

# Phytopathology

## Strain-level identification of bacterial tomato pathogens directly from metagenomic sequences

Journal:	<i>Phytopathology</i>
Manuscript ID	PHYTO-09-19-0351-R.R1
Manuscript Type:	Research
Date Submitted by the Author:	n/a
Complete List of Authors:	<p>Mechan Llontop, Marco; Virginia Tech, School of Plant and Environmental Sciences</p> <p>Sharma, Parul; Virginia Tech, School of Plant and Environmental Sciences; Virginia Tech, Graduate program in Genomics, Bioinformatics, and Computational Biology</p> <p>Aguilera Flores, Marcela; Virginia Tech, School of Plant and Environmental Sciences; Virginia Tech, Graduate program in Genomics, Bioinformatics, and Computational Biology</p> <p>Yang, Shu ; Virginia Tech, School of Plant and Environmental Sciences</p> <p>Pollock, Jill; Virginia Tech, School of Plant and Environmental Sciences; Virginia Tech, Eastern Shore Agricultural Research and Extension Center</p> <p>Tian, Long; Virginia Tech, School of Plant and Environmental Sciences; Virginia Tech, Graduate program in Genomics, Bioinformatics, and Computational Biology</p> <p>Huang , Chengjie ; Virginia Tech, Computer Sciences</p> <p>Rideout, Steven; Virginia Tech, School of Plant and Environmental Sciences; Virginia Tech, Eastern Shore Agricultural Research and Extension Center</p> <p>Heath, Lenwood; Virginia Tech, Computer Science</p> <p>Li, Song; Virginia Tech, School of Plant and Environmental Sciences</p> <p>Vinatzer, Boris; Virginia Tech, School of Plant and Environmental Sciences</p>
Keywords:	Bacteriology, Disease control and pest management, Techniques
<p>Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.</p> <p>Supplementary Table 4 revised.xlsx</p>	

1  
2  
3 **1 Strain-level identification of bacterial tomato pathogens directly from metagenomic**  
4  
5 **2 sequences**  
6

7  
8  
9  
10 4 Marco E. Mechan Llontop<sup>1\*</sup>, Parul Sharma<sup>1,2\*</sup>, Marcela Aguilera Flores<sup>1,2\*</sup>, Shu Yang<sup>1</sup>, Jill Pollok<sup>1,3</sup>,  
11 5 Long Tian<sup>1</sup>, Chenjie Huang<sup>4</sup>, Steve Rideout<sup>1,3</sup>, Lenwood S. Heath<sup>4</sup>, Song Li<sup>1</sup>, Boris A. Vinatzer<sup>1</sup>  
12  
13  
14 6

15  
16 7 <sup>1</sup> School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA  
17

18 8 <sup>2</sup> Graduate program in Genetics, Bioinformatics, and Computational Biology, Virginia Tech,  
19 Blacksburg, VA  
20 9

21  
22 10 <sup>3</sup> Virginia Tech Eastern Shore Agricultural Research and Extension Center, Painter, VA, USA  
23

24 11 <sup>4</sup> Department of Computer Sciences, Virginia Tech, Blacksburg, VA  
25  
26  
27 12

28 13 \*These authors contributed equally  
29  
30  
31 14

32  
33 15 Corresponding authors: Boris A. Vinatzer and Song Li  
34

35 16 E-mail addresses: [vinatzer@vt.edu](mailto:vinatzer@vt.edu) [songli@vt.edu](mailto:songli@vt.edu)  
36

37 17 Phone number: +1 540 231 2126  
38

39 18 B.A. Vinatzer ORCID: 0000-0003-4612-225X  
40  
41 19  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

## 20 Abstract

21 Routine strain-level identification of plant pathogens directly from symptomatic tissue could  
22 significantly improve plant disease control and prevention. Here we tested the Oxford Nanopore  
23 Technologies (ONT) MinION™ sequencer for metagenomic sequencing of tomato plants either  
24 artificially inoculated with a known strain of the bacterial speck pathogen *Pseudomonas syringae*  
25 *pv. tomato (Pto)*, or collected in the field and showing bacterial spot symptoms caused by either  
26 one of four *Xanthomonas* species. After species-level identification using ONT's WIMP software  
27 and the third party tools Sourmash and MetaMaps, we used Sourmash and MetaMaps with a  
28 custom database of representative genomes of bacterial tomato pathogens to attempt strain-level  
29 identification. In parallel, each metagenome was assembled and the longest contigs were used  
30 as query with the genome-based microbial identification Web service LINbase. Both the read-  
31 based and assembly-based approaches correctly identified *Pto* strain T1 in the artificially  
32 inoculated samples. The pathogen strain in most field samples was identified as a member of  
33 *Xanthomonas perforans* group 2. This result was confirmed by whole genome sequencing of  
34 colonies isolated from one of the samples. Although in our case, metagenome-based pathogen  
35 identification at the strain-level was achieved, caution still needs to be exerted when interpreting  
36 strain-level results because of the challenges inherent to assigning reads to specific strains and  
37 the error rate of nanopore sequencing.

38

## 39 Introduction

40 Early detection of plant disease outbreaks and accurate plant disease diagnosis are prerequisites  
41 of efficient plant disease control and prevention (Tinivella et al. 2008). In many cases, an  
42 experienced plant pathologist can quickly diagnose a disease based on symptoms. However,  
43 visual diagnosis does not identify the causative agent at the strain-level. For example, three  
44 different strains of the plant pathogen *Pseudomonas syringae* pathovar (pv.) *tomato* (*Pto*) cause  
45 indistinguishable bacterial speck disease symptoms in tomato (Cai et al. 2011). Sometimes, visual  
46 diagnosis cannot even identify a pathogen at the species level. For example, four different species  
47 of the genus *Xanthomonas* cause indistinguishable bacterial spot disease symptoms on tomato  
48 (*Solanum lycopersicum*) leaves (Jones et al. 2004). Note that in this article, we use the term  
49 “strain” as an intraspecific, monophyletic group of bacteria, which have a very recent common  
50 ancestor and are thus genotypically and phenotypically more similar to each other than to other  
51 members of the same species (Dijkshoorn et al. 2000). To avoid confusion, we use the term  
52 “isolate” instead of “strain” when referring to a pure culture of bacteria isolated on a specified date  
53 at a specified geographic location from a specific plant.

54 While most disease control measures may be the same for different pathogen strains or  
55 species, depending on the precise identity of the pathogen, additional control measures may need  
56 to be undertaken. For example, different strains of the same pathogen species may have different  
57 host ranges. Therefore, it may be necessary to avoid certain crop rotations or to eliminate certain  
58 weeds depending on the identity of the strain that causes a disease and its specific host range.  
59 In the case of *Pto*, strain T1 causes disease only in tomato while strain DC3000 causes disease  
60 in tomato and in leafy greens of the family *Brassicaceae* (Yan et al. 2008). Strain DC3000 could  
61 thus spread from tomato fields to leafy green fields, cause disease in a leafy green planted after  
62 tomato, and/or survive in weeds that belong to the *Brassicaceae* family. In other cases, identifying  
63 a pathogen to strain level could even trigger eradication procedures to stop further spread of the  
64 disease. For example, this would happen if the select agent *Ralstonia solanacearum* Race 3

1  
2  
3 65 Biovar 2 were to be identified as the causative agent of bacterial wilt disease outbreak in the USA  
4  
5 66 (Williamson et al. 2002). Fast strain-level plant pathogen identification would thus add significant  
6  
7 67 value to plant disease diagnostics.  
8

9 68 Many molecular tools have been developed over the years for pathogen identification and  
10  
11 69 they all have their strengths and weaknesses (Fang and Ramasamy 2015). Many of them depend  
12  
13 70 on a pure pathogen culture and thus require lengthy procedures to isolate and culture the  
14  
15 71 pathogen from the plant tissue. Moreover, many of them cannot identify pathogens at the strain  
16  
17 72 level. Gene sequence-based techniques, such as multilocus sequence typing/analysis (MLST/A)  
18  
19 73 (Almeida et al. 2010), can identify a pathogen to strain-level but usually require pure cultures.  
20  
21 74 Moreover, gene sequence-based techniques depend on previous species-level identification  
22  
23 75 because different species require different primers to amplify the genes to be sequenced by  
24  
25 76 polymerase chain reaction (PCR), for example see (Rees-George et al. 2010). One alternative  
26  
27 77 gene-based method is to amplify the 16S rRNA gene directly from DNA extracted from plant tissue  
28  
29 78 and to identify the putative pathogen based on its 16S rRNA sequence. We have recently tested  
30  
31 79 this method but not found it to be suitable because of its low resolution (Mechan-Llontop et al.  
32  
33 80 2019).  
34  
35

36  
37 81 Whole genome sequencing (WGS) does not require PCR and strain-level identification is  
38  
39 82 now routine practice in the surveillance of food-borne pathogen outbreaks in several countries  
40  
41 83 (Nadon et al. 2017). With the drop of sequencing cost and development of genome databases  
42  
43 84 that contain strain-level classification of plant pathogens, WGS now represents a real possibility  
44  
45 85 in plant disease diagnostics. For example, LINbase at [linbase.org](http://linbase.org) (Tian et al. 2019) contains  
46  
47 86 precise genome-based circumscriptions for many bacterial plant pathogens from the genus level  
48  
49 87 to the strain level. Genome sequences of unknown isolates can be identified as members of  
50  
51 88 circumscribed plant pathogens based on how similar they are at the whole genome level,  
52  
53 89 measured as Average Nucleotide Identity (ANI) (Konstantinidis and Tiedje 2005), to the other  
54  
55 90 members of these taxa. However, the limitation of WGS is its dependence on pure cultures.  
56  
57  
58  
59  
60

1  
2  
3 91 Metagenomic sequencing consists in extracting DNA directly from plant tissue followed by  
4  
5 92 sequencing all DNA present in the sample. Compared to WGS, the two main advantages of this  
6  
7 93 approach are that (1) it is much faster because it does not require lengthy pathogen isolation and  
8  
9 94 culturing procedures; and (2) it does not require much prior knowledge about the pathogen since  
10  
11 95 any pathogen, besides RNA viruses, can be detected with this method. However, the main  
12  
13 96 challenge of this approach is that the obtained DNA sequences also contain host plant sequences  
14  
15 97 and microbe sequences that do not belong to the pathogen. Therefore, obtaining sufficient  
16  
17 98 sequences of the causative agent and identifying the causative agent among all the other potential  
18  
19 99 causative agents present in the same plant requires optimized experimental methods for DNA  
20  
21 100 extraction and sequencing and optimized algorithms and genome databases for precise pathogen  
22  
23 101 identification.

