparative genomics will provide further evidence for evolutionary mechanisms with a particular focus on understanding the extensive infrasubspecific diversity and race differentiation characteristic of *Xoo*. Furthermore, the whole genome sequence of its native host has also been completed, and analysis of the host–parasite interaction on the basis of the two genomes can be expected to foster exciting progress in understanding plant bacterial interactions and the evolution of race-cultivar specificity.

Acknowledgements

We thank A. Bogdanove for a critical reading of the manuscript, S. Tsuyumu and I. Toth for valuable discussion, and N. Katsura, K. Higo, T. Sasaki, A. Hasebe, and J. Kurisaki for their suggestions and encouragement. We also thank all the technicians who helped with this project. This work was supported by special coordination funds for promoting science and technology (from Ministry of Education, Culture, Sports, Science and Technology of Japan).

Data deposition: The sequence has been deposited in DDBJ under accession number AP008229. The sequences and gene information described in this paper will be accessible through the Web database, *Xanthomonas oryzae* pv. *oryzae* Genome Database ORF Viewer, at http://micro.dna.affrc.go.jp/xan/orf/.

References

1. Adhikari, T. B., Mew, T. W. & Leach, J. E. (1999) Genotypic and pathotypic diversity in *Xanthomonas oryzae* pv. *oryzae* in Nepal. *Phytopathology*, **89**, 687–694.
2. Altschul, S. F. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25,** 3389–3402.
3. Bai, J. et al. (2000) *Xanthomonas oryzae* pv. *oryzae* avirulence genes contribute differently and specifically to pathogen aggressiveness. *Mol. Plant-Microbe Interact.*, **13**, 1322–1329.
4. Buell, C. R. et al. (2003) The complete genome sequence of the *Arabidopsis* and tomato pathogen *Pseudomonas syringae* pv. *tomato* DC3000. *Proc. Natl. Acad. Sci. USA*, **100,** 10181–10186.
5. Burns, D. L. (1999) Biochemistry of type IV secretion. *Curr. Opin. Microbiol.*, **2,** 25–29.
6. Büttner, D. & Bonas, U. (2002) Getting across-bacterial type III effector proteins on their way to the plant cell. *EMBO J.*, **21,** 5313–5322.
7. Chain, P. S. G. et al. (2004) Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia pseudotuberculosis*. *Proc. Natl. Acad. Sci. USA*, **101**, 13826–13831.
8. Cornelis, G. R. & Gijsegem, V. (2000) Assembly and function of type III secretory systems. *Annu. Rev. Microbiol.*, **54,** 735–774.
9. Da Silva, A. C. R. et al. (2002) Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. *Nature*, **417,** 459–463.

