Improved annotation of a plant pathogen genome *Xanthomonas oryzae* pv. *oryzae* PXO99A

Yang Lei,1, Shou-Kai Kang,1, Junxiang Gao, Xi-Shuai Jia and Ling-Ling Chen*

*aNational Key Laboratory of Agricultural Microbiology, College of Life Science and Technology, Center for Bioinformatics, Huazhong Agricultural University, No. 1 Shizishan Street, Hongshan District, Wuhan, 430070, P.R. China; \( ^b \)College of Science, Huazhong Agricultural University, No. 1 Shizishan Street, Hongshan District, Wuhan, 430070, P.R. China

Communicated by Ramaswamy H. Sarma

(Received 31 October 2011; final version received 4 May 2012)

Many bacterial genomes have been sequenced and stored in public databases now, of which Reference Sequence (RefSeq) is the most widely used one. However, the annotation in RefSeq is still unsatisfactory. The present analysis is focused on the re-annotation of an important plant pathogen genome *Xanthomonas oryzae* pv. *oryzae* PXO99A (Xoo PXO99A), which is the causal agent of bacterial blight on rice. Based on the parameters of 28 nucleotide frequencies and support vector machine algorithm, 41 originally annotated hypothetical genes were recognized as noncoding sequences, which were further supported by principal component analysis and other evidence. Ten of them were tested with reverse transcription-polymerase chain reaction experiments (RT-PCR), and all of them were confirmed to be non-coding sequences. Furthermore, 197 potential new genes not annotated in RefSeq were both recognized by two ab initio gene finding programs. Most of them only have sequence similarities with part of the known genes in other species, so they are unlikely to be protein-coding genes. Twelve potential new genes have high full-length sequence similarities with function-known genes, which are very likely to be true protein-coding genes. All the 12 potential genes were tested with RT-PCR, and 11 of them (92%) were successfully amplified in cDNA template. The RT-PCR experiments confirm that our theoretical prediction has high accuracy. The improvement of Xoo PXO99A annotation is helpful for the research of lifestyle, metabolism, and pathogenicity of this important plant pathogen. The improved annotation can be obtained from http://211.69.128.148/Xoo.

**Keywords:** nucleotide frequencies; hypothetical genes; principal component analysis (PCA); support vector machine (SVM); open reading frames (ORFs)

**Introduction**

With the rapid progress of prokaryotic genome sequencing projects, more than 1900 bacterial genomes have been sequenced up to the written of this paper, thus providing an unprecedented opportunity to study the genetics, biochemistry, and evolution of bacterial species. Such analyses strongly depend on accurate annotation. Most sequenced bacterial genomes are stored in public databases including National Center for Biotechnology Information (NCBI) GenBank (Benson, Boguski, Lipman, & Ostell, 1997), European Bioinformatics Institute (EBI, http://www.ebi.ac.uk/genomes/bacteria.html), J. Craig Venter Institute (JCVI) Comprehensive Microbial Resource (CMR, http://cmr.jcvi.org), etc. Furthermore, NCBI's Reference Sequence (RefSeq) database is generated to provide reference standards for genome annotation, which can provide more curated, nonredundant, and rich annotation for bacterial species (Pruitt, Tatusova, Klimke, & Maglott, 2009). The first step in bacterial genome annotation is to use gene finding programs, which scan the genome and identify potential protein-coding genes and other functional RNA products. The identified genes are then compared to public databases in order to identify related sequences. If hits of a certain similarity are identified, information about their function is transferred to the new sequence (Guo, Ou, & Zhang, 2003; Skovgaard, Jensen, Brunak, Ussery, & Krogh, 2001; Stothard & Wishart, 2006). Since most of the genes are identified with gene finding programs and not verified by experiments, there are still many problems in current bacterial genome annotation (Guo et al., 2003; Nielsen & Krogh, 2005; Skovgaard et al., 2001; Stothard & Wishart, 2006). Firstly, many genomes have false-positive gene identification, i.e. some open reading frames (ORFs) are...
incorrectly annotated as protein-coding genes, most of them are short ORFs (<150 bp) without functional information. Secondly, many annotated genes have incorrect translation initiation sites (TISs). Finally, some actual protein-coding genes are still missed in the current annotation (Guo et al., 2003; Nielsen & Krogh, 2005; Skovgaard et al., 2001; Stothard & Wishart, 2006).

_**Xanthomonas oryzae pv. oryzae** (Xoo)_ belongs to the gamma subdivision of _Proteobacteria_, which is the causal agent of bacterial blight (BB) on rice (Oryza sativa L.) (Lee et al., 2005; Nino-Liu, Ronald, & Bogdanove, 2006; Salzberg et al., 2008). BB is the most serious bacterial disease in tropical Asian countries where high yielding rice cultivars are often susceptible to this disease. In severely infected fields, BB can cause yield losses as high as 50% (Lee et al., 2005). Several factors that contribute to fitness and virulence of Xoo have been identified. Since rice is a major crop for Asian population and a model plant for cereal biology, a better understanding of pathogenesis caused by Xoo remains a pressing goal both for controlling of BB and for understanding of bacterial–plant interactions (Salzberg et al., 2008).

