SHORT TERM SCIENTIFIC MISSION (STSM) SCIENTIFIC REPORT

This report is submitted for approval by the STSM applicant to the STSM coordinator

Action number: CA16107
STSM title: "*Xanthomonas arboricola* pv. *juglandis* genomics and comparative genomics"
STSM start and end date: 19/02/2018 to 23/02/2018
Grantee name: Camila Borges Fernandes

PURPOSE OF THE STSM

Xanthomonas arboricola pv. *juglandis* (*Xaj*) is the bacteria phytopathogen responsible for diseases outbreaks on walnut orchards often associated with economic important losses for walnut producing worldwide (Frutos 2010; Lamichhane 2014). Efficient implementation of disease control and prevention measures are needed but is difficult when epidemiological knowledge is limited.
The purpose of this Short Term Scientific Mission (STSM) was to unveil the main genomic differences of the xanthomonads population found within a single walnut tree host. Starting from the raw sequence data from five genomes recently sequenced, the main research interest was to determine dissemination fitness of different *Xaj* lineages and unveil epidemiological pattern.
This STSM was also extremely important to have training in genomics analyses, being possible to provide an introduction to some of dedicated bioinformatics software tools used for filter comprehensively large datasets from whole-genome sequencing.

DESCRIPTION OF WORK CARRIED OUT DURING THE STSMS

The work programme of this STSM included *de novo* genome assemblies, automatic annotations, sequence alignments for genome comparison and evolutionary analysis of the five genomes newly sequenced. First, the total number of reads for each sample (Genome Sequencer Illumina HiSeq, 2×150bp paired-end reads, **Table 1**) were *de novo* assembled using SeqMan NGen from the Lasergene genomics package (version 12.1.0; DNAStar, Madison, WI).

Table 1 – Genomic library raw data information.

Sample Code	Strain	Sequenced reads	Sequenced bases	%GC
XANT1	CPBF 367	12,218,570	3,665,571,000	65.6
XANT2	CPBF 424	12,672,550	3,801,765,000	65.7
XANT3	CPBF 426	9,383,710	2,815,113,000	65.7
XANT4	CPBF 427	11,009,360	3,302,808,000	65.1
XANT5	CPBF 1521	10,113,730	3,034,119,000	64,9

Comparison of contig sequences was performed in order to assemble multiple contigs into a single one, using SeqMan Pro (DNAStar) under default parameters values through the method Pro Assemblers. Consistent connections between different contigs were edited manually and paired-end reads information was also used to evaluate reassemble. Misassigned bases were identified with script executed to map single nucleotide polymorphisms (SNPs) using contig coordinates. All SNP reports were manually checked and base substitutions in each position were performed using another custom script. The resulting set of contigs for each genome (CPBF 367, CPBF 424, CPBF 426, CPBF 427 and CPBF 1521) were manually reviewed to detect the presence of similar region at the 5' and 3' ends using the Gepard software (Krumsiek et al. 2007). Additionally, each set of contigs was also rearranged to a defined start position taking into account the location of the chromosomal replication initiator protein (DnaA) associated to bacterial replication origins

(Mackiewicz et al. 2004). Contigs were searched against the aminoacid sequence of DnaA protein of *Xanthomonas campestris* pv. campestris ATCC 33913 (AAM39320) through BLAST analysis using tblastn. Furthermore, contigs from each genome were ordered and oriented according to the related reference genome of *Xanthomonas arboricola* pv. *juglandis* (strain CFBP 2528, NZ_JZEF00000000.1) using the Move Contigs function in MAUVE Version 2.4.0 (Darling et al. 2004). Annotations were performed with Prokka (Seemann 2014) based on *de novo* discovery of genes using a database of *Xanthomonas* spp.. Furthermore, a BLAST database was created with protein sequences previously known to be involved in pathogenesis and virulence of bacterial pathogens, mainly in *Xanthomonas spp.* genus. Homologs to proteins associated with xanthan gum production, the flagellar system, components of different secretion systems (T2SS, T3SS and T4SS) and type III effectors (T3Es) were predicted in the five genomes using tblastn. Comparative genomics analysis was initiated using progressive MAUVE software with default parameters. Genome alignment between the five genomes and the reference CFBP 2528 was performed in order to detect conserved gene clusters and genome rearrangements. BLAST Ring Image Generator (BRIG) (Alikhan et al. 2011) was also used to visualize the whole genome similarity, determined with blastn, between the reference genome of *Xaj* strain and the five Portuguese genomes. At last, the average nucleotide identity (ANI) (Konstantinidis and Tiedje 2005) was calculated to infer robust genomic relatedness between the five genomes, together with published genomes of *Xanthomonas* spp., including genomes from *X. arboricola* species. Based on multiple alignment of shared genes, the total similarities are defined in terms of the amount of sequence identity and two genomes may be considered as belonging to the same species if they share ANI values higher than 95% (Gupta and Sharma 2015).

