Phytopathology • 2020 • 110:969-972 • https://doi.org/10.1094/PHYTO-01-20-0012-A

e-Xtra*

Complete Circularized Genome Data of Two Spanish strains of *Xylella fastidiosa* (IVIA5235 and IVIA5901) Using Hybrid Assembly Approaches

Luis F. Arias-Giraldo,[1] Annalisa Giampetruzzi,[2] Madis Metsis,[3] Ester Marco-Noales,[4] Juan Imperial,[5] María P. Velasco-Amo,[1] Miguel Román-Écija,[1] and Blanca B. Landa[1,†]

[1] Institute for Sustainable Agriculture, Consejo superior de Investigaciones Científicas (CSIC), Córdoba, Spain
[2] Dipartimento di Scienze del Suolo della Pianta e degli Alimenti, Università degli Studi di Bari Aldo Moro, Bari, Italy
[3] Testsystems OU, Tallinn, Estonia
[4] Centro de Protección Vegetal y Biotecnología, Instituto Valenciano de Investigaciones Agrarias (IVIA), Moncada, Spain
[5] Institute of Agricultural Sciences, CSIC, Madrid, Spain

Abstract

Xylella fastidiosa is an economically important plant pathogenic bacterium of global importance associated, since 2013, with a devastating epidemic in olive trees in Italy. Since then, several outbreaks of this pathogen have been reported in other European member countries including Spain, France, and Portugal. In Spain, the three major subspecies (subsp. *fastidiosa, multiplex*, and *pauca*) of the bacterium have been detected in the Balearic Islands, but only subspecies *multiplex* in the mainland (Alicante). We present the first complete genome sequences of two Spanish strains: *X. fastidiosa* subsp. *fastidiosa* IVIA5235 from Mallorca and *X. fastidiosa* subsp. *multiplex* IVIA5901 from Alicante, using Oxford Nanopore and Illumina sequence reads, and two hybrid approaches for genome assembly. These completed genomes will provide a resource to better understand the biology of these *X. fastidiosa* strains.

Funding
This work was funded by European Union's Horizon 2020 Framework Research Programme Projects XF-ACTORS (Xylella fastidiosa Active Containment Through a Multidisciplinary Oriented Research Strategy grant 727987) and POnTE (Pest Organisms Threatening Europe grant 635646), and Desarrollo de estrategias de erradicación, contención y control de Xylella fastidiosa en España: Diagnóstico, estructura genética y gama de huéspedes (Project E-RTA2017-00004-C06-02) from Programa Estatal de I+D+I Orientada a los Retos de la Sociedad of the Spanish Government and COST Action CA16107 EuroXanth supported by European Cooperation in Science and Technology.

Keywords
bacteriology, hybrid assembly, Illumina, Oxford Nanopore, population biology, *Xylella fastidiosa*

Genome Announcement

The plant pathogen *Xylella fastidiosa* has been associated with various recent epidemics in Europe including Italy, France, and Spain, affecting agricultural crops such as almond, grapevine, and olive, but also endemic species occurring in natural forest landscapes and ornamental plants (Landa et al. 2018). The first identified case of *X. fastidiosa* in Spain was reported in Mallorca Island in 2016, associated with cherry trees (*Prunus avium*) (Landa et al. 2018). Later on, in 2017, *X. fastidiosa* subsp. *multiplex* ST6 was detected for the first time in mainland Spain, in almond trees in the Guadalest Valley (Alicante province, Valencian Community) (Giampetruzzi et al. 2019).

In a previous study, a strain of *X. fastidiosa* subsp. *fastidiosa* belonging to sequence type 1 (ST1) and designated IVIA5235 was isolated from an infected cherry tree in Mallorca and its draft genome sequence, including that of its native plasmid (pXFAS_5235), were obtained (Landa et al. 2018). This strain presented a high genome sequence similarity to that of an already characterized American strain, *X. fastidiosa* subsp. *fastidiosa* M23 (NC_010577), harboring a similar plasmid (NC_010579). In a different study, the draft genome sequences of two strains (named ESVL and IVIA5901) isolated from almond trees infected by *X. fastidiosa*

†Corresponding author: B. B. Landa; blanca.landa@csic.es

*The e-Xtra logo stands for "electronic extra" and indicates that one supplementary table is published online.