The Xoo PXO99A genome has a single circular chromosome of 5,240,075 bp. The original annotation was submitted to GenBank with the accession number CP000967, which contained 5083 protein-coding genes (Salzberg et al., 2008). Subsequently, a comprehensive annotation containing 4988 protein-coding genes was provided in NCBI RefSeq (NC_010717). The annotation of Xoo PXO99A in EBI bacteria genomes also contained 4988 genes. However, the annotation of Xoo PXO99A in CMR contains 5200 genes (http://cmr.jcvi.org/tigr-scripts/CMR/GenomePage.cgi?org=Xoo). Since Xoo PXO99A genome has quite different annotation in different databases, it is highly necessary to re-annotate it and provide more accurate annotation for related researchers.

**Materials and methods**

**Date collection**

The sequence and annotation of Xoo PXO99A genome were downloaded from NCBI RefSeq since it can provide a comprehensive and relatively precise annotation (Pruitt et al., 2009). The 4988 annotated protein-coding genes can be classified into two groups: the first contains 3452 genes with confirmed functions, and the second group contains 1536 genes with “putative,” “probable,” or “hypothetical” functions. The first 3452 genes have high sequence similarity with function-known genes in public databases, which were used for training data-set. Some of the 1536 genes with “putative,” “probable,” or “hypothetical” functions might not be protein-coding genes, whose coding status was identified in the current analysis. Furthermore, potential new protein-coding genes not occurred in the RefSeq annotation was predicted by two _ab initio_ gene finding programs, i.e. Prodigal (Hyatt et al., 2010) and FgenesB, (http://linux1.softberry.com/berry.phtml?topic=fgenesb&group=programs&subgroup=gfdb), respectively.

**Method for recognizing noncoding ‘hypothetical ORFs’**

The method adopted here was based on nucleotide frequencies. The algorithm was presented briefly as follows. The frequencies of bases A, C, G, and T occurring in an ORF or a fragment of DNA sequence with bases at positions 1, 4, 7, ..., 2, 5, 8, ..., 3, 6, 9, ..., were denoted by $a_1$, $c_1$, $g_1$, $t_1$; $a_2$, $c_2$, $g_2$, $t_2$; $a_3$, $c_3$, $g_3$, $t_3$ (12 parameters), respectively. In addition to the frequencies of phase-dependent single nucleotides, the frequencies of phase-independent di-nucleotides were also considered. Let the frequencies of the 16 di-nucleotides $aa$, $ac$, ..., and $tt$ of an ORF or a fragment of DNA sequence be denoted by $p(aa)$, $p(ac)$, ..., $p(tt)$ respectively. All the 28 components of the space $V$ were defined as follows:

$$
\begin{align*}
    u_1 &= a_1, u_2 = g_1, u_3 = c_1, u_4 = t_1 \\
    u_5 &= a_2, u_6 = g_2, u_7 = c_2, u_8 = t_2 \\
    u_9 &= a_3, u_{10} = g_3, u_{11} = c_3, u_{12} = t_3 \\
    u_{13} &= p(aa), u_{14} = p(ag), u_{15} = p(ac), u_{16} = p(at) \\
    u_{17} &= p(ga), u_{18} = p(gg), u_{19} = p(gc), u_{20} = p(gt) \\
    u_{21} &= p(ca), u_{22} = p(cg), u_{23} = p(cc), u_{24} = p(ct) \\
    u_{25} &= p(ta), u_{26} = p(tg), u_{27} = p(tc), u_{28} = p(tt)
\end{align*}
$$

Therefore, an ORF or a fragment of DNA sequence can be represented by a point or a vector in the 28-dimensional space $V$. Note that $u_i \in [-1, +1]$, $i = 1, 2, ..., 28$. The space $V$ was a 28-dimensional super-cube with the side length of 2.

To complete the algorithm, two groups of samples were needed. One was a set of positive samples corresponding to protein-coding genes, and the other was a set of negative samples corresponding to noncoding sequences. The two groups of samples constituted the training set used in the support vector machine (SVM) algorithm described below. Since coding bases account for about 85% of Xoo PXO99A genomes (Salzberg et al., 2008), it was rather difficult to prepare an appropriate set of noncoding sequences. Therefore, each of the 3452 protein-coding sequences were randomly shuffled 100,000 times, and then served as negative samples.

**Support vector machine**

The general principle of SVM was to perform a classification by constructing an n-dimensional hyper plane that optimally separates the positive and negative samples.
into two groups. SVM minimized the empirical classification error and maximized the geometric margin. The margin was defined as the distance from the separating hyper plane to its nearest sample. SVM was based on the structural risk minimization principle which allows the building of predictive models, even if descriptors were numerous and redundant. The structural risk was a trade-off factor between the training error and the model complexity. Moreover, SVM was not significantly affected by unbalanced active and inactive data-sets.

