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 campestris* pv. *campestris* genomics and comparative genomics"
STSM start and end date: 03/02/2020 to 07/02/2020
Grantee name: Eliška Peňázová

PURPOSE OF THE STSM:

Xanthomonas campestris pv. *campestris* (Xcc) is a causal agent of black rot disease on a wide range of Brassicaceae plants including vegetable crops, ornamental crucifers as well as weeds. The disease is described through the world and is considered to be one of the most destructive diseases of cruciferous plants.
The main purpose of this Short Term Scientific Mission (STSM) was to obtain complete genomes of six isolates of bacterium Xcc from the raw data generated by whole genome sequencing (MiniSeq, Illumina) and their comparison as a part of the study of epidemiology of Xcc in the Czech Republic. The other aim of the STSM was to gain the ability to evaluate sequencing data by the pipeline appropriate for bacterial genomes and to learn the basics in bioinformatics regarding genomic analysis (commands, software, related tools for description of bacterial genomes).

DESCRIPTION OF WORK CARRIED OUT DURING THE STSMS

The work was performed with six natural strains of Xcc originated from different regions of the Czech Republic. The identity of all isolates was tested by MALDI-TOF MS approach and the results were evaluated by the SARAMIS (Spectral ARchiveAnd Microbial Identification System) application and the database PAPMID (Putative Assigned Protein Masses for Identification Database).
The quality of data generated by MiniSeq instrument (2×150 bp) was examined by FastQC software and the *de novo* assembly of reads was performed using the Unicycler assembler. For visualization of results, the Bandage software was used.
To achieve complete genome sequences, the sequencing of isolates by Oxford Nanopore technology was added. High molecular weight genomic DNA isolated by Puregene Yeast/Bact. Kit B (Qiagen) was used to prepare libraries using the 1D native barcoding genomic DNA kit (Oxford Nanopore) according to the manufacturer's instructions. The library were then sequenced for 16 hours on a MinION device. Obtained data were basecalled and demultiplexed according to the used barcodes with Guppy. The data from MiniSeq and MinION instruments were combined in hybrid *de novo* assembly using Unicycler. Final genomic sequences were visualized in Bandage. As the presence of plasmids was evidenced and Unicycler failed in detecting the start codon of plasmid partition protein (e.g. ParA, RepA), blastx analysis were performed and the sequences of plasmids were rearranged according to this start position when found. For chromosomal sequence, the presence of CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) regions and type III effectors were evaluated.
The whole genome similarity of Xcc genomes was subsequently visualized in MAUVE and BRIG (Blast Ring Image Generator) software. For BRIG, the Xcc reference sequence (NC_003902) was downloaded from GenBank/NCBI database. The ANI (average nucleotide identity) expressing the percentage of sequence identity between two genomes were counted by the JSpecies WS platform using Mummer (ANIm). At the end of the STSM, the annotation of genomes was done using Prokka. The usage

of comparative genomics tools such as the EDGAR and OrthoVenn platforms was also introduced.

DESCRIPTION OF THE MAIN RESULTS OBTAINED

1. MALDI-TOFMS analysis

Six strains of bacterium Xcc were tested by whole cell MALDI-TOF MS approach. The first analysis confirmed the contamination of one strain when the obtained profile corresponded to the spectrum profile of the *Bacillus* genus. From the spectra obtained for remaining Xcc strains (four replicates), the similarity tree was constructed in SARAMIS. The results showed two clusters when only the isolate MEND-B-000002 appeared in a distance from the others. Isolates did not form separate branches reflecting the geographical location and the tree displayed an overall similarity of all *Xanthomonas* strains.

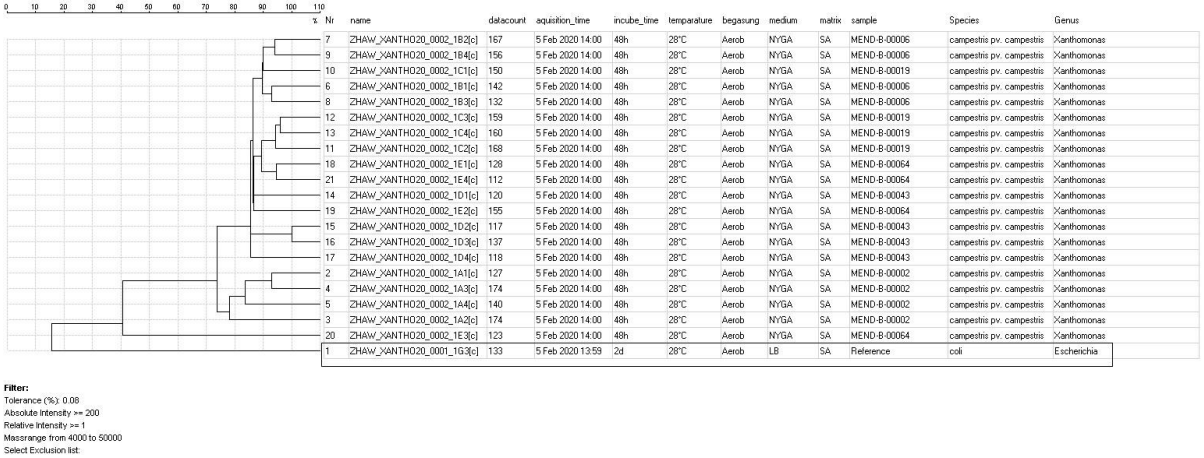

Fig. 1 The similarity tree of Xcc isolates based on MALDI-TOF mass spectra.