24  
25  
26 102 The sequencing method that is currently most attractive for metagenomics-based  
27  
28 103 pathogen identification is nanopore sequencing with the Oxford Nanopore Technologies (ONT)  
29  
30 104 MinION™ device (Jain et al. 2016). The main strengths of this method are that (1) DNA can be  
31  
32 105 prepared for sequencing with relatively short protocols ranging from a few hours to less than an  
33  
34 106 hour (protocols.io), (2) the MinION™ sequencer is not much larger than a USB stick and can be  
35  
36 107 used with a desktop or a laptop computer in the lab or even in the field, (3) it provides the first  
37  
38 108 sequencing results within minutes from the start of a sequencing run, and (4) the output can reach  
39  
40 109 over 10 gigabases of DNA sequences (more than 1000 times the size of an individual bacterial  
41  
42 110 genomes) after 48 hours (MinION brochure 2019a). However, the major weaknesses are (1) the  
43  
44 111 currently high sequencing error rate of approximately 10% (Tedersoo et al. 2019; Loit et al. 2019)  
45  
46 112 and (2) that the sequencing hardware only works once at full capacity limiting reuse (MinION  
47  
48 113 brochure 2019b).

49  
50  
51 114 Metagenomic sequencing with the MinION™ has already been used on several crops for  
52  
53 115 identification of various pathogens (Chalupowicz et al. 2019) using ONT's software WIMP (Juul  
54  
55 116 et al. 2015) and on wheat to identify various fungal pathogens (Hu et al. 2019) using the sequence

1  
2  
3 117 alignment tool BLASTN (Camacho et al. 2009) in combination with custom databases. The  
4  
5 118 MinION™ has also been used for plant pathogen detection and identification starting from  
6  
7 119 extracted RNA or DNA in combination with general or specific primers to increase the quantity of  
8  
9 120 input for the MinION™ (Loit et al. 2019; Badial et al. 2018). However, in none of these studies,  
10  
11 121 was strain-level identification attempted directly from sequencing metagenomic DNA without prior  
12  
13 122 amplification.

14  
15 123 Here we tested the MinION™ with tomato plants artificially inoculated with different strains  
16  
17 124 of *Pseudomonas syringae*, including isolates of the *Pto* strains T1 and DC3000 (Cai et al. 2011),  
18  
19 125 and with plants from tomato fields showing symptoms of natural infection with bacterial spot for  
20  
21 126 which we did not know the *Xanthomonas* species that caused the infection. We then explored the  
22  
23 127 precision of identification that can be achieved when using ONT's WIMP software and the third  
24  
25 128 party tools Sourmash (Brown and Irber 2016) and MetaMaps (Dilthey et al. 2019) in combination  
26  
27 129 with default and custom reference databases. We also assembled metagenomic sequences into  
28  
29 130 contigs that were used as input to BLASTN (Camacho et al. 2009) and the LINbase Web service  
30  
31 131 for genome-based microbial identification (Tian et al. 2019).

32  
33  
34  
35 132

## 36 37 133 **Materials and Methods**

### 38 39 134 **Laboratory-infected tomato plants**

40  
41 135 Seeds of tomato (*Solanum lycopersicum*) 'Rio Grande' were germinated in potting mix soil  
42  
43 136 (Miracle-grow, OH, USA) under laboratory conditions with a long day period (16-h photoperiod)  
44  
45 137 and infected at 4 weeks of age. *Pto* isolate K40 (belonging to strain T1), *Pto* isolate DC3000  
46  
47 138 (belonging to strain DC3000) (Cai et al. 2011), *P. syringae* pv. *syringae* B728a (Feil et al. 2005),  
48  
49 139 and *P. syringae* 642 (Clarke et al. 2010) were grown in King's B solid medium at 28°C for 24  
50  
51 140 hours. Isolate *Pto* K40 was suspended at a concentration corresponding to an OD600 of 0.001 in  
52  
53 141 10 mM MgSO<sub>4</sub> for single-strain inoculation. For the mixed-strain inoculation, all four isolates were  
54  
55 142 suspended at an OD600 of 0.001 in 10 mM MgSO<sub>4</sub> and pooled together in equal amounts before

1  
2  
3 143 inoculation. Silwet L-77 was added to bacterial suspensions (0.025% vol/vol) to facilitate bacterial  
4  
5 144 infection. Plants were placed in ziplock plastic bags for high humidity conditions for 24 hours  
6  
7 145 before inoculation. After plants were spray-inoculated with 10 ml of bacterial suspensions, they  
8  
9 146 were placed back into the plastic bags for another 24 hours. Plants were processed for DNA  
10  
11 147 extraction four days after inoculation. Inoculation with 10mM MgSO<sub>4</sub> was included as a mock  
12  
13 148 treatment.  
14

15  
16 149

### 17 18 150 [Naturally infected tomato plants](#)

19  
20 151 Five tomato plants with bacterial spot symptoms, one plant with symptoms of Septoria leaf spot,  
21  
22 152 and one plant without symptoms were collected on August 10, 2018, on the Eastern Shore of  
23  
24 153 Virginia (Accomack and Northampton counties). The diagnosis of bacterial spot was made by  
25  
26 154 matching symptomology in the field (chlorotic haloes surrounding leaf lesions) with the presence  
27  
28 155 of bacterial streaming microscopically at 200X. Additionally, bacteria were cultured on King's  
29  
30 156 medium B (KB) media and were found to be non-fluorescent and of deep yellow color leading to  
31  
32 157 their identification as *Xanthomonas*. The diagnosis of Septoria leaf spot was confirmed by  
33  
34 158 microscopic identification of conidia at 200X. Plants were then shipped overnight to the Virginia  
35  
36 159 Tech campus in Blacksburg, VA, where they were processed for DNA extraction. Another set of  
37  
38 160 plants with bacterial spot symptoms were collected in May, 2019. Bacteria were isolated from  
39  
40 161 symptomatic leaves on KB. Plants and plates were shipped to the Virginia Tech campus overnight  
41  
42 162 where plants and bacterial colonies were processed for DNA extraction.  
43  
44

45 163

### 46 47 164 [DNA extraction](#)

48  
49 165 All plant samples used for DNA extraction are listed in Table 1. DNA extraction was performed  
50  
51 166 according to (Ottesen et al. 2013) with the following modifications. Briefly, wearing gloves, the top  
52  
53 167 of each plant sample (6 to 10 leaves from the top with or without stems) was collected using  
54  
55 168 clippers. The weight of samples was between 5 to 10 grams. After removing all the dirt from the  
56  
57  
58  
59  
60



1  
2  
3 169 plant surface by shaking vigorously, each sample was placed in a 6-1/2"× 5-7/8" Ziploc® bag  
4  
5 170 together with 300 ml sterilized double-distilled water (DDW). Samples were sonicated for 15  
6  
7 171 minutes using a Branson 1510 Ultrasonic Cleaner. DNA was extracted with DNeasy®  
8  
9 172 PowerWater® Kit (QIAGEN; Catalog # 14900-50-NF). All steps for DNA extraction were  
10  
11 173 performed according to the kit's specifications, except that after adding 1 mL of the kit's solution  
12  
13 174 PW1, the tube was incubated at 65°C for 15 minutes and then vortexed for 20 minutes.

15  
16 175 DNA from two plant-isolated bacteria was extracted from cultures derived from single  
17  
18 176 colonies with the Gentra® Puregene® Cell and Tissue Kit (Gentra Systems; Catalog # D5000).  
19  
20 177 All steps for DNA extraction were performed according to the Gram-negative Bacteria protocol,  
21  
22 178 except that cells were collected in 1 mL of sterilized DDW in a 1.5 ml microcentrifuge tube for the  
23  
24 179 lysis step. For both extraction procedures, the concentration and purity of DNA was measured  
25  
26 180 using a Thermo Scientific™ NanoDrop™ One<sup>C</sup> Spectrophotometer.