Radial basis function (RBF) is a commonly used kernel function with one parameter $\gamma$.

$$K(x_i, x_j) = e^{-\gamma|x_i-x_j|^2}$$  \hspace{1cm} (2)

For the selected kernel function, the learning task is to solving the following convex quadratic programming (QP) problem,

$$\max \sum_{i=1}^{N} a_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j y_i y_j K(x_i, x_j)$$  \hspace{1cm} (3)

Subject to:

$$0 \leq a_i \leq C, \quad \sum_{i=1}^{N} a_i y_i = 0,$$  \hspace{1cm} (4)

where labels, $y_i = +1, -1$, stand for the positive label and negative label, respectively. Tool package used in the current analysis is LIBSVM (Chang & Lin, 2001).

**Method for identifying new functional genes**

In this analysis, two ab initio gene finding programs were performed to identify new protein-coding genes not annotated in RefSeq. Prodigal (Prokaryotic dynamic programming gene-finding algorithm) was an recently developed highly accurate microbial gene finding program, which has high speed, low false-positive rate, and high accuracy in locating the TISs (Hyatt et al., 2010). FgenesB was another accurate ab initio prokaryotic gene prediction program, which is based on Markov chain models of coding regions, translation, and termination sites. FgenesB included simplified prediction of operons based only on distances between predicted genes. Combining the predicted result of the two ab initio programs, new protein-coding genes not annotated in RefSeq were identified in Xoo PXO99A genome.

**Strain cultivation and nucleic acid isolation**

*Xoo PXO99A* was grown in 5 mL nutrient agar (NA, 5 g polypeptone, 1 g yeast power, 3 g beef extract, 15 g sucrose per liter) at 28 °C for 24–36 h with 200 rpm shaking. Then, 250 μL of the suspension were added to 25 mL of NA and grown at 28 °C to logarithmic growth phase. The sodium dodecyl sulfate (SDS) method was used for DNA extraction. The total RNA was extracted by RiboPure™-Bacteria (Ambion) and treated with DNase I to remove genomic DNA contamination.

**PCR and sequence validation**

The total DNA, total RNA, and cDNA were used for polymerase chain reaction (PCR) analysis. The 50 μL PCR mixture contained 5 μL 10× PCR Buffer, 0.2 mM dNTP (Takara), 0.02 μM primers, 1 μL total DNA, total RNA or cDNA, 1 μL taq DNA polymerase (Takara), and nuclease-free water. The samples were incubated with the following cycles: 94 °C for 3 min, 30 cycles of 94 °C for 30 s, annealing for 30 s, 72 °C for 1 min, and a final extension of 72 °C for 10 min. First strand cDNA was synthesized using SuperScript™II RT (Invitrogen) and then amplified following the protocol as remarked above. The PCR primers for the chosen sequences were designed by Primer Premier 5.0 software (Premier BioSoft International, Palo Alto, CA). The PCR reaction condition for every primer set was optimized with DNA sample in repeated PCR experiments with 2–10 °C lower than the predicted annealing temperature until a single amplified band was obtained. 16S rRNA gene was used as positive controls for multicopy gene. Each of the PCR products was purified using the PCR products purification kits (QIAGEN). The purified DNA fragments were ligated with pMD®18-T Vector Systems (Takara) and transformed into competent cells of *Escherichia coli* DH5α. The positive clones were sequenced by Beijing AuGCT DNA-SYN Biotechnology Co., Ltd. (Beijing, China).

**Results and discussion**

**Identification of 41 noncoding ORFs**

In the process of hypothetical ORFs re-annotation, the 1536 ORFs with “putative,” “probable,” or “hypothe-

cical” functions were re-identified based on 28 parameters, including 12 single nucleotide frequencies and 16 di-

nucleotide frequencies mentioned above. Firstly, the 3452 function-known genes were randomly divided into two equal parts. The former served as a training set to calculate the discrimination parameters, and the latter served as a test set to assess the accuracy of the algo-

rithm. Both the training and the test sets should include positive and negative samples. In the genome of *Xoo PXO99A*, about 85% of the whole DNA sequences were coding and the remaining intergenic regions were domi-

nated by structural RNA sequences, so it was difficult to prepare an appropriate set of negative samples. Thus, the following procedures were taken to produce negative
samples. Each of the known genes was randomly shuffled 100,000 times, so that it was transformed into a random sequence. The shuffled sequences then served as negative samples. The performance of each test was measured by the following benchmark criteria, sensitivity ($s_n$), specificity ($s_p$), accuracy (Ac), and Matthew’s correlation coefficient (MCC):

$$s_n = \frac{TP}{TP + FN}, \quad s_p = \frac{TN}{TN + FP}, \quad Ac = \frac{TP + TN}{TP + TN + FP + FN}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}$$

where TP, TN, FP, and FN are fractions of true positive, true negative, false positive, and false negative predictions, respectively. After performing thirtyfold cross-validation tests, the mean sensitivity, specificity, accuracy and MCC, and the standard deviations were obtained (Table 1). The average prediction accuracy was as high as 99.5%. Then the 1536 hypothetical ORFs in Xoo PXO99A were re-identified. A total of 41 hypothetical ORFs were recognized as noncoding by all the thirtyfold cross-validation tests, which are listed in Table 2.