10. De Feyter, R., Yang, Y. & Gabriel, D. W. (1993) Gene-for-gene interactions between cotton *R* genes and *Xanthomonas campestris* pv. *malvacearum avr* genes. *Mol. Plant-Microbe Interact.*, **6**, 225–237.
11. Delcher, A. L. et al. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27,** 4636–4641.
12. Deng, W. et al. (2002) Genome sequence of *Yersinia pestis* KIM. *J. Bacteriol.*, **184**, 4601–4611.
13. Ezuka, A. & Kaku, H. (2000) A historical review of bacterial blight of rice. *Bull. Natl. Inst. Agrobiol. Resour.*, **15,** 1–207.
14. Keen, N. T. (1990) Gene-for-gene complementarity in plant-pathogen interactions. *Annu. Rev. Genet.*, **24,** 447–463.
15. Kim, J. -G. et al. (2003) Characterization of the *Xanthomonas axonopodis* pv. *glycines hrp* pathogenicity island. *J. Bacteriol.*, **185,** 3155–3166.
16. Felsenstein, J. (1993) PHYLIP (Phylogeny inference package) Version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.
17. Göttfert, M., Grob, P. & Hennecke, H. (1990) Proposed regulatory pathway encoded by the *nodV* and *nodW* genes, determinants of host specificity in *Bradyrhizobium japonicum*. *Proc. Natl. Acad. Sci. USA*, **87,** 2680–2684.
18. Heidelberg, J. F. et al. (2000) DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature*, **406,** 477–483.
19. Kurtz, S. et al. (2004) Versatile and open software for comparing large genomes. *Genome Biol.*, **5,** R12.
20. Leach, J. E. & White, F. F. (1996) Bacterial avirulence genes. *Annu. Rev. Phytopathol.*, **34,** 153–179.
21. Lee, B. -M. et al. (2005) The genome sequence of *Xanthomonas oryzae* pathovar *oryzae* KACC10331, the bacterial blight pathogen of rice. *Nucleic Acids Res.*, **33**, 577–586.
22. Lowe, T. M. & Eddy, S. R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25,** 955–964.
23. Nelson, R. J. et al. (1994) Relationship between phylogeny and pathotype for the bacterial blight pathogen of rice. *Appl. Environ. Microbiol.*, **60**, 3275–3283.
24. Nierman, W. C. et al. (2004) Structual flexibility in the *Burkholderia mallei* genome. *Proc. Natl. Acad. Sci. USA*, **101**, 14246–14251.
25. Noda, T. et al. (1996) Pathogenic races of *Xanthomonas oryzae* pv. *oryzae* in South and East Asia. *JIRCAS J.*, **3,** 9–15.
26. Ochiai, H. et al. (2000) Genetic diversity of *Xanthomonas oryzae* pv. *oryzae* strains from Sri Lanka. *Phytopathology*, **90**, 415–421.
27. Ochiai, H. et al. (2001) Construction and characterization of a *Xanthomonas oryzae* pv. *oryzae* bacterial artificial chromosome library. *FEMS Microbiol. Lett.*, **200,** 59–65.
28. Petes, T. D. & Hill, C. W. (1988) Recombination between repeated genes in microorganisms. *Annu. Rev. Genet.*, **22**, 147–168.
29. Ray, S. K. et al. (2002) A high-molecular-weight outer membrane protein of *Xanthomonas oryzae* pv. *oryzae* exhibits similarity to non-fimbrial adhesins of animal pathogenic bacteria and is required for optimum virulence. *Mol. Microbiol.*, **46,** 637–647.
30. Sakiyama, T. et al. (2000) An automated system for genome analysis to support microbial whole-genome shotgun sequencing. *Biosci. Biotech. Biochem.*, **64,** 670–673.
31. Salanoubat, M. et al. (2002) Genome sequence of the plant pathogen *Ralstonia solanacearum*. *Nature*, **415,** 497–502.
32. Sasaki, T. et al. (2002) The genome sequence and structure of rice chromosome 1. *Nature*, **420,** 312–316.
33. Simpson, A. J. G. et al. (2000) The genome sequence of the plant pathogen *Xylella fastidiosa*. *Nature*, **406,** 151–157.
34. Stachel, S. E. & Zambryski, P. C. (1986) *virA* and *virG* control the plant-induced activation of the T-DNA transfer process of *Agrobaterium tumefaciens*. *Cell*, **46,** 325–333.
35. Wengelnik, K., Ackerveken, G. V. d. & Bonas, U. (1996) HrpG, a key hrp regulatory protein of *Xanthomonas campestris* pv. *vesicatoria* is homologous to two-component response regulators. *Mol. Plant-Microbe Interact.*, **9,** 704–712.
36. Wood, D. W. et al. (2001) The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. *Science*, **294**, 2317–2323.
37. Yada, T. & Hirosawa, M. (1996) Detection of short protein coding regions within the cyanobacterium genome: application of the hidden markov model. *DNA Res.*, **3,** 355–361.
38. Yang, Y., De Feyter, R. & Gabriel, D. W. (1994) Host-specific symptoms and increased release of *Xanthomonas citri* and *X. campestris* pv. *malvacearum* from leaves are determined by the 102-bp tandem repeats of *pthA* and *avrb6*, respectively. *Mol. Plant-Microbe Interact.*, **7**, 345–355.
39. Yang, Y. & Gabriel, D. W. (1995) Intragenic recombination of a single plant pathogen gene provides a mechanism for the evolution of new host specificities. *J. Bacteriol.*, **177**, 4963–4968.
40. Yap, I. V. & Nelson, R. J. (1996) WinBoot: A program for performing bootstrap analysis of binary data to determine the confidence limits of UPGMA-based dendrograms. *IRRI Discussion Paper Series 14*. International Rice Research Institute, Manila, the Philippines.
41. Yasuta, T. et al. (2001) DNA sequence and mutational analysis of rhizobitoxine biosynthesis genes in *Bradyrhizobium elkanii*. *Appl. Environ. Microbiol.*, **67,** 4999–5009.