The author(s) declare no conflict of interest.

Accepted for publication 21 February 2020.

Table 1. Relevant features of the complete genome assemblies of Spanish *Xylella fastidiosa* strains IVIA5235 and IVIA5901

Strain	Reference	Software[a]	Number of contigs	Sequence type	Size (bp)	GC content (%)	Number of CDS genes[b]	Number of rRNA	Number of tRNA	Noncoding RNAs
IVIA5235	Landa et al. 2018	SPAdes	106	Chromosome	2,491,574	51.60	2,258	6	48	4
		+ PGAAP	1	Plasmid	38,297	49.21	42	0	0	
	This study	Unicycler	2	Chromosome	2,537,917	51.73	2,812	6	48	
		+ RAST		Plasmid	38,297	49.21	43	0	0	
	This study	Unicycler	2	Chromosome	2,537,917	51.73	2,211	6	49	3
		+ PGAAP		Plasmid	38,297	49.21	41	0	0	
IVIA5901	Giampetruzzi et al. 2019	SPAdes + PGAAP	141	Chromosome	2,493,558	51.8	2,236	6	51	4
	This study	Flye/RaconILx2	2	Chromosome	2,559,157	52.0	2,867	6	47	
		+ RAST		Phage	35,839	56.9	46	0	0	
	This study	Flye/RaconILx2	2	Chromosome	2,559,157	52.0	2,268	6	49	3
		+ PGAAP		Phage	35,839	56.9	46	0	0	

[a] PGAAP = NCBI Prokaryotic Genome Automatic Annotation Pipeline. Unicycler hybrid mode = combined use of SPAdes for genome assembly, *Pilon* for polishing with Illumina reads, and *dnaA* gene search for genome circularization with BlastX.
[b] Protein encoding genes.

subsp. *multiplex* ST6 were obtained (Giampetruzzi et al. 2019). The draft assemblies of those Spanish isolates are currently incomplete (over 100 contigs). Whole genome sequencing methods based on short-read sequencing platforms such as Illumina do not always allow complete sequence assembly and circularization of bacterial chromosomes, especially for chromosomes with a high number of repetitive elements in their sequences, a situation observed for *X. fastidiosa* genomes belonging to subspecies *multiplex* in self-alignment dot plots.

In recent years, with the arrival of long-read sequencing platforms, several studies have explored the combined use of short, but accurate, Illumina reads, with long, but less accurate, reads generated by Oxford Nanopore Technologies (ONT) and/or Pacific Biosciences (PacBio) sequencing platforms to create high-quality full genome reconstructions (Sović et al. 2016). In this work, we sequenced the genomes of *X. fastidiosa* subsp. *fastidiosa* ST1 strain IVIA5235 and *X. fastidiosa* subsp. *multiplex* ST6 strain IVIA5901 using ONT, and combined the ONT reads with the previous sequenced Illumina libraries. The use of hybrid assembly approaches greatly improved the quality of the final assembly and yielded complete circularized assemblies of their genomes. Contrary to partial genome sequences, complete genome assemblies allow detail examination of minor structural genomic alterations among different *X. fastidiosa* isolates, and its characterization may be critical for epidemiological studies ongoing in different outbreak areas of Europe.