**Evidence of the 41 recognized ORFs as noncoding ORFs**

Since protein-coding genes can fold into stable and functional proteins, many constraints were imposed on protein-coding sequences. In most prokaryotic and eukaryotic species, protein-coding genes showed similar base usage patterns in the first and second codon positions, while the base usage in the third codon position was species specific (Chiusano et al., 2000; Gupta, Majumdar, Bhattacharya & Ghosh, 2000; Trifonov, 1987). Generally, the base usage pattern was RGN type, where R, G and N denoted purine, nonguanine, and any bases ate the first, second, and third codon positions, respectively (Chen & Zhang, 2003). It was observed that the first and second codon positions were related to the biosynthetic pathway and protein secondary structures (Chiusano et al., 2000; Gupta et al., 2000). On the other hand, the negative samples were shuffled sequences of function-known genes (random sequences which have the same nucleotide composition as protein-coding genes), so the base frequencies at each ‘codon’ position were almost equal. The difference of the base usage between protein-coding genes and noncoding sequences forms the basis of the present method to distinguish the two types of samples.

The difference between coding and noncoding sequences can be intuitively viewed by principal component analysis (PCA). PCA defines the correlation among the variables of given data. The first derived direction was chosen to maximize the standard deviation of the derived variable and the second was to maximize the standard deviation among directions uncorrelated with the first, and so forth (Dillon & Goldstein, 1984). Figure 1 shows the distribution of points on the principal plane spanned by the first two principal components. The coding and noncoding sequences were represented by open circles and triangles, respectively. The first and second principal axes possessed 25.7% and 17.4% of the total inertia of the 28-dimensional space. It was observed that the two principal axes can separate the coding and noncoding sequences into almost nonoverlapping clusters. The recognized noncoding ORFs were represented with filled stars, which distribute far from the core of function-known genes, and close to the random sequences, implying that the 41 ORFs listed in Table 2 were very unlikely to encode proteins.

The Clusters of Orthologous Groups (COGs) of proteins were involved in the RefSeq annotation. COG was a group of three or more proteins that were inferred to

### Table 1. The genome feature, sensitivity, specificity, and accuracy of thirtyfold cross-validation tests for Xoo PXO99A.

| Sensitivity (%) | Specificity (%) | Accuracy (%) | MCC |
|----------------|----------------|--------------|-----|
| 99.44 ± 0.0015| 99.58 ± 0.0015| 99.51 ± 0.001| 0.9903 ± 0.002 |

*The figure following ± is the standard deviation.

### Table 2. The PID of the 41 recognized noncoding ORFs by the thirtyfold cross-validation tests.

| PID          | PID          | PID          | PID          | PID          | PID          | PID          | PID          |
|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| 188,574,379  | 188,574,412  | 188,574,456  | 188,574,480  | 188,574,630  | 188,574,729  | 188,574,750  |
| 188,574,782  | 188,574,830  | 188,575,054  | 188,575,087  | 188,575,134  | 188,575,236  | 188,575,309  |
| 188,575,596  | 188,575,621  | 188,575,749  | 188,575,894  | 188,575,981  | 188,576,038  | 188,576,084  |
| 188,576,269  | 188,576,412  | 188,576,417  | 188,576,472  | 188,576,799  | 188,576,804  | 188,576,847  |
| 188,576,848  | 188,576,991  | 188,576,997  | 188,577,039  | 188,577,040  | 188,577,763  | 188,577,763  |
| 188,578,759  | 188,578,764  | 188,578,923  | 188,578,946  | 188,579,038  | 188,579,110  | 188,579,318  |
be orthologs, i.e. they have evolved from a common ancestor (Tatusov et al., 2003). Analysis of complete bacterial genomes showed that prokaryotic proteins were generally highly conserved, with about 70% of them containing ancient conserved regions shared by homologs from distantly related species (Tatusov et al., 2003). Therefore, an ORF within a COG was highly likely to be a protein-coding gene, which has homologs from other species. In the 3452 function-known genes, 2875 contained COG codes, however, none of the 41 recognized noncoding ORFs contained COG codes. Since most of the over annotation are short ORFs, the average length of the 3452 function-known genes in the first group and the 41 recognized noncoding ORFs were compared. The average length of the recognized noncoding ORFs (189 bp) was much shorter than that of the function-known genes (1012 bp, Table 3). In addition, the ab initio gene finding programs Prodigal and FgenesB have not identified the 41 ORFs. Based on these facts, it was highly possible that the 41 “hypothetical genes” annotated in Xoo PXO99A genome were not protein-coding genes but over-annotated short ORFs.