27  
28  
29 181

## 30 182 DNA library preparation

31  
32  
33 183 Library preparation was performed according to the '1D Native barcoding genomic DNA protocols  
34  
35 184 (EXP-NBD104, EXP-NBD114, and SQK-LSK108 or SQK-LSK109) provided by ONT. Sequencing  
36  
37 185 libraries were prepared using the Ligation Sequencing Kit (ONT Ltd.; SQK-LSK109). For each  
38  
39 186 run, NEBNext® Ultra™ II End Repair/dA-Tailing Module (New England Biolabs, Inc.; Catalog #  
40  
41 187 E7546S) was used for DNA repair and end-prep for each sample. Repaired DNA was cleaned up  
42  
43 188 by 1.5 volumes of AMPure XP beads, washed on a magnetic rack using freshly made 70%  
44  
45 189 ethanol, and eluted with 25 µL nuclease-free water. 22.5 µL elute was used for barcoding by  
46  
47 190 mixing with the Blunt/TA Ligase Master Mix (New England Biolabs, Inc.; Catalog # M0367S) and  
48  
49 191 Native Barcode (Oxford Nanopore Technologies Ltd.; Native Barcoding Expansion Kit EXP-  
50  
51 192 NBD104), followed by another wash step using 1.5 volumes of AMPure XP beads, and DNA was  
52  
53 193 eluted in 26 µL nuclease-free water. Equimolar amounts of barcoded DNA were then pooled into  
54  
55 194 a 1.5 mL microcentrifuge for ligation. Adapter ligation was performed by mixing the pooled

1  
2  
3 195 barcoded sample with Adapter Mix (Oxford Nanopore Technologies Ltd.; SQK-LSK109),  
4  
5 196 NEBNext® Quick Ligation Reaction Buffer (New England Biolabs, Inc.; Catalog # B6058S) and  
6  
7 197 Quick T4 DNA Ligase (New England Biolabs, Inc.; Catalog # M2200S). Ligated DNA was cleaned  
8  
9 198 up by one volume of AMPure XP beads, washed on a magnetic rack using Long Fragment Buffer  
10  
11 199 (Oxford Nanopore Technologies Ltd.; SQK-LSK109), and eluted with 15 µL Elution Buffer (Oxford  
12  
13 200 Nanopore Technologies Ltd.; SQK-LSK109).

15  
16 201 Sequencing was performed on ONT MinION™ flow cells (FLO-MIN106 R9 Version)  
17  
18 202 connected to a Mk1B device (ONT Ltd.; MIN-101B) operated by the MinKNOW software (version  
19  
20 203 3.3.2). Each flow cell was primed with the priming buffer prepared by mixing 30 µL Flush Tether  
21  
22 204 (ONT Ltd.; EXP-FLP001) with a tube of Flush Buffer (ONT Ltd.; EXP-FLP001). 12 µL of the final  
23  
24 205 library mixed with Sequencing Buffer (ONT Ltd.; SQK-LSK109) and Library Loading Beads (ONT  
25  
26 206 Ltd.; SQK-LSK109) was loaded onto the SpotON sample port of the flow cell in a dropwise  
27  
28 207 fashion. The sequencing run was stopped after 48 hours.

30  
31 208

### 32 209 [Illumina genome sequencing and assembly](#)

34  
35 210 Genomic DNA from isolated bacteria was used to prepare 350bp insert DNA libraries and  
36  
37 211 sequence on an Illumina platform PE150 at Novogene Corporation Inc (Sacramento, CA). FastQC  
38  
39 212 was used to assess the quality of the raw sequencing data (Andrews 2010). Adapter-trimming  
40  
41 213 was performed using BBduk with the parameters 'k=23, mink=9, hdist=1, ktrim=r, minlength=100'  
42  
43 214 (Bushnell 2015). Unicycler v0.4.7 with default parameters was used to *de novo* assemble the  
44  
45 215 bacterial genomes (Wick et al. 2017).

47 216

### 49 217 [Read-based metagenomic analysis](#)

#### 51 218 *Guppy*

53  
54 219 For all samples, the Fast5 files containing raw reads were base-called with the base-calling ONT  
55  
56 220 software Guppy-cpu (v3.3.2) available via the Nanopore website, which uses neural networks to

1  
2  
3 221 translate raw signals into DNA sequences in fastq format, with default parameters. All fastq files  
4  
5 222 were deposited in the NCBI SRA database under PRJNA588037.  
6

7 223 *What's in my pot? (WIMP)*  
8

9 224 The ONT workflow WIMP (v2019.7.9), which uses Centrifuge (Kim et al. 2016) to assign taxonomy  
10  
11 225 to reads in real-time, was used for species level identification in all samples. The workflow uses  
12  
13 226 bacterial, viral, and fungal genomes present in Refseq as the reference database.  
14  
15

16 227

17  
18 228 *Sourmash*  
19

20 229 Sourmash is a command-line tool used for k-mer based taxonomic classification of genomes and  
21  
22 230 metagenomes. It uses a MinHash sketching algorithm (Ondov et al. 2016) to create signatures,  
23  
24 231 which are compressed representations of DNA sequences that are then used to assign taxonomic  
25  
26 232 annotations. The *gather* function in this software was used for taxonomic classification at the  
27  
28 233 species- and strain-level. For species-level classification, the default Genbank LCA (Lowest  
29  
30 234 Common Ancestor) database (v.2018.03.29, k=31) containing 100,000 microbial genomes was  
31  
32 235 used. For strain level-classification, a custom library with 245 microbial genomes of representative  
33  
34 236 tomato plant pathogens and close relatives was used. A complete list of genomes used in the  
35  
36 237 custom reference library is provided in Supplementary Table 1. For all samples, signatures were  
37  
38 238 computed at 31 k-mer size (for species level) and 51 k-mer size (for strain level) and abundance  
39  
40 239 filtering was performed to exclude k-mers with an abundance of 1 (Brown and Irber 2016).  
41  
42 240 Sourmash was run on Virginia Tech's High Performance Computing system, Advanced Research  
43  
44 241 Computing (ARC), with Intel Broadwell, 2.1GHz CPU, 32 cores and 128GB memory.  
45  
46

47 242

48  
49 243 *MetaMaps*  
50

51 244 MetaMaps (Dilthey et al. 2019) was used for taxonomic classification at the species-level using  
52  
53 245 the miniSeq+H database, which includes more than 12,000 microbial genomes and is included  
54  
55 246 with the software package. For strain-level classification, the custom library described above for  
56  
57

1  
2  
3 247 Sourmash was used. However, the list of genomes was reduced to 149 to include only those  
4  
5 248 genomes that had NCBI taxonomy IDs as per a prerequisite for MetaMaps. MetaMaps was also  
6  
7 249 run on Virginia Tech's High Performance Computing system, Advanced Research Computing  
8  
9 250 (ARC), with Intel Broadwell, 2.1GHz CPU, 32 cores and 128GB memory.

### 251 *Metagenome-assembled genome analysis*

13  
14 252 The reads of each metagenome were mapped against each other to find overlaps using minimap2  
15  
16 253 (Li 2018) with the -x and ava-ont parameters. *De novo* assembly was performed for each  
17  
18 254 metagenome using the long reads assembler miniasm with default parameters (Li 2016).  
19  
20 255 Assembly correction was achieved by two iterations of racon (v1.4.7) with default parameters  
21  
22 256 (Vaser et al. 2017).

### 257 *BLAST*

26  
27 258 The assemblies of each metagenome were used as input to the command-line version of BLASTN  
28  
29 259 (Camacho et al. 2009) against the bacterial tomato pathogens custom database described above  
30  
31 260 and with the parameter of e-value set to less than or equal to 0.01. The top hit was determined to  
32  
33 261 be the alignment with the longest length for each contig.

### 262 *LINbase*

34  
35  
36  
37 263 The two longest contigs of each metagenome assembly were used as input to LINbase at  
38  
39 264 linbase.org (Tian et al. 2019) to identify the pathogens at the strain level with the function "Identify  
40  
41 265 using a genome sequence".  
42

43 266

## 45 267 **Results**

### 47 268 **Read-based pathogen identification after single-strain inoculation in the laboratory**

49 269 Tomato plants inoculated with *Pto* isolate K40 (strain T1) in the laboratory showed bacterial speck  
50  
51 270 symptoms four days after inoculation (Figure 1A), at which time DNA was extracted from a leaf  
52  
53  
54 271 wash fluid after sonication.

1  
2  
3 272 The quantity and quality of the extracted DNA are listed in Table 2. An entire MinION™  
4  
5 273 flow cell was used to sequence this sample (called L-K40). Of all the sequencing reads, 1,377,617  
6  
7 274 reads (approximately 60% of the total number of reads) were base-called after the run was  
8  
9 275 completed using the guppy software. The base-called reads had a total length of approximately  
10  
11 276 4.2 Gigabases (Gbp) with the longest read measuring 66,000 bp (see more details about reads  
12  
13  
14 277 in Table 1).