2. *De novo* assembly (Illumina, Nanopore and hybrid assembly)

De novo assembly of Illumina reads for five Xcc genomes led between 117 to 175 contigs with an average genome length around 5.09 Mb. Due to the visualization of reads in Bandage software, the total genome length and the lower average GC content, the presence of a contamination in the isolate MEND-B-00019 was found. Based on this result, subsequent work with this isolate was done only with the Nanopore data that did not show any contamination. For other isolates, the hybrid assembly was used. The most important genome metrics generated for hybrid assembly are stated in Table 1. The sequence of whole circular chromosome was achieved for two isolates (MEND-B-00006 and MEND-B-00043). For the isolate MEND-B-00019, the new library and sequencing on MiniSeq instrument will be done to complete current data. The presence of plasmid was confirmed for two isolates using BLAST (blastx database) and their sequences were arranged to start with the starting codon of ParA protein.

Table 1 Genome assembly metrics – hybrid assembly (Illumina and Nanopore).

Strain	*Number of contigs*	*Total genome length (bp)*	*GC %*	*N_{50} (bp)*	*L_{50}*	*N_{90} (bp)*	*Note*
MEND-B-00002	1	5'118'546	65.20	5'118'546	NA	NA	Non-circular chromosome
MEND-B-00006	2	5'044'795	65.05	4'966'863	NA	77'932	Closed, circular chromosome, starts with *dnaA*; one circular plasmid, starts with *parA*
MEND-B-00019 (contaminated)	-	9'859'975	60.31	2'139'058	1	208'351	Non-circular chromosome
MEND-B-00043	1	5'034'241	65.10	5'034'241	NA	NA	Closed, circular chromosome, starts with *dnaA*
MEND-B-00064	2	5'157'558	64.97	5'117'651	NA	39'907	Non-circular; one circular plasmid, starts with *parA*

3. Comparative genomic analysis

The similarities of genome sequences were analyzed with MAUVE and BRIG software. Both tools showed differences between individual strains and also with the Xcc reference genome when using BRIG. Regarding the reference sequence, the higher variation was observed for isolates MEND-B-00002, MEND-B-00006, MEND-B-00019 and MEND-B-00064 than for the isolate MEND-B-00043 (Fig. 2).

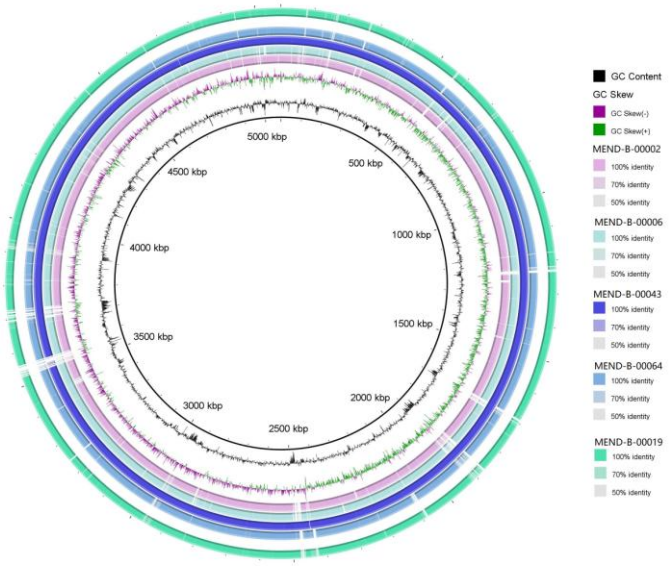

Fig. 2 Genome similarities of five Xcc strains showed by BRIG software using the genome sequence of Xcc ATCC 33913 as a reference.

4. CRISPR, type III effectors and ANI evaluation

The presence of CRISPR regions was evaluated using the online tool CRISPR finder. Sequences connected with repeated regions were analyzed by BLAST and except of one isolate the identity with ice nucleation protein (InaX) was proved. The presence of type III effectors was confirmed using tblastx analysis; the data are still being further analyzed.

The ANIm values higher than 97 % were calculated for all sequence combinations which corresponds with the MALDI-TOF MS results. Also the similarities in range of 97 to 99 % were observed when obtained sequences were compared with the Xcc reference.

5. Genome annotation

In all genomes, more than 4'000 coding sequences (CDS) were detected (Table 2). Results of annotation show the presence of six ribosomal RNA, one transfer-messenger RNA and 55-59 transfer RNA in each genome. Only isolate MEND-B-00002 contained a CRISPR array as detected by minced.

Table 2 Genome annotation results

Strain	*Contigs*	*Number of bases*	*Number of CDS*	*Number of genes*	*rRNA*	*tRNA*	*tmRNA*	*mics_RNA*	*Repeat region*
MEND-B_00002	1	5'118'546	4'347	4'526	6	56	1	116	1
MEND-B-00006	2	5'044'795	4'276	4'447	6	55	1	109	0
MEND-B-00019	2	5'117'056	4'353	4'525	6	57	1	108	0
MEND-B-00043	1	5'034'241	4'227	4'399	6	57	1	108	0
MEND-B-00064	2	5'157'558	4'437	4'608	6	59	1	105	0

FUTURE COLLABORATIONS (if applicable)

The future collaboration will contain the sequencing of isolates MEND-B-00019 on MiniSeq and MEND-B-0001 on MinION devices to obtained whole genome sequences of six Xcc isolates and the completion of analysis described above. Obtained data will be incorporated to the scientific article evaluating the epidemiology and genomic variability of the bacterium Xcc in the Czech Republic and published in cooperation of both institutions. The application of bilateral project reflecting SNSF-GAČR call was discussed and the project proposal is now being prepared.