To test our theoretical prediction, 10 noncoding ORFs were randomly selected to perform RT-PCR experiments. Information of the 10 noncoding ORFs and the designed primers are listed in Table 4 and the PCR results are shown in Fig. 2(A)–(C). The RT-PCR using total DNA as template confirmed that all the DNA segments could be amplified precisely (Fig. 2(A)). In Fig. 2(C), water was used as template for detecting if the total RNA sample had DNA contamination or not, and negative result was obtained. Fig. 2(B) showed the RT-PCR result of cDNA as template. It was observed that except for the positive control of 16S rRNA gene (for multicopy gene), all the tested noncoding ORFs were not successfully amplified, confirming that they were true noncoding ORFs. The DNA sequencing results confirmed that the PCR products were the correct target gene sequences (data not shown). The RT-PCR results verified that the theoretical prediction accuracy was very high.

**Newly predicted protein coding genes not annotated in NCBI RefSeq**

The ab initio programs employed in NCBI RefSeq annotation included Glimmer (Delcher, Bratke, Powers, & Salzberg, 2007), GeneMarks (Besemer & Borodovsky, 2005), and recently developed Prodigal (Hyatt et al., 2010). However, the Prodigal prediction result has not incorporated into the RefSeq annotation. FgenesB was another accurate gene prediction program not used in RefSeq annotation. In this analysis, Prodigal and FgenesB were used to find new protein-coding genes not present in RefSeq annotation in Xoo PXO99A genome. Prodigal and FgenesB predicted 4980 and 5074 ORFs in Xoo PXO99A genome, respectively. Since each gene finding program has its false-positive prediction, ORFs predicted simply by Prodigal or FgenesB were regarded as false-positive ORFs. The Venn Diagram of Prodigal, FgenesB, compared with JCVI CMR annotation is shown in Figure 3. The three systems predicted 4543 ORFs in common, 4346 of them were contained in NCBI RefSeq annotation, and the other 197 ORFs were not included. Since the above three systems commonly predicted these 197 ORFs, and the method proposed in this paper can also recognize these ORFs, they were highly possible to be protein-coding genes but missed in RefSeq annotation.

Then BLAST search was performed to find potential functions for these ORFs, the prediction results can be

---

**Table 3.** The average length and percentage of ORFs with COG code for 3452 function-known protein-coding genes and 41 recognized noncoding ORFs in Xoo PXO99A.

| Feature                  | 3452 genes with known functions | 41 recognized noncoding ORFs |
|--------------------------|---------------------------------|------------------------------|
| Number of genes with COG| 2875                            | 0                            |
| Average length (bp)      | 1012                            | 189                          |

Figure 1. The distribution of points on the principal plane spanned by the first (x) and second (y) principal axes using the PCA in Xoo PXO99A. The open circles denoted function-known genes, the open triangles represented the corresponding negative samples and the filled stars denoted the hypothetical ORFs recognized as noncoding sequences. The first and second principal axes accounted for 25.7% and 17.4% of the total inertia of the 28-dimensional space, respectively. Note that the distribution of the open circles was well separated from that of the open triangles, indicating that coding and noncoding sequences were well distinguished. Furthermore, most of the identified noncoding ORFs distributed far from the core of protein-coding genes, and closed to the random sequences, implying that the 41 recognized noncoding ORFs listed in Table II were very unlikely to encode proteins.
classified into several groups. Sixty-nine ORFs predicted to encode 25 transposases and 44 truncated transposases (Supplementary Table I). The original RefSeq annotation of Xoo PXO99A genome contained 765 transposases, adding the above new annotated transposases, Xoo PXO99A encoded more than 820 transposases, accounting for about 16% of its gene number, which was much higher than the average 1.1% transposase content in bacterial genome (Aziz, Breitbart, & Edwards, 2010). Twenty-nine ORFs encoding hypothetical proteins or had no sequence similarity in public databases (Supplementary Table II). Eighty-seven ORFs were predicted with concrete functions (Supplementary Table III). It was very surprising that many ORFs encoded truncated genes, and sometimes several neighboring genes could align to different part of the same homolog gene in other bacteria (the alignment part was listed in the “identity” column in Supplementary Table III). Comprehensive investigation showed that there were transposases flanking most of the truncated genes. Transposases were enzymes that bind to the ends of transposons and catalyze the movement of transposons to other parts of the genome by cut-and-paste mechanism or replicative transposition mechanism, without the need for homology between transposons and the new DNA target sites (Curcio & Derbyshire, 2003). As mentioned above, Xoo PXO99A contained about

Table 4. The information of 10 noncoding sequences and corresponding primers.