15  
16 278 The base-called reads were used as input to WIMP, which classified 89% of reads as of  
17  
18 279 bacterial origin. This result showed that our DNA extraction method starting from sonicated leaf  
19  
20 280 washes was successful at minimizing host DNA contamination. Of these reads, WIMP identified  
21  
22 281 77.47% as *P. syringae* genomospecies 3, a genome similarity group of which *Pto* is a member.  
23  
24 282 This genome similarity group was never validly published as a named species and is thus referred  
25  
26 283 to with the number 3 instead of a name (Gardan et al. 1999). Also NCBI's taxonomy database  
27  
28 284 (Sayers et al. 2009) includes this taxon as *P. syringae* genomospecies 3. The next most abundant  
29  
30 285 species were identified as *P. syringae* (9.39%), *P. cerasi* (2.09%), and *P. savastanoi* (1.60%).  
31  
32 286 Figure 2 shows a screenshot of the WIMP result. The composition analysis is shown in Figure 3A  
33  
34 287 (see Supplementary Table 2 for all relative abundance values for all composition analyses shown  
35  
36 288 in Figure 3 and 4).

37  
38  
39 289 Next, the reads were used as input for composition analysis using Sourmash (Brown and  
40  
41 290 Irber 2016) and MetaMaps (Dilthey et al. 2019) using the default reference libraries provided by  
42  
43 291 these programs. Results are shown in Figure 3A. Sourmash identified 56.84% of the reads as *P.*  
44  
45 292 *syringae* genomospecies 3 while MetaMaps identified over 91.53% of the reads as *P. syringae*  
46  
47 293 genomospecies 3. Similarly to WIMP, both programs identified *P. syringae* as the next most  
48  
49 294 abundant species (14.41% and 4.17%, respectively). All other species were found at a relative  
50  
51 295 abundance of 2% or below. Therefore, WIMP, MetaMaps, and Sourmash all correctly identified  
52  
53 296 the pathogen used in the inoculation as a member of *P. syringae* genomospecies 3.  
54  
55 297 Supplementary Table 3 reports the run times for the three tools for this sample.

1  
2  
3 298 In an attempt to reach strain level resolution (since WIMP is limited to species-level  
4  
5 299 identification), we built Sourmash and MetaMaps custom reference libraries consisting of genome  
6  
7 300 sequences of representative bacterial tomato pathogen isolates and closely related isolates that  
8  
9 301 do not cause disease on tomato. The libraries included multiple isolates of the *Pto* strains DC3000  
10  
11 302 and T1 (Supplementary Table 2). When using these custom libraries, Sourmash identified 71.64%  
12  
13 303 of the sequences in the sample as *Pto* isolate T1 (the isolate after which strain T1 is named) and  
14  
15 304 the remaining sequences as other *P. syringae* isolates that are not pathogens of tomato (Table  
16  
17 305 2). Only 0.9% of the sequences were misidentified as *Pto* DC3000. MetaMaps in combination  
18  
19 306 with the same custom library identified 70.93% as *Pto* isolate T1, 15.90% as *Pto* isolate  
20  
21 307 NCPPB1108 (another isolate belonging to strain T1), and 7.81% as *Pto* isolate DC3000.  
22  
23 308 Therefore, both Sourmash and MetaMaps identified most of the reads correctly as an isolate  
24  
25 309 belonging to *Pto* strain T1 but MetaMaps misidentified many more reads as *Pto* strain DC300  
26  
27 310 compared to Sourmash.  
28  
29  
30  
31

311

### 312 [Read-based pathogen identification after multi-strain inoculation in the laboratory](#)

313 Next, we wanted to test the bioinformatics pipelines established with the single-strain inoculation  
34  
35 314 by using a mixed inoculum consisting of the *Pto* isolate K40 (strain T1) and the *Pto* isolate DC3000  
36  
37 315 (strain DC3000) of *P. syringae* genomospecies 3 together with two additional isolates of the  
38  
39 316 species *P. syringae* that do not cause disease on tomato: the bean pathogenic isolate *Psy* B728a  
40  
41 317 and the non-pathogenic isolate *Psy* 642. DNA was again extracted on day four after inoculation  
42  
43 318 and sequenced on an entire flow cell. All details for this sample (called L-mix) are listed in Table  
44  
45 319 1. Approximately 1 million reads of a total length of 4.2 Gbp were obtained with the longest read  
46  
47 320 measuring 67,000 bp. Since this time 100% of reads were base-called, the number of base-called  
48  
49 321 reads and the total length of reads were very similar to the single strain inoculation sample.  
50  
51  
52

53 322 The caveat with this sample is that we did not know the relative abundance of the 4 isolates  
54  
55 323 in the sample. However, since *Pto* isolates T1 and DC3000 are tomato pathogens while *Psy*

1  
2  
3 324 isolates B728a and 642 are not, we expected that most sequences would be identified again as  
4  
5 325 *P. syringae* genomospecies 3. In fact, WIMP identified 79.61% of all bacterial sequences (which  
6  
7 326 constituted 95% of all reads) as *P. syringae* genomospecies 3 (Figure 3B), similar to the 77.47%  
8  
9 327 identified in the single-strain inoculation sample. Compared to WIMP, Sourmash and MetaMaps  
10  
11 328 showed the same trend as with the single strain inoculation sample: Sourmash found a lower  
12  
13 329 relative abundance of *P. syringae* genomospecies 3 (43.24%) compared to WIMP and MetaMaps  
14  
15 330 found a higher relative abundance compared to WIMP (91.09%) (Figure 3B).

16  
17  
18 331 Since both *Psy* isolates used in the inoculation belong to the species *P. syringae*, we  
19  
20 332 expected a slightly higher relative abundance of *P. syringae* compared to the single strain  
21  
22 333 inoculation sample. Interestingly, this expectation came true for Sourmash (36.87% versus  
23  
24 334 14.4%) but for WIMP and MetaMaps the relative abundance of *P. syringae* only increased  
25  
26 335 marginally from 9.38% to 10.01% and from 4.17% to 5.39%, respectively (Figure 3B).

27  
28 336 We then used the custom reference libraries of representative tomato pathogens to see if  
29  
30 337 Sourmash and MetaMaps could distinguish isolate K40 (of strain T1) from isolate DC3000 (of  
31  
32 338 strain DC3000). Sourmash did identify isolate T1 of strain T1 at a relative abundance of 65.98%  
33  
34 339 and isolate DC3000 of strain DC3000 at a relative abundance of 16.01% (Table 2) while  
35  
36 340 MetaMaps identified 84.71% of the reads as isolates that belong to strain T1 and 5.61% as isolate  
37  
38 341 DC3000 (not shown in Table 2 since only the top three hits are shown for each sample).

39  
40  
41 342 Since we did not know the correct relative abundances of strains in this inoculated plant  
42  
43 343 sample and could thus not determine how accurate the results were, we decided to sequence an  
44  
45 344 additional sample (called L-culture-mix) that consisted of DNA extracted from an equal mixture of  
46  
47 345 the same four strains after they were grown separately overnight in liquid culture. Approximately  
48  
49 346 54,000 reads of a total length of 150 Mbp were obtained on 1/6th of a flow cell with the longest  
50  
51 347 read measuring 76,000 bp. WIMP classified 95% of the reads as bacterial. WIMP, MetaMaps,  
52  
53 348 and Sourmash identified both, *P. syringae* and *P. syringae* genomospecies 3 in this sample, which  
54  
55 349 we expected to be present at 50% each. WIMP over-estimated *P. syringae* compared to *P.*



1  
2  
3 350 *syringae* genomospecies 3 (56% compared to 28%) and identified some other species at low  
4  
5 351 relative abundance (Figure 3C). MetaMaps also overestimated *P. syringae* compared to *P.*  
6  
7 352 *syringae* genomospecies 3: 65.58% vs 32.19%. Sourmash came the closest to the expected 1 to  
8  
9 353 1 ratio finding 52.20% of *P. syringae* and 41.68% of *P. syringae* genomospecies 3 (Figure 3C).  
10  
11 354 When using the custom reference libraries of tomato pathogens with MetaMaps and Sourmash,  
12  
13 355 MetaMaps outperformed Sourmash since it identified DC3000 and T1 close to the expected 25%  
14  
15 356 abundance: 38.89% and 27.48%, respectively (Table 2). Sourmash instead assigned a much  
16  
17 357 higher abundance to strain DC3000 (75.1%) compared to strain T1 (19.63%) (Table 2).  
18

19  
20 358 Finally, we sequenced the leaf wash from a tomato plant grown in the lab that was not  
21  
22 359 inoculated with any pathogen (called sample L-mock). Since the DNA concentration of this sample  
23  
24 360 was very low, only approximately 82,000 base-called reads were obtained on 1/7th of a flow cell  
25  
26 361 with a total length of 103 Mb. The longest read was only 19,000 bp long. Only 8% of the reads  
27  
28 362 were classified as bacterial showing that this lab-grown plant was not colonized by many bacteria,  
29  
30 363 which was probably also the reason for the low DNA concentration. WIMP, Sourmash, and  
31  
32 364 MetaMaps provided very different results for this sample (Figure 3D). Importantly, as expected  
33  
34 365 from a non-inoculated plant, none of the reads were identified by either of the three tools as *P.*  
35  
36 366 *syringae* or *P. syringae* genomospecies 3.  
37  
38

39 367

#### 40 41 368 [Read-based pathogen identification in naturally infected tomato field samples](#)

42  
43 369 After obtaining promising results in regard to strain-level identification with laboratory samples,  
44  
45 370 we used DNA extracted from tomato field samples that were collected on the Eastern Shore of  
46  
47 371 Virginia to test our pipelines with naturally infected plants (Table 1). The samples came from  
48  
49 372 tomato plants that either showed symptoms of bacterial spot (samples F1-bs, F2-bs, F4-bs, F7-  
50  
51 373 bs, F8-bs; see Figure 1B), symptoms of the fungal disease *Septoria* leaf spot (sample F5-  
52  
53 374 *Septoria*) or no signs of any disease (F6-healthy). We also obtained one sample (F3-bs) with  
54  
55 375 symptoms of bacterial spot. However, colonies obtained by culturing bacteria from this plant  
56  
57  
58  
59  
60



1  
2  
3 376 during the initial diagnosis (not used for sequencing) had been identified as a mixture of  
4  
5 377 *Pseudomonas* and *Xanthomonas*.