DESCRIPTION OF THE MAIN RESULTS OBTAINED

1. *De novo* assembly, contigs editing and ordering results:

The *de novo* genome assembly resulted in 15 to 53 contigs per genome with similar GC contents (66%) and a total size between 3,934,004 and 5,199,703 bp. **Table 2** summarizes the most important genome metrics.

Table 2 – Genome assembly metrics

Strain	Contigs	Genome size (bp)	Max/Min contig length (bp)	Mean contig length (bp)	GC (%)	Unresolved (inc. Ns)	Ns	N_{50} (bp)	N_{90} (bp)	L_{50}
CPBF 367	37	4,935,637	1,816,398/274	129,885	65.8	5	0	1,347,650	322,220	1
CPBF 424	18	4,341,324	1,191,592/248	228,491	65.8	5	0	637,211	260,759	2
CPBF 426	15	4,085,974	1,219,198/246	255,373	66.1	3	0	669,305	284,111	2
CPBF 427	21	3,934,004	758,932/251	178,818	65.7	2	0	396,511	158,209	3
CPBF 1521	53	5,199,703	694,552/245	96,291	65.4	2	0	338,137	67,693	5

Through Gepard software, no visual evidence of circularization was observed, except for a single contig in CPBF 424 representing repeated sequence at the ends of the contig. Nucleotide BLAST (blastn) of such contig showed no indication of plasmid-related genes and revealed that sequence is probably located on the bacterial chromosome.

The putative replication origin was found in the five genomes studied. Highly conserved sequence (query coverage – 100%, Identity – 93%, *e*-value 0.0) for DnaA protein was identified and the start of Contig 1 on its forward strand was defined for the five genomes.

2. Genome annotation results:

A total of 4’075, 3’570, 3’356, 3’220 and 4’351 coding sequences (CDSs) was detected for CPBF 367, CPBF 424, CPBF 426, CPBF 427 and CPBF 1521 genomes, respectively. No plasmid was found in the sequenced strains as previously predicted using Gepard. Ribosomal RNA, transfer RNA features and other miscellaneous RNA were also identified in the five genomes (**Table 3**).

Table 3 – General features of the five annotated genomes

Strain	No. of CDSs	No. of genes	rRNA	tmRNA	tRNA	Misc_RNA
CPBF 367	4,075	4,186	5	1	58	47
CPBF 424	3,570	3,661	2	1	51	37
CPBF 426	3,356	3,450	3	1	51	39
CPBF 427	3,220	3,310	3	0	37	50

CPBF 1521	4,351	4,479	4	1	55	68

No differences were observed for the *gum* genes involved in xanthan gum synthesis. The flagellar system seems to be incomplete in CPBF 426 and CPBF 427 genomes due the absence of genes related with bacterial flagella motility (*motA* and *motB*). Nevertheless, major differences were found in the repertoire of T3SS and T3E genes among the five genomes (**Table 4**).

Table 4 – Presence and absence of secretion systems (T2SS, T3SS and T4SS) and type III effectors (T3Es) across the five genome studied.