| Primer name | PID     | Annealing temperature (°C) | Primer sequence                  |
|-------------|---------|-----------------------------|----------------------------------|
| 16S rRNA-F  | -       | 55                          | 5’AGAGTTTGATCCTGGCTCAG3’         |
| 16S rRNA-R  |         |                             | 5’GGTTAATGTTCGAACCACCTA3’        |
| NC-1-F      | 188,574,630 | 55                     | 5’CAGCGCAATGACAGCCATCAG3’        |
| NC-1-R      |         |                             | 5’GGCCGCAGTACAGCAGCTT3’          |
| NC-2-F      | 188,575,134 | 56                     | 5’AAATATTGTGCAGCGAAGCG3’         |
| NC-2-R      |         |                             | 5’TTCCTTACAGCGCGCTTGTT3’         |
| NC-3-F      | 188,578,318 | 56                     | 5’ATGGAAACGCAAGCCGTCAG3’         |
| NC-3-R      |         |                             | 5’TCCGACTTGCTTTTTGCCCA3’         |
| NC-4-F      | 188,576,804 | 57                     | 5’AACACACGCAGCTTTAGAA3’          |
| NC-4-R      |         |                             | 5’CTGGACTTGCTTTTTGCCCA3’         |
| NC-5-F      | 188,578,946 | 20                     | 5’GGGATGCTGCTCAAGCCTAA3’         |
| NC-5-R      |         |                             | 5’CGCCGATCTTGGTGAGCA3’           |
| NC-6-F      | 188,575,759 | 57                     | 5’ATGGATCAGCGACAAAGGCTCA3’       |
| NC-6-R      |         |                             | 5’AGTGATCGCAATCTGCTTT3’          |
| NC-7-F      | 188,574,456 | 54                     | 5’CATCCCTTCCTACGACATT3’          |
| NC-7-R      |         |                             | 5’TGGACTAGCCTGTCATGCT3’          |
| NC-8-F      | 188,547,480 | 55                     | 5’TGAAGATGCTGTCAACGTCAA3’        |
| NC-8-R      |         |                             | 5’TAATCTAAAGAAGCCGGGCCAG3’       |
| NC-9-F      | 188,574,729 | 55                     | 5’ATGGGAAATAGGGGGAGCA3’          |
| NC-9-R      |         |                             | 5’TACACACATGTGCAAGCCAT3’         |
| NC-10-F     | 188,575,087 | 53                      | 5’ATGGTAAACGACGACGACTAG3’        |
| NC-10-R     |         |                             | 5’TATGGCAGCGCATCTGCT3’           |

Table 5. The 12 newly predicted complete function-known genes in Xoo PXO99A genome.

| Start … Stop | Strand | E value | Query length (aa) | Subject length (aa) | Identity (%) | Product                                      |
|-------------|--------|---------|-------------------|---------------------|--------------|----------------------------------------------|
| 135,036 … 135,758 | - | 2e−136  | 240               | 240                 | 99            | ATP-binding protein of ABC transporter       |
| 455,492 … 455,917 | - | 2e−75   | 141               | 141                 | 99            | LysR family transcriptional regulator       |
| 955,017 … 957,914 | + | 1.0      | 965               | 965                 | 100           | Truncated beta-galactosidase                |
| 2,148,219 … 2,149,142 | + | 4e−172  | 307               | 316                 | 99            | ABC transporter ATP-binding protein         |
| 3,011,885 … 3,012,619 | − | e−138   | 244               | 244                 | 99            | MlrB                                         |
| 3,647,997 … 3,648,284 | − | 8e−47   | 92                | 92                  | 100           | Integral membrane protein                   |
| 4,132,639 … 4,134,276 | + | 0       | 545               | 552                 | 99            | Transcriptional regulator                   |
| 4,234,496 … 4,234,735 | + | e−37    | 79                | 79                  | 99            | Putative transposase                        |
| 4,567,717 … 4,568,469 | − | 3e−140 | 250               | 250                 | 99            | Transcuded serine protease                  |
| 4,703,171 … 4,704,424 | + | 0       | 417               | 417                 | 100           | Histidine kinase                            |
| 4,839,183 … 4,840,163 | − | 5e−168  | 326               | 313                 | 99            | Arabinose efflux permease                   |
| 4,840,470 … 4,840,799 | + | 2e−54   | 109               | 113                 | 100           | Transcriptional regulator l lysR family     |
16% transposase genes, which was the most abundant in current sequenced bacteria genomes.