6  
7 378 DNA from all tomato field samples were barcoded and sequenced together with other  
8  
9 379 samples by multiplexing them on the same flow cell. Therefore, the number of reads (between  
10  
11 380 35,923 for samples F6-healthy and 137,497 for F1-bs) and total read length (between 66  
12  
13 381 megabases (Mb) for F6-healthy and 588 Mb for F1-bs) for these samples were much lower  
14  
15 382 compared to the laboratory samples (Table 1).

16  
17  
18 383 Detailed results for all samples are reported in Figure 4. Similar to the lab-inoculated  
19  
20 384 samples, the majority of reads in the field samples that had symptoms of bacterial disease were  
21  
22 385 classified as bacteria by WIMP (between 78 and 81%). Importantly, WIMP and Sourmash agreed  
23  
24 386 that *X. perforans* was the species with the highest relative abundance in these samples (between  
25  
26 387 25.82% and 56.44% for WIMP and between 18.51 and 66.01% for Sourmash) suggesting that *X.*  
27  
28 388 *perforans* was the causative agent. Sample F3-bs, which had a mixed  
29  
30 389 *Xanthomonas/Pseudomonas* infection based on culturing, was found by both WIMP and  
31  
32 390 Sourmash to still be dominated by *X. perforans* (21.98% and 19.55% respectively) followed by  
33  
34 391 either *P. oryzihabitans* (10.11%) and *P. fluorescens* (5.09%) based on WIMP or *P. putida*  
35  
36 392 (16.98%) based on Sourmash. Therefore, the presence of a mixed infection was confirmed by  
37  
38 393 both tools.

39  
40  
41 394 In contrast to the results from WIMP and Sourmash, MetaMaps identified *X. euvesicatoria*  
42  
43 395 and *X. alfalfae* instead of *X. perforans* as the two species with the highest relative abundance in  
44  
45 396 all samples with bacterial spot symptoms. This is because *X. perforans* was missing from the  
46  
47 397 MetaMaps reference library.

48  
49 398 Interestingly, even the non-symptomatic tomato sample (F6-healthy) was found to include  
50  
51 399 *X. perforans* as the species with the highest relative abundance based on WIMP and Sourmash.  
52  
53 400 However, the relative abundance values were lower (6.89% and 18.54%, respectively). This  
54  
55 401 suggests that this plant might have been infected with *X. perforans* but was asymptomatic

1  
2  
3 402 because of lower bacterial titer. This non-symptomatic sample also included a number of species  
4  
5 403 at relatively high abundance that were rarely found in the samples with bacterial spot symptoms,  
6  
7 404 for example, *P. oleovorans*, *Sphingomonas parapaucimobilis*, *Microbacterium sp.* Leaf203, and  
8  
9 405 *Methylobacterium populi*.

11 406 The sample with *Septoria* leaf spot symptoms (F5-Septoria), probably infected by the plant  
12  
13 407 pathogenic fungus *Septoria lycopersici*, carried a diverse bacterial population consisting of  
14  
15 408 species in the genera *Pseudomonas*, *Xanthomonas*, *Pantoea*, *Curtobacterium*,  
16  
17 409 *Methylobacterium*, and *Sphingomonas*. The genome of *Septoria lycopersici* is not publicly  
18  
19 410 available and other species of the genus *Septoria* were not included in any of the reference  
20  
21 411 libraries. Identification of this fungal pathogen was thus not pursued any further.

24 412 When we analyzed our samples with Sourmash and MetaMaps using our custom  
25  
26 413 database of representative bacterial tomato pathogens as reference libraries, *X. perforans*  
27  
28 414 isolates TB9, TB15, and Xp9-5 were identified as the top hits in all plants with bacterial spot  
29  
30 415 symptoms with the exception of F3-bs, which had the mixed *Pseudomonas/Xanthomonas*  
31  
32 416 infection. In this sample, isolate Xp17-12 was identified by both Sourmash and MetaMaps as top  
33  
34 417 hit. Interestingly, isolates TB9, TB15, and Xp9-5 are all members of the same intraspecific group,  
35  
36 418 *X. perforans* group 2, based on core genome phylogeny (Schwartz et al. 2015), suggesting that  
37  
38 419 the *X. perforans* strain infecting the tomatoes with bacterial spot symptoms on the Eastern Shore  
39  
40 420 of Virginia was also a member of *X. perforans* group 2.

43 421 For sample F8-bs, we also isolated *Xanthomonas* bacteria to compare the results from  
44  
45 422 the culture-independent, read-based metagenomic approach with a culture-dependent genomic  
46  
47 423 approach. DNA was extracted from two colonies and sequenced using Illumina HiSeq. The two  
48  
49 424 genome sequences were assembled into 87 and 86 contigs, respectively, with a total length of  
50  
51 425 5,340,265 bp and 5,339,287 bp. We used the LINbase Web service for genome-based microbial  
52  
53 426 identification and found isolate GEV1063 to be the best match for both genomes with 99.98% ANI  
54  
55 427 and both genomes were identified by LINbase as members of *X. perforans* group 2, which is

1  
2  
3 428 circumscribed in LINbase as an intraspecific taxon. Therefore, the culture-dependent genome-  
4  
5 429 based identification confirmed the culture-independent read-based strain-level identification of *X.*  
6  
7 430 *perforans* group 2 as the causative agent in sample F8-bs.  
8  
9  
10 431

### 11 432 [Metagenome assembly-based pathogen identification](#)

12  
13 433 In parallel to the read-based pipelines described above, we also assembled each metagenomic  
14  
15 434 sample using all reads that had a minimum length of 1,000 bp followed by two iterations of the  
16  
17 435 error correcting tool racon (Vaser et al. 2017). The results are summarized in Table 3. The non-  
18  
19 436 inoculated tomato sample from the lab (L-mock), the healthy tomato sample from the field (F6-  
20  
21 437 healthy), and the sample of the tomato plant with Septoria leaf spot (F5-Septoria) had the lowest  
22  
23 438 number of contigs (between 4 and 9) with the shortest total length of contigs (between 21,390 bp  
24  
25 439 and 122,956 bp). This was probably a result of the low number of bacterial reads in these samples  
26  
27 440 (Table 1).  
28  
29

30  
31 441 The samples with symptoms of either bacterial speck or bacterial spot had a wide range  
32  
33 442 in regard to contig number (10 to 131 contigs) and total length of contigs (5.2 to 12.8 Mbp). For  
34  
35 443 our goal of identifying the causative agent in each symptomatic plant to strain level, we focused  
36  
37 444 on the two longest contigs in each sample since these contigs were the most likely to be of the  
38  
39 445 causative pathogenic agents. It was very promising to see that in some of the symptomatic  
40  
41 446 samples the longest contig was of a size similar to an entire bacterial genome, for example,  
42  
43 447 6.08Mbp in the tomato lab sample inoculated with Pto isolate K40 (L-K40), and 5.03Mbp for the  
44  
45 448 field sample F7-bs showing bacterial spot symptoms (Table 3). We then used the genome  
46  
47 449 alignment tool MUMmer (Marçais et al. 2018) to determine how much of the published genome  
48  
49 450 sequences these contigs covered. We found that in the case of sample L-K40, the longest contig  
50  
51 451 aligned with 97.90% of the published genome sequence of isolate K40. For F7-bs, the longest  
52  
53 452 contig aligned with 95.81% of the published *X. perforans* genome TB15 (the genome identified  
54  
55 453 by Sourmash with the highest abundance in this sample).  
56  
57  
58  
59  
60

1  
2  
3 454 To obtain a preliminary identification of all contigs we used BLASTN (Camacho et al. 2009)  
4  
5 455 in combination with our custom tomato pathogen database. The results were mostly in agreement  
6  
7 456 with the reads-based analysis at the species level (Figure 5) but *X. euvesicatoria* was identified  
8  
9 457 as species instead of *X. perforans* in some of the samples with bacterial spot.