Both strains were grown as pure cultures in PD2 agar medium, and DNA extraction was performed using the DNeasy kit (Qiagen). ONT sequencing libraries were prepared using the transposase-based rapid sequencing kit following manufacturer's recommendations and generated in-house using an R9.4 flow cell with the MinION device. Each library was sequenced in a standard run protocol of ~3 h for IVIA5235 and ~16 h for IVIA5901. The preprocessing of the raw data generated by MinION, which includes the filtering and trimming steps, was performed with *Filtlong* v0.2.0 (https://github.com/rrwick/Filtlong) and *Porechop* v0.2.3 (https://github.com/rrwick/Porechop), removing adapters and reads shorter than 2 kb. Previously obtained WGS data (2 × 150 bp paired-end libraries sequenced with a HiSeq4000 platform) (Landa et al. 2018; Giampetruzzi et al. 2019) for both strains consisted of 5,008,800 reads for IVIA5235 (NCBI BioProject PRJNA488161) and 3,956,773 reads for IVIA5901 (NCBI BioProject PRJNA482385). Both high-throughput datasets were combined using two different de novo assembly pipelines. We evaluated the open-source assemblers *Unicycler* v0.4.7 (Wick et al. 2017) and *Flye* v2.6 (Kolmogorov et al. 2019) that were run with the default options or following the steps recommended in their respective manuals. In order to evaluate the quality of the assembly, the polished genomes were analyzed with *CheckM* v1.0.13 (Parks et al. 2015) using the lineage-specific workflow (lineage_wf) applied to each strain. The starting positions of circular genomes were fixed by the *dnaA* gene using *Circlator* v1.5.5 (Hunt et al. 2015). For functional annotation and classification using subsystems, genomes were submitted to the *RAST Server* (Rapid Annotation using Subsystem Technology) (Aziz et al. 2008; Brettin et al. 2015) and to the *PGAAP*, the NCBI Prokaryotic Genome Automatic Annotation Pipeline (Haft et al. 2018; Tatusova et al. 2016). We uploaded the fasta genome files to the *PHASTER* server (http://phaster.ca/) to identify phage-related genes and their integrity.

Total reads obtained after quality control of the raw data generated by ONT were 7,754 and 291,275 reads, with a mean length of 5.8 and 4.2 kb, and equivalent to a depth of genome coverage of ~17× and ~470× for IVIA5235 and IVIA5901, respectively. Compared with currently available draft assemblies for those strains (106 and 141 contigs for IVIA5235 and IVIA5901, respectively) (Table 1), hybrid assemblies allowed us to obtain the complete chromosome sequences for both strains, although the approaches followed were different. Thus, using the *Unicycler* hybrid approach, the complete genome assembly of strain IVIA5235 was obtained in two circular contigs, a chromosome of 2.53 Mb and a plasmid of 38.3 kb (Table 1). However, with even greater sequence coverage, this approach did not yield a complete assembly for strain IVIA5901, obtaining a total of eight noncircularized bacterial contigs (three of them ranging from 1,430 to 393 kb, and the rest ranging from 36 to 1 kb), and a circularized contig of 35.84 kb. The failure in obtaining a complete assembly for this strain was probably due to the high number of long repetitive sequences, many of them phage sequences, present in its genome. Instead, for this later strain, we used a long-read-only assembly with *Flye*, which was run with the flags *–asm-coverage 40* and *–genome-size 2.6m* to reduce coverage using a subset of the longest reads for initial assembly and to estimate the final size of the genome assembly. This was followed by a polishing step with Illumina reads using *Racon*, which resulted in a successful genome assembly of IVIA5901 into two contigs: a complete circular chromosome of 2.56 Mb and a circular contig of 35.84 kb (Table 1). This second contig was identified as a bacteriophage sequence that presents 94.25% average nucleotide identity with the genome of *Xylella* phage *xfas53* (NC_013599.1), and likely belongs to the same species.

CheckM estimated the completeness of both genomic assemblies at 99.64% using the specific set of markers for the C_gammaproteobacteria lineage consisting of 481 genes. The polishing step carried out by both approaches considerably improved the quality of the genome assemblies. The *Unicycler* pipeline underwent a total of nine rounds of polishing with *Pilon*, until no changes could be made to the genome of IVIA5235 obtaining a completeness of 99.64%, while *Racon* did not require more than two rounds to increase the completeness score of IVIA5901 genome to 99.64%.