| Primer Name | START | STOP | Annealing temperature (°C) | Primer sequence |
|-------------|-------|------|---------------------------|-----------------|
| 16S rRNA-F  | –     | –    | 55                        | 5′AGAATTTGATCTTGCTGAG3′ |
| 16S rRNA-R  | –     | –    | 55                        | 5′GGTTACCTTTTACACTT3′ |
| PC-1-F      | 135,036 | 135,758 | 54                  | 5′ATATGACTGCGGTTACCT3′ |
| PC-1-R      | –     | –    | 54                        | 5′ATCCAGGTATTGCTGCTGAG3′ |
| PC-2-F      | 2,148,219 | 2,149,142 | 53               | 5′ATGGATGACCCGAAATGGA3′ |
| PC-2-R      | –     | –    | 53                        | 5′ACCGAAAACGCCAGAAAAGAT3′ |
| PC-3-F      | 3,011,885 | 3,012,619 | 55              | 5′ATCTTGCAAGCGAGCTAATC3′ |
| PC-3-R      | –     | –    | 55                        | 5′TGCGATCTTGAGATGATTT3′ |
| PC-4-F      | 4,132,639 | 4,134,276 | 52          | 5′GGTTGATCCGCTGTTACAAG3′ |
| PC-4-R      | –     | –    | 52                        | 5′ACTGGTGCTGCGCTGAGAA3′ |
| PC-5-F      | 4,567,717 | 4,568,469 | 53                  | 5′CCATTCTTGGAAGACACCA3′ |
| PC-5-R      | –     | –    | 53                        | 5′AGCATTCCACACCACCCGATA3′ |
| PC-6-F      | 4,703,171 | 4,704,424 | 53              | 5′ATTGCCGACACCCGCAA3′ |
| PC-6-R      | –     | –    | 53                        | 5′ATCGATGCGCATGAGCTAC3′ |
| PC-7-F      | 4,839,183 | 4,840,163 | 54          | 5′CTGCAATGCGCCGACATCT3′ |
| PC-7-R      | –     | –    | 54                        | 5′ATAAGACCACGTCAGCCCTT3′ |
| PC-8-F      | 455,492 | 455,917 | 54                  | 5′CAATCAATGGCAAACGGAT3′ |
| PC-8-R      | –     | –    | 54                        | 5′ATTGCCGACACCCGCAA3′ |
| PC-9-F      | 955,017 | 957,914 | 55              | 5′ATCGATGCGCATGAGCTAC3′ |
| PC-9-R      | –     | –    | 55                        | 5′CTGCAATGCGCCGACATCT3′ |
| PC-10-F     | 3,647,997 | 3,648,284 | 50         | 5′TAATAGTGCAGCCGCGCTCT3′ |
| PC-10-R     | –     | –    | 50                        | 5′AGGCGGCTGCCAGAGATTCT3′ |
| PC-11-F     | 4,234,496 | 4,234,735 | 50     | 5′GCAGTTCAGAATGTGTAAGAA3′ |
| PC-11-R     | –     | –    | 50                        | 5′TTGCAGCACAATGGAATTG3′ |
| PC-12-F     | 4,840,470 | 4,840,799 | 50  | 5′ACGGTGTTGACCGAGGAAG3′ |
| PC-12-R     | –     | –    | 50                        | 5′ACGGTGTTGACCGAGGAAG3′ |

Figure 2. (A)–(C) The PCR results of the DNA fragments re-annotated as noncoding ORFs. (A) The expected products of PCR used total DNA as a template were all obtained with the right sizes, 16S rRNA gene (1204 bp), NC-1 (428 bp), NC-2 (440 bp), NC-3 (236 bp), NC-4 (288 bp), NC-5 (327 bp), NC-6 (175 bp), NC-7 (184 bp), NC-8 (196 bp), and NC-9 (195 bp). (B) In the RT-PCR with cDNA as templates, the positive controls of 16S rRNA gene (for multicopy gene) was obtained. The tested samples were all got negative amplification results as predicted. (C) When the total RNA was used as templates in the PCR, the products were all failed for amplification.

Figure 3. Venn diagram of ORFs predicted by Prodigal, FgenesB, and JCVI CMR. The three systems predicted 4543 ORFs in common, 4346 of them were contained in NCBI RefSeq annotation, whereas the other 197 ORFs were not included in NCBI RefSeq annotation, they were very likely to be protein-coding genes.

16% transposase genes, which was the most abundant in current sequenced bacteria genomes. Xoo PXO99A genome had multiple copies of different transposase families.
that might be involved in maintaining genetic diversity through genome rearrangements. Since these genes were truncated functional genes, most of them might not encode proteins. In addition, the other two sequenced Xoo strains, KACC10331 and MATF311018, contain 662 (16%) and 675 (15%) transposases, respectively. Salzberg et al. (2008) also stated that most of the rearrangements between Xoo PXO99A and MATF311018 were mediated by a diverse set of transposable elements.

Table 5 listed the remaining potential 12 protein-coding genes, which had high sequence identity and alignment coverage with function-known genes. Thus, they were very likely to be new functional genes in Xoo PXO99A genome. All the 12 protein-coding genes were tested with RT-PCR experiments (Table 6), and the results were shown in Fig. 4(A)–(C). The RT-PCR using total DNA as template confirmed that all the DNA segments could be amplified precisely (Fig. 4(A)). Fig. 4(C) showed that the total RNA sample had no DNA contamination. Fig. 4(B) showed that when cDNA as template, 11 (92%) of the tested genes were successfully amplified except PC-2. The RT-PCR results verified that most of the 12 predicted potential new genes were true protein-coding genes.