		Absent in:				
	Present in all	**CPBF 367**	**CPBF 424**	**CPBF 426**	**CPBF 427**	**CPBF 1521**
T2SS	*xcsN*		*xpsD, xpsE*		*xpsD, xpsE*	
T3SS		*hpa2, hpaB, hrcC, hrcQ, hrcR, hrcT, hrcU, hrpB1, hrpB2, hrpB4, hrpB5, hrpB7, hrpD5, hrpD6, hrpF, hrpB3, hrpB6, hrpB8, hrpD1, hrpD2, hrpD3, hrpW, YscR*		*hpa2, hpaB, hrcC, hrcQ, hrcR, hrcT, hrcU, hrpB1, hrpB2, hrpB4, hrpB5, hrpB7, hrpD5, hrpD6, hrpF, hrpG, hrpX, hrpB3, hrpB6, hrpB8, hrpD1, hrpD2, hrpD3, hrpW, YscR*	*hrpF*	
T4SS	*virB10, virB11, virB2, virB4, virB8, virD4,*	*virB6*	*virB6*	*virB6*		
T3E	*xopR*	*hpaA, xopF1, xopF2, xopZ2, avrBs2, avrXccA2, xopG, xopK, xopL, xopN, xopQ, xopV, xopX, xopZ1, xopC, xopE*	*avrBs2, avrXccA2, xopG, xopK, xopL, xopN, xopQ, xopV, xopX, xopZ1, xopC, xopE*	*hpaA, xopF1, xopF2, xopZ2, avrBs2, avrXccA2, xopG, xopK, xopL, xopN, xopQ, xopV, xopX, xopZ1, xopC, xopE*	*xopZ2, xopC, xopE4*	*xopZ2*

3. Comparative genomic analysis results using MAUVE and BRIG:

Multiple alignment of the genomes allowed the identification of regions that are conserved and unique among genomes (**Figure 1**). Most of the rearranged and deleted regions found in the alignments are associated with insertion sequences elements. In addition, the number of transposable elements and integrase genes detected through MAUVE was different among genomes. The highest number was found in CPBF 1521 genome, with 82 transposases and 10 integrases whereas in CPBF 426 it was only found seven transposases and three integrases.

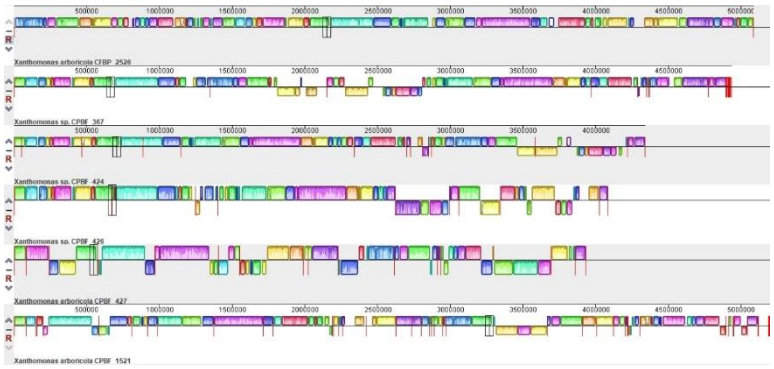

Figure 1 - MAUVE alignment results of the genome sequences of reference *Xaj* strains CFBP 2528 (row1) with genomes of isolates CPBF 367 (row2), CPBF 424 (row3), CPBF 426 (row4), CPBF 427 (row5) and CPBF 1521 (row6). Boxes with identical colours across the genomes, represent homologous genomic regions. Inside each box, the average level of conservation of sequence across genomes is shown with vertical bars: conserved and highly related regions are colored and low-identity unique regions are in white (colorless). Regions outside coloured boxes lack detectable homology among the genomes.

The overall sequence similarity among genomes was easily observed with the graphical output obtained from the BRIG analysis (**Figure 2**). Despite high identity between genomes, some genomic regions of CFBP 2528 are absent among the other genomes.