Table 1 shows a summary of features for both assemblies. For IVIA5235, RAST annotation pipeline resulted in a total of 2,909 features (coding sequences [CDS] and RNA sequences) and 241 subsystems, while for IVIA5901, a total of 2,920 features (CDS and RNA sequences) and 243 subsystems were obtained. In both cases, most genes were associated with amino acids, cofactors and related derivatives, protein processing, and energy metabolism. An alternative annotation pipeline, PGAAP, resulted in 2,268 CDS and 58 RNA sequences for IVIA5235, and 2,211 CDS and 58 RNA sequences for IVIA5901. As with RAST annotation, most PGAAP annotated genes were associated with amino acids, cofactors and related derivatives, protein processing, and energy metabolism.

PHASTER identified eight prophage regions in the genome of IVIA5901, of which six regions were intact and two regions were questionable, with no incomplete regions; likewise, nine prophage regions were identified in the genome of IVIA5235, of which four regions were intact, two regions were incomplete, and three regions were questionable (Supplementary Table S1).

Data availability. The complete genome sequence of *X. fastidiosa* subsp. *multiplex* IVIA5901 has been deposited in GenBank under accession number CP047134 and the phage sequence under accession number MT084350. The complete chromosome and plasmid sequences of *X. fastidiosa* subsp. *fastidiosa* IVIA5235 have been deposited in GenBank under accession numbers CP047171 (chromosome) and CP047172 (plasmid pXFAS_5235). Accession numbers are associated to BioProject numbers PRJNA488161 and PRJNA482385 for strains IVIA5235 and IVIA5901, respectively. Both strains are deposited in the Spanish Type Culture Collection (CECT).

Literature Cited

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., et al. 2008. The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics 9:75.

Brettin, T., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Olsen, G. J., Overbeek, R., Parrello, B., Pusch, G. D., Shukla, M., Thomason, J. A., III, Stevens, R., Vonstein, V., Wattam, A. R., and Xia, F. 2015. RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Sci. Rep. 5:8365.

Giampetruzzi, A., Velasco-Amo, M. P., Marco-Noales, E., Montes-Borrego, M., Román-Écija, M., Navarro, I., Monterde, A., Barbe, S., Almeida, R. P. P., Saldarelli, P., Saponari, M., Montilon, V., Savino, V. N., Boscia, D., and Landa, B. B. 2019. Draft genome resources of two strains ("ESVL" and "IVIA5901") of *Xylella fastidiosa* associated with almond leaf scorch disease in Alicante, Spain. Phytopathology 109: 219-221.

Haft, D. H., DiCuccio, M., Badretdin, A., Brover, V., Chetvernin, V., O'Neill, K., Li, W., Chitsaz, F., Derbyshire, M. K., Gonzales, N. R., Gwadz, M., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Yamashita, R. A., Zheng, C., Thibaud-Nissen, F., Geer, L. Y., Marchler-Bauer, A., and Pruitt, K. D. 2018. RefSeq: An update on prokaryotic genome annotation and curation. Nucleic Acids Res. 46:D851-D860.

Hunt, M., Silva, N. D., Otto, T. D., Parkhill, J., Keane, J. A., and Harris, S. R. 2015. Circlator: Automated circularization of genome assemblies using long sequencing reads. Genome Biol. 16:294.

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. 2019. Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. 37:540-546.

Landa, B. B., Velasco-Amo, M. P., Marco-Noales, E., Olmo, D., López, M. M., Navarro, I., Monterde, A., Barbe, S., Montes-Borrego, M. Roman-Ecija, M., Saponari, M., and Giampetruzzi, A. 2018. Draft genome sequence of *Xylella fastidiosa* subsp. *fastidiosa* strain IVIA5235, isolated from *Prunus avium* in Mallorca Island, Spain. Microbiol. Resour. Announc. 7:e01222-e18.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. 2015. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25:1043-1055.

Sović, I., Križanović, K., Skala, K., and Šikić, M. 2016. Evaluation of hybrid and non-hybrid methods for de novo assembly of nanopore reads. Bioinformatics 32:2582-2589.

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., Lomsadze, A., Pruitt, K. D., Borodovsky, M., and Ostell, J. 2016. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res. 44:6614-6624.

Wick, R. R., Judd, L. M., Gorrie, C. L., and Holt, K. E. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLOS Comput. Biol. 13:e1005595.