11 458 To attempt identification of the longest contigs to strain level, we used these contigs as  
12  
13 459 queries with the “Identify using a genome sequence” function in the LINbase Web service (Tian  
14  
15 460 et al. 2019). Table 4 lists the results that were obtained for the longest two contigs (separately  
16  
17 461 and merged) for each sample. When using the longest contig of the tomato plant inoculated with  
18  
19 462 isolate K40 of *Pto* strain T1 (sample L-K40), the *Pto* strain T1 isolate BAV1020 was identified as  
20  
21 463 best hit with an ANI of 99.76% compared to the query sequence. This very high ANI value shows  
22  
23 464 that the error-correcting tool racon (Vaser et al. 2017) was successful in correcting most  
24  
25 465 sequencing errors in the assembly. K40 was expected to be the best hit for this sample since this  
26  
27 466 is the isolate that was used in the inoculation. We do not know why isolate BAV1020 was identified  
28  
29 467 as best hit. However, isolates BAV1020 and K40 have a reciprocal ANI of over 99.75%, both were  
30  
31 468 isolated from tomato plants in Virginia, and both belong to *Pto* strain T1, making it irrelevant for  
32  
33 469 strain-level identification which isolate was the best hit. Most importantly, since genome-  
34  
35 470 sequenced isolates of *Pto* strain T1 have pair-wise ANI values of 99.75% or higher and the ANI  
36  
37 471 between the longest contig of L-K40 and its best hit BAV1020 had an ANI of over 99.76%, we  
38  
39 472 were able to identify L-K40 as member of *Pto* strain T1.

43 473 For the tomato plant inoculated with the four-strain mix (sample L-mix), the longest contig  
44  
45 474 was again identified as *Pto* strain T1 based on the same best hit to *Pto* isolate BAV1020 with an  
46  
47 475 ANI value of 99.77%. Interestingly, using the two longest contigs together in a single query, isolate  
48  
49 476 K40 was identified as the best hit. No contig of significant length was identified as either *Pto* isolate  
50  
51 477 DC3000 or the other two *Psy* isolates used in the inoculation. This may have been due to poor  
52  
53 478 growth of these isolates in tomato compared to isolate K40 (as suggested by the read-based  
54  
55 479 analysis above). Moreover, since the genomes of *Pto* isolates DC3000 and K40 are over 98.5%

1  
2  
3 480 identical to each other, some DC3000 reads may have been assembled together with K40 reads  
4  
5 481 into the same contigs.  
6

7 482 For the longest contigs in the tomato field samples that showed bacterial spot symptoms,  
8  
9 483 different isolates of *X. perforans* were the best hits: GEV1063, GEV2116, and TB6 (Table 4).  
10  
11 484 Isolates GEV1063 and TB6 both belong to *X. perforans* group 2 (Schwartz et al. 2015) and this  
12  
13 485 results is thus in line with the read-based results described above. Only the second-longest contig  
14  
15 486 in sample F4-bs contradicted the read-based results: *X. perforans* isolate 91-118, a member of  
16  
17 487 *X. perforans* group 1B (Schwartz et al. 2015), was the best hit for this contig.  
18  
19

20 488 Since for sample F8-bs we also had the genome sequences of the two cultured isolates  
21  
22 489 (sequenced with Illumina and assembled with Unicycler; see previous section), we could again  
23  
24 490 directly compare the metagenomic assembly-based approach with the culture-dependent  
25  
26 491 genomic approach. The best match in LINbase for both approaches was the isolate GEV1063 of  
27  
28 492 *X. perforans* group 2. The ANI value of 99.76% between the longest contig of F8-bs and isolate  
29  
30 493 GEV1063 was almost as high as the ANI value between the Illumina-sequenced isolates cultured  
31  
32 494 from F8-bs and isolate GEV1063, which was 99.98%. However, since genome-sequenced  
33  
34 495 isolates of *X. perforans* group 2 have pair-wise ANI values of over 99.9% and the ANI between  
35  
36 496 the longest contig of F8-bs and its best hit, isolate GEV1063, was 99.76%, we could not identify  
37  
38 497 the causative agent in sample F8-bs with high confidence as member of *X. perforans* group 2.  
39  
40

41 498

## 42 43 499 **Discussion**

44  
45 500 Sensitive detection and precise identification of pathogens in real time directly from symptomatic  
46  
47 501 organisms, or even better from infected but still asymptomatic organisms, without the need for  
48  
49 502 pathogen isolation and culturing, is the ultimate goal in control and prevention of infectious  
50  
51 503 diseases of humans, animals, and plants.  
52

53  
54 504 As a step towards this goal in plant pathology, here we used the ONT MinION™ for precise  
55  
56 505 identification of two bacterial tomato pathogens by sequencing metagenomic DNA directly  
57

1  
2  
3 506 extracted from symptomatic plants and analyzing the obtained sequences with a set of different  
4  
5 507 tools and databases. However, we neither attempted to maximize sensitivity of detection nor to  
6  
7 508 minimize the time necessary for identification.  
8

9  
10 509 Several other reports describing the use of the MinION™ in culture-independent  
11  
12 510 metagenomic DNA sequencing for plant pathogen identification have recently been published.  
13  
14 511 Most of these reports either focused on species-level identification (Hu et al. 2019) and/or on  
15  
16 512 accelerating the identification protocol (Loit et al. 2019). Only one report focused on strain-level  
17  
18 513 identification but after polymerase chain reaction with primers specific to loci of a single pathogen  
19  
20 514 species, which increased the sensitivity of detection and resolution of identification but restricts  
21  
22 515 the approach to a single pathogen species at the time (Radhakrishnan et al. 2019). Our goal  
23  
24 516 instead was to develop an experimental and bioinformatics pipeline that can be used for any  
25  
26 517 bacterial plant pathogen, and, with modifications, possibly for fungal and oomycete pathogens as  
27  
28 518 well.  
29

30  
31 519 The first critical step in metagenomic-based pathogen identification is DNA extraction.  
32  
33 520 There are mainly two possibilities: extracting DNA directly from plant tissue or extracting DNA  
34  
35 521 from water used to wash the plant (after sonication to help dislocate the pathogen from the tissue).  
36  
37 522 The first approach has the advantage that large quantities of high-quality DNA can be extracted.  
38  
39 523 The obvious disadvantage is that a large fraction of the extracted DNA is plant DNA. The second  
40  
41 524 approach is the approach we decided to use since it is widely used for plant microbiome analysis,  
42  
43 525 for example (Ottesen et al. 2013). Based on the results from our DNA sequence analysis, this  
44  
45 526 approach allowed us to obtain DNA that was over 80% of bacterial origin for the naturally infected  
46  
47 527 tomato field samples and over 90% of bacterial origin for the artificially inoculated tomato plants  
48  
49 528 grown in the laboratory. This value was as high as the fraction of bacterial DNA when extracting  
50  
51 529 DNA directly from a bacterial culture. Therefore, we conclude that for metagenome-based  
52  
53 530 identification of bacterial foliar pathogens in symptomatic plant tissue extracting DNA from wash  
54  
55 531 water after sonication is an excellent solution. Importantly, even the wash water of our healthy  
56  
57  
58  
59  
60



1  
2  
3 532 field sample still contained 30% of bacterial DNA, making this approach possibly still a good  
4  
5 533 choice even for asymptomatic leaves with relatively low bacterial titers.  
6

7 534 Because in this project we were not interested in speed, we used the slower, higher  
8  
9 535 yielding DNA sequencing library preparation protocol, as suggested by ONT, without significant  
10  
11 536 modifications. Also for the sequencing protocol itself, we followed ONT's instructions without  
12  
13 537 modifications. The first critical step after sequencing the DNA, is base-calling, which is the process  
14  
15 538 of translating the raw electrical signals measured by the MinION™ into nucleotide sequences.  
16  
17 539 Since base-calling is computationally intensive and takes longer than sequencing itself, base-  
18  
19 540 calling needed to be completed after the sequencing runs themselves were completed. We used  
20  
21 541 the ONT Guppy base-calling tool without any polishing.  
22  
23

24 542 The actual assignment of sequencing reads to specific bacterial species and strains was  
25  
26 543 done using a total of five tools: 1. ONT's WIMP software with graphical user interface, which is  
27  
28 544 intuitive to use and uses the software Centrifuge (Kim et al. 2016) to rapidly identify and assign  
29  
30 545 taxonomy to the reads coming from the sequencing base calling in real-time, 2. the command-  
31  
32 546 line tool Sourmash (Brown and Irber 2016) that computes hash sketches from DNA sequences  
33  
34 547 and includes k-mer based taxonomic classification for genomic and metagenomic analysis, 3. the  
35  
36 548 command line tool MetaMaps (Dilthey et al. 2019) which uses approximate mapping algorithm to  
37  
38 549 map long-read metagenomic sequences to comprehensive databases, 4. the command line  
39  
40 550 version of BLASTN (Camacho et al. 2009) was used to speed up the identification of pathogens  
41  
42 551 with a custom built database after metagenome assembly, 5. after metagenome assembly  
43  
44 552 performed with minimap 2 and miniasm (Li 2016), the two longest contigs of each metagenome  
45  
46 553 assembly were used for taxonomy assignment with LINbase (Tian et al. 2019). Moreover,  
47  
48 554 Sourmash and MetaMaps were used both with default and custom libraries.  
49  
50

51 555 For species-level identification, the three read-based tools performed similarly well with  
52  
53 556 the lab samples in regard to accuracy with Sourmash coming the closest to the expected 1 : 1  
54  
55 557 ratio of *P. syringae* genomospecies 3 : *P. syringae* in the sample L-culture-mix. For the field  
56  
57  
58  
59  
60

1  
2  
3 558 samples, the absence of *X. perforans* in the MetaMaps default reference library did not allow  
4  
5 559 MetaMaps to identify *X. perforans* while WIMP and Sourmash performed similarly well. Both  
6  
7 560 identified *X. perforans* as the most abundant species in all samples with bacterial spot symptoms.  
8