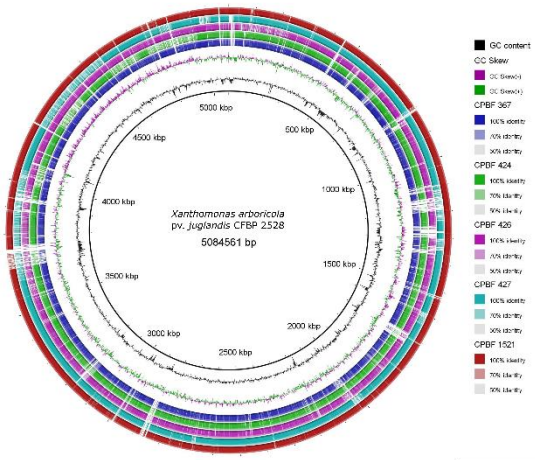

Figure 2 - Circular representation of blastn results obtained for the five genomes, CPBF 367 (blue ring), CPBF 424 (green ring), CPBF 426 (purple ring), CPBF 427 (light blue ring) and CPBF 1521 (red ring), using the genome of *Xaj* CFBP 2528 as reference (the inner black ring). The second and third innermost rings shown GC content and GC skew, respectively. Identity values are represented according a graduated scale showed on the right of the image. Genomic regions of CFBP 2528 with no blastn matches appeared as blank spaces across ring of the other genomes.

4. Average nucleotide identity results:

ANI values higher than 95% was observed for the isolates CPBF 427 and CPBF 1521 when genome sequences was compared with reference genomes of *X. arboricola* strains, being more closely related to *X. arboricola* strains of pathovar *juglandis* sharing ANI values higher than 97%. Conversely, the other three isolates (CPBF 367, CPBF 424 and CPBF 426) showed ANI values lower than 94% when compared to the strains of *X. arboricola*, being only closely related to each other. In addition, another ANI analysis was performed with *Xanthomonas* strains belong to the species *X. cassavae*, *X. gardneri*, *X. hortorum* and *X. fragariae*. Isolates CPBF 367, CPBF 424 and CPBF 426 showed always values ≤ than 90% when compared to the other *Xanthomonas* species.

FUTURE COLLABORATIONS

Analysis of the obtained information is still in progress. The quality of *de novo* assemblies performed with SeqMan Pro are being compared with *de novo* assemblies obtained with MIRA v.4.0 in order to check and correct misassemblies. The genome sequence information will be further compared based on functional studies in order to understand the co-colonization of the five *Xaj* clonal lineages within the same plant host. Furthermore, the work started during this STSM will be published in an international scientific journal with peer review, together with the ongoing outcomes, and will be discussed at the second EuroXanth annual conference in Germany in July 2018 which will be linked with the 6th Xanthomonas Genomics Conference.

REFERENCES

Alikhan, N., F. Petty, N., K. Ben Zakour, N., L. Beatson, S., A (2011). “BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons”. BMC Genomics 12: 402 - 412.

Darling, A., C. Mau, B. Blattner, F., R. Perna, N., T (2004) “Mauve: multiple alignment of conserved genomic sequence with rearrangements”. Genome Research 14: 1394 -1403.

Frutos, D. (2010). "Bacterial diseases of walnut and hazelnut and genetic resources." Journal of Plant Pathology 92(1): S79-S85.

Gupta, A. Sharma, V., K. (2015). “Using the taxon-specific genes for the taxonomic classification of bacterial genomes”. BMC Genomics, 16(1), 396 - 405.

Konstantinidis, K., T. Tiedje, J., M (2005) “Genomic insights that advance the species definition for prokaryotes”. Proc Natl Acad Sci U S A 102: 2567-2572.

Krumsiek, J. Arnold, R. Rattei, T (2007) “Gepard: a rapid and sensitive tool for creating dotplots on genome scale”. Bioinformatics 23:1026-1028.

Lamichhane, J. R. (2014). “*Xanthomonas arboricola* diseases of stone fruit, almond, and walnut trees: Progress toward understanding and management.” Plant Disease 98(12): 1600-1610.

Mackiewicz, P. Zakrzewska-Czerwińska, J. Zawilak, A. Dudek, M., R. Cebrat, S (2004) “Where does bacterial replication start? Rules for predicting the *oriC* region”. Nucleic Acids Research 32(13): 3781-3791.

Seemann, T. (2014) “Prokka: rapid prokaryotic genome annotation”. Bioinformatics 30: 2068-2069.