**Conclusion**

The present analysis focused on the re-annotation of plant pathogen Xoo PXO99A, which was the causal agent of bacterial blight on rice. Forty-one originally annotated hypothetical genes were recognized as noncoding sequences, which were supported by PCA, other theoretical evidence and RT-PCR experiment. In addition, 12 potential protein-coding genes were predicted with functions, and 11 (92%) were verified to be true protein-coding genes with RT-PCR experiment. Another important finding was that diversities among different Xoo strains were mainly caused by high amount of transposases. The improved annotation of Xoo PXO99A genome will benefit the research of lifestyle, metabolism, and pathogenicity of this important plant pathogen.

**Acknowledgements**

The present study was supported by the National Natural Science Foundation of China (31071659), the National Basic Research Program of China (2012AA10A304) and the Fundamental Research Funds for the Central Universities (2011PY071). We would like to express our thanks to Dr. Lifang Ruan for providing Xanthomonas oryzae pv. oryzae PXO99A and technical assistance.

**Supplementary material**

The supplementary material for this paper is available online at http://dx.doi.10.1080/07391102.2012.698218.

**References**

Aziz, R.K., Breitbart, M., & Edwards, R.A. (2010). Transposases are the most abundant, most ubiquitous genes in nature. Nucleic Acids Research, 38, 4207–4217.

Benson, D.A., Boguski, M.S., Lipman, D.J., & Ostell, J. (1997). GenBank. Nucleic Acids Research, 25, 1–6.

Besemer, J., & Borodovsky, M. (2005). GeneMark: Web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Research, 33, W451–W454.

Chang, C.C., & Lin, C.J. (2001). LIBSVM: A library for support vector machines. Retrieved from http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chen, L.L., & Zhang, C.T. (2003). Gene recognition from questionable ORFs in bacterial and archaeal genomes. Journal of Biomolecular Structure & Dynamics, 21, 99–110.

Chiusano, M.L., Alvarez-Valin, F., Di Giulio, M., D’Onofrio, G., Ammirato, G., Colonna, G., & Bernardi, G. (2000). Second codon positions of genes and the secondary structures of proteins. Relationships and implications for the origin of the genetic code. Gene, 261, 63–69.

Crammer, K., & Singer, Y. (2002). On the algorithmic implementation of multiclass kernel-based vector machines. Journal of Machine Learning Research, 2, 265–292.

Curcio, M.J., & Derbyshire, K.M. (2003). The outs and ins of transposition: From Mu to Kangaroo. Nature Reviews Molecular Cell Biology, 4, 865–877. doi: 10.1038/nrm1241
Delcher, A.L., Bratke, K.A., Powers, E.C., & Salzberg, S.L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, 23, 673–679.

Dillon, W.R., & Goldstein, M. (1984). *Multivariate analysis – methods and applications*. Wiley series in probability and mathematical statistic. New York: Wiley.

Guo, F.B., Ou, H.Y., & Zhang, C.T. (2003). ZCURVE: A new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Research*, 31, 1780–1789.

Gupta, S.K., Majumdar, S., Bhattacharya, T.K., & Ghosh, T.C. (2000). Studies on the relationships between the synonymous codon usage and protein secondary structural units. *Biochemical and Biophysical Research Communications*, 269, 692–696.

Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F. W., & Hauser, L.J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11, 119.

Lee, B.M., Park, Y.J., Park, D.S., Kang, H.W., Kim, J.G., Song, E.S., ... Go, S.J. (2005). The genome sequence of *Xanthomonas oryzae* pathovar oryzae KACC10331, the bacterial blight pathogen of rice. *Nucleic Acids Research*, 33, 5775–5786.

Nielsen, P., & Krogh, A. (2005). Large-scale prokaryotic gene prediction and comparison to genome annotation. *Bioinformatics*, 21, 4322–4329.

Nino-Liu, D.O., Ronald, P.C., & Bogdanove, A.J. (2006). Xanthomonas oryzae pathovars: model pathogens of a model crop. *Molecular Plant Pathology*, 7, 303–324.

Pruitt, K.D., Tatusova, T., Klimke, W., & Maglott, D.R. (2009). NCBI Reference Sequences: Current status, policy, and new initiatives. *Nucleic Acids Research*, 37, D32–D36.

Salzberg, S.L., Sommer, D.D., Schatz, M.C., Phillippy, A.M., Rabinowicz, P.D., Tsuge, S., ... Bogdanove, A.J. (2008). Genome sequence and rapid evolution of the rice pathogen *Xanthomonas oryzae* pv. oryzae PXO99A. *BMC Genomics*, 9, 204.

Skovgaard, M., Jensen, L.J., Brunak, S., Ussery, D., & Krogh, A. (2001). On the total number of genes and their length distribution in complete microbial genomes. *Trends in Genetics*, 17, 425–428.

Stothard, P., & Wishart, D.S. (2006). Automated bacterial genome analysis and annotation. *Current Opinion in Microbiology*, 9, 505–510.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., ... Natale, D.A. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4, 41.

Trifonov, E.N. (1987). Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16S rRNA nucleotide sequences. *Journal of Molecular Biology*, 194, 643–652.