9 561 As for run time, only WIMP is set up to provide real-time results starting minutes after runs  
10  
11 562 are initiated and results are updated as more sequencing reads are base-called. However, since  
12  
13 563 base-calling cannot keep up with the amount of raw data that is being generated during a run,  
14  
15 564 WIMP needs to be re-run when base-calling is completed after a run ends in order to analyze all  
16  
17 565 data. This took over 36 hours for our largest sample, L-K40 (Supplementary Table 3). The  
18  
19 566 advantage is that users do not need any significant local computing resources to do this since  
20  
21 567 WIMP runs on ONT's cloud. For the same L-K40 sample, it took Sourmash only 35 minutes to  
22  
23 568 calculate the k-mer signature and perform species-level classification while MetaMaps completed  
24  
25 569 the same run in 6-8 hours. Both tools were run on Virginia Tech's ARC high-performance  
26  
27 570 computing system. Therefore, Sourmash is significantly faster than MetaMaps and WIMP but still  
28  
29 571 requires significant computing resources.  
30  
31

32 572 In regard to ease of use, WIMP stands out because of its intuitive graphical user interface.  
33  
34 573 Although both Sourmash and MetaMaps are command-line tools, Sourmash beats MetaMaps  
35  
36 574 because of the extensive tutorials provided on the Sourmash website. The added ease of making  
37  
38 575 custom reference libraries and adding genomes to existing libraries also makes Sourmash more  
39  
40 576 user-friendly compared to MetaMaps, which requires NCBI taxIDs (or creation of custom taxIDs)  
41  
42 577 for all genomes in custom reference libraries.  
43  
44

45 578 Assembling reads into contigs before identification did not provide any advantages for  
46  
47 579 species-level identification since species-level identification was successful with read-based tools  
48  
49 580 and read-based identification is generally faster since it does not require prior assembly of reads  
50  
51 581 into contigs. However, this advantage of speed may diminish with an increasing number of reads  
52  
53 582 since mapping of a smaller number of assembled contigs might be faster than mapping a large  
54  
55 583 number of reads individually.  
56  
57



1  
2  
3 584 For strain-level identification, WIMP cannot be used since it only reaches species-level  
4  
5 585 resolution. When comparing MetaMaps with Sourmash, MetaMaps misidentified a larger number  
6  
7 586 of reads as strain *Pto* DC3000 compared to Sourmash in the single strain inoculation sample L-  
8  
9 587 K40, which we knew did not contain any DNA of strain *Pto* DC3000. Instead in the sample L-  
10  
11 588 culture-mix with known equal concentrations, it was Sourmash that overestimated strain *Pto*  
12  
13 589 DC3000 compared to strain *Pto* T1. For field sample F8-bs for which we had also a culture-  
14  
15 590 dependent result indicating *X. perforans* group 2 as causative agent, both software identified the  
16  
17 591 same best hit in the custom database that was also a member of *X. perforans* group 2. Therefore,  
18  
19 592 we conclude that Sourmash and MetaMaps did equally well in regard to strain accuracy. In regard  
20  
21 593 to run time, Sourmash's run time increased to 1-3 hours when using a k-mer size of 51, which is  
22  
23 594 required for strain-level identification. Run time for MetaMaps decreased to 3-4 hours because of  
24  
25 595 the smaller size of the custom library in comparison to default databases. However, Sourmash  
26  
27 596 still performed better than MetaMaps in regard to computing time.

30  
31 597 The challenge when using either Sourmash or MetaMaps for strain-level identification is  
32  
33 598 that we had to interpret the results based on prior knowledge of which isolates in our custom  
34  
35 599 database belonged to which pathogen strain. For example, only by checking Figure 1 in (Schwartz  
36  
37 600 et al. 2015), were we able to identify the best matches found by Sourmash and MetaMaps in our  
38  
39 601 custom database as members of *X. perforans* group 2. Moreover, a best match with an isolate  
40  
41 602 that belongs to a certain strain, or any other group or taxon for that matter, still does not  
42  
43 603 necessarily mean that the query is a member of the same group as well. To make such a  
44  
45 604 conclusion, it is necessary to determine (1) the genomic breadth of the group, for example,  
46  
47 605 99.75% for *X. perforans* group 2, and (2) the genomic distance of the query to a representative  
48  
49 606 member of that group with this distance needing to be smaller than the genomic breadth of the  
50  
51 607 group. Alternatively, a phylogenetic analysis could be performed to determine if the unknown is a  
52  
53 608 member of the clade that corresponds to the specific group. Because species have a standard  
54  
55 609 genomic breadth of 95% ANI, WIMP, Sourmash, and MetaMaps can infer species membership

1  
2  
3 610 from metagenomic reads relatively easily. However, strains (and any other group smaller than a  
4  
5 611 species) do not have a standard ANI breadth. Therefore, Sourmash and MetaMaps would need  
6  
7 612 to be given genomic circumscriptions of strains as part of the reference library information in order  
8  
9 613 to precisely assign reads to strains.

11 614 Since the MinION™ outputs long reads, we were successful in assembling reads into  
12  
13 615 contigs almost as long as entire bacterial genomes, which could then be used for genome-based  
14  
15 616 identification. We specifically developed the LINbase Web service for identifying microbes as  
16  
17 617 members of taxa at any genomic breadth below the rank of genus (Tian et al. 2019) and we had  
18  
19 618 circumscribed both *Pto* strain T1 and *X. perforans* group 2 as taxa in LINbase with genomic  
20  
21 619 breadths of 99.75% and 99.9% ANI, respectively. These ANI thresholds were chosen because  
22  
23 620 genome-sequenced isolates of *Pto* strain T1 deposited in LINbase have pair-wise ANI values of  
24  
25 621 over 99.75% and genome-sequenced isolates of *X. perforans* group 2 deposited in LINbase have  
26  
27 622 pairwise ANI values of over 99.9%. Since ANI between the longest contig of sample L-K40 and  
28  
29 623 the isolate T1 of *Pto* strain T1 was 99.76% (and thus above the 99.75% ANI threshold at which  
30  
31 624 *Pto* strain T1 is circumscribed in LINbase), we were able to correctly identify the causative agent  
32  
33 625 in sample L-K40 as a member of *Pto* strain T1. For the field samples, this was not possible since  
34  
35 626 ANI between the longest contigs and the most similar isolates of *X. perforans* group 2 in LINbase  
36  
37 627 was between 99.62% and 99.84% (and thus below the 99.9% ANI threshold at which *X. perforans*  
38  
39 628 group 2 is circumscribed in LINbase). We expect that a modest reduction in the current error rate  
40  
41 629 of the MinION or a small improvement in the error correction step would probably allow strain-  
42  
43 630 level identification even in this case.

47 631 In conclusion, using either the Sourmash and MetaMaps tools for metagenomic read-  
48  
49 632 based strain identification or LINbase for assembly-based strain-level identification, putative  
50  
51 633 strain-level identification was possible and was confirmed by culture-dependent genome-based  
52  
53 634 identification. However, it was not yet possible to reach the same high-confidence strain-level  
54  
55 635 identification of culture-dependent genome-based identification because of the absence of

1  
2  
3 636 appropriate strain-level databases for the read-based tools and because of the currently still high  
4  
5 637 error rate of the MinION™(version 19.05.0) when using assembly-based identification. Therefore,  
6  
7 638 at this point, we consider culture-independent metagenomic sequencing with the MinION™ an  
8  
9 639 excellent approach to obtain results when high confidence strain-level identification is not required  
10  
11 640 or when a culture-dependent genome-based identification is used as a follow-up. However,  
12  
13 641 considering the large and active user community of the MinION™ sequencer and the continued  
14  
15 642 development of new versions of the MinION™, we expect improvements in the precision at which  
16  
17 643 the MinION™ can distinguish nucleotides from each other, in base-calling algorithms, in error  
18  
19 644 correction, and in tool development for read-based identification. Together, these improvements  
20  
21 645 can be expected to take us to high-confidence strain-level identification of bacterial plant  
22  
23 646 pathogens from metagenomic sequences in the near future.  
24  
25

26 647

#### 27 28 648 **Author contributions**

29  
30 649 BAV and SL developed the project. MEML performed most of the wet-lab experiments. MAF and  
31  
32 650 PS did most of the bioinformatics analyses. SY contributed to the wet-lab experiments. LT and  
33  
34 651 CH, under supervision from BAV and LSH, developed LINbase. BAV, with contributions from  
35  
36 652 MEML, MAF, PS, and SL wrote the manuscript. All authors read and approved the final version  
37  
38 653 of the manuscript.  
39  
40

41 654

#### 42 43 655 **Conflict of Interest**

44  
45 656 LINbase uses the trademarks Life Identification Number® and LIN®, which are registered by This  
46  
47 657 Genomic Life, Inc. LSH and BAV report in accordance with Virginia Tech policies and procedures  
48  
49 658 and their ethical obligation as researchers that they have a financial interest in This Genomic Life,  
50  
51 659 Inc. Therefore, their financial interests may be affected by the research reported in this  
52  
53 660 manuscript. They have disclosed those interests fully to Virginia Tech, and they have in place an  
54  
55 661 approved plan for managing any potential conflicts arising from this relationship.  
56  
57

662

**Funding**

This study was supported by the College of Agriculture and Life Sciences at Virginia Polytechnic Institute and State University and by the National Science Foundation (IOS-1754721). Funding to BAV and SL was also provided in part by the Virginia Agricultural Experiment Station and the Hatch Program of the National Institute of Food and Agriculture, US Department of Agriculture.

668

**Acknowledgements**

The authors acknowledge Advanced Research Computing (ARC) at Virginia Tech for providing computational resources and technical support that have contributed to the results reported within this paper. URL: <http://www.arc.vt.edu>

673

**Literature cited**

Almeida, N. F., Yan, S., Cai, R., Clarke, C. R., Morris, C. E., Schaad, N. W., et al. 2010.

PAMDB, a multilocus sequence typing and analysis database and website for plant-associated microbes. *Phytopathology*. 100:208–215.

Andrews, S. 2010. Babraham bioinformatics-FastQC a quality control tool for high throughput sequence data. URL: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

(accessed 06. 12. 2018). Available at:

<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.

Badial, A. B., Sherman, D., Stone, A., Gopakumar, A., Wilson, V., Schneider, W., et al. 2018.

Nanopore Sequencing as a Surveillance Tool for Plant Pathogens in Plant and Insect

Tissues. *Plant Disease*. 102:1648–1652 Available at: [http://dx.doi.org/10.1094/pdis-04-17-](http://dx.doi.org/10.1094/pdis-04-17-0488-re)

0488-re.

Brown, C. T., and Irber, L. 2016. sourmash: a library for MinHash sketching of DNA. *J. Open*

Source Software. 1:27.

- 1  
2  
3 688 Bushnell, B. 2015. BBMap. Available at: <https://sourceforge.net/projects/bbmap/>.
- 4  
5 689 Cai, R., Lewis, J., Yan, S., Liu, H., Clarke, C. R., Campanile, F., et al. 2011. The plant pathogen  
6  
7 690 *Pseudomonas syringae* pv. *tomato* is genetically monomorphic and under strong selection  
8  
9 691 to evade tomato immunity. PLoS Pathog. 7:e1002130.
- 10  
11 692 Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. 2009.  
12  
13 693 BLAST+: architecture and applications. BMC Bioinformatics. 10:421.
- 14  
15 694 Chalupowicz, L., Dombrovsky, A., Gaba, V., Luria, N., Reuven, M., Beerman, A., et al. 2019.  
16  
17 695 Diagnosis of plant diseases using the Nanopore sequencing platform. Plant Pathol. 68:229–  
18  
19 696 238.
- 20  
21 697 Clarke, C. R., Cai, R., Studholme, D. J., Guttman, D. S., and Vinatzer, B. A. 2010.  
22  
23 698 *Pseudomonas syringae* strains naturally lacking the classical *P. syringae* *hrp/hrc* Locus are  
24  
25 699 common leaf colonizers equipped with an atypical type III secretion system. Mol. Plant.  
26  
27 700 Microbe. Interact. 23:198–210.
- 28  
29 701 Dijkshoorn, L., Ursing, B. M., and Ursing, J. B. 2000. Strain, clone and species: comments on  
30  
31 702 three basic concepts of bacteriology. J. Med. Microbiol. 49:397–401.
- 32  
33 703 Dilthey, A. T., Jain, C., Koren, S., and Phillippy, A. M. 2019. Strain-level metagenomic  
34  
35 704 assignment and compositional estimation for long reads with MetaMaps. Nat. Commun.  
36  
37 705 10:3066.
- 38  
39 706 Fang, Y., and Ramasamy, R. P. 2015. Current and Prospective Methods for Plant Disease  
40  
41 707 Detection. Biosensors. 5:537–561.
- 42  
43 708 Feil, H., Feil, W. S., Chain, P., Larimer, F., DiBartolo, G., Copeland, A., et al. 2005. Comparison  
44  
45 709 of the complete genome sequences of *Pseudomonas syringae* pv. *syringae* B728a and pv.  
46  
47 710 *tomato* DC3000. Proc. Natl. Acad. Sci. U. S. A. 102:11064–11069.
- 48  
49 711 Gardan, L., Shafik, H., Belouin, S., Broch, R., Grimont, F., and Grimont, P. A. 1999. DNA  
50  
51 712 relatedness among the pathovars of *Pseudomonas syringae* and description of  
52  
53 713 *Pseudomonas tremae* sp. nov. and *Pseudomonas cannabina* sp. nov. (ex Sutic and Dowson

- 1  
2  
3 714 1959). *Int. J. Syst. Bacteriol.* 49 Pt 2:469–478.  
4  
5 715 Hu, Y., Green, G. S., Milgate, A. W., Stone, E. A., Rathjen, J. P., and Schwessinger, B. 2019.  
6  
7 716 Pathogen Detection and Microbiome Analysis of Infected Wheat Using a Portable DNA  
8  
9 717 Sequencer. *Phytobiomes Journal.* 3:92–101.  
10  
11 718 Jain, M., Olsen, H. E., Paten, B., and Akeson, M. 2016. The Oxford Nanopore MinION: delivery  
12  
13 719 of nanopore sequencing to the genomics community. *Genome Biol.* 17:239.  
14  
15 720 Jones, J. B., Lacy, G. H., Bouzar, H., Stall, R. E., and Schaad, N. W. 2004. Reclassification of  
16  
17 721 the Xanthomonads associated with bacterial spot disease of tomato and pepper. *Syst. Appl.*  
18  
19 722 *Microbiol.* 27:755–762.  
20  
21 723 Juul, S., Izquierdo, F., Hurst, A., Dai, X., Wright, A., Kulesha, E., et al. 2015. What's in my pot?  
22  
23 724 Real-time species identification on the MinION. *bioRxiv.* :030742.  
24  
25 725 Kim, D., Song, L., Breitwieser, F. P., and Salzberg, S. L. 2016. Centrifuge: rapid and sensitive  
26  
27 726 classification of metagenomic sequences. *Genome Res.* 26:1721–1729.  
28  
29 727 Konstantinidis, K. T., and Tiedje, J. M. 2005. Genomic insights that advance the species  
30  
31 728 definition for prokaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 102:2567–2572.  
32  
33 729 Li, H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 34:3094–  
34  
35 730 3100.  
36  
37 731 Li, H. 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long  
38  
39 732 sequences. *Bioinformatics.* 32:2103–2110.  
40  
41 733 Loit, K., Adamson, K., Bahram, M., Puusepp, R., Anslan, S., Kiiker, R., et al. 2019. Relative  
42  
43 734 performance of Oxford Nanopore MinION vs. Pacific Biosciences Sequel third-generation  
44  
45 735 sequencing platforms in identification of agricultural and forest pathogens. *bioRxiv.* :592972  
46  
47 736 Available at: <https://www.biorxiv.org/content/10.1101/592972v1.abstract> [Accessed  
48  
49 737 September 8, 2019].  
50  
51 738 Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., and Zimin, A. 2018.  
52  
53 739 MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.*  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 740 14:e1005944.  
4  
5 741 Mechan-Llontop, M. E., Tian, L., Bernal-Galeano, V., Reeves, E., Hansen, M. A., Bush, E., et al.  
6  
7 742 2019. Assessing the potential of culture-independent 16S rRNA microbiome analysis in  
8  
9 743 disease diagnostics: the example of *Dianthus gratianopolitanus* and *Robbsia andropogonis*.  
10  
11 744 European Journal of Plant Pathology. Available at: <http://dx.doi.org/10.1007/s10658-019->  
12  
13 745 01850-8 [Accessed September 16, 2019].  
14  
15 746 MinION brochure. 2019a. Oxford Nanopore Technologies. Available at:  
16  
17 747 <http://nanoporetech.com/resource-centre/minion-brochure> [Accessed September 14, 2019].  
18  
19 748 MinION brochure. 2019b. Oxford Nanopore Technologies. Available at:  
20  
21 749 <http://nanoporetech.com/resource-centre/minion-brochure> [Accessed September 14, 2019].  
22  
23 750 Nadon, C., Van Walle, I., Gerner-Smidt, P., Campos, J., Chinen, I., Concepcion-Acevedo, J., et  
24  
25 751 al. 2017. PulseNet International: Vision for the implementation of whole genome sequencing  
26  
27 752 (WGS) for global food-borne disease surveillance. Euro Surveill. 22 Available at:  
28  
29 753 <http://dx.doi.org/10.2807/1560-7917.ES.2017.22.23.30544>.  
30  
31 754 Ondov, B.D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S. and  
32  
33 755 Phillippy, A. M., 2016. Mash: fast genome and metagenome distance estimation using  
34  
35 756 MinHash. Genome biology. 17(1):132.  
36  
37 757 Ottesen, A. R., González Peña, A., White, J. R., Pettengill, J. B., Li, C., Allard, S., et al. 2013.  
38  
39 758 Baseline survey of the anatomical microbial ecology of an important food plant: *Solanum*  
40  
41 759 *lycopersicum* (tomato). BMC Microbiol. 13:114.  
42  
43 760 protocols.io. High molecular weight DNA extraction from all kingdoms. Available at:  
44  
45 761 <https://www.protocols.io/groups/high-molecular-weight-dna-extraction-from-all-kingdoms>  
46  
47 762 [Accessed November 13, 2019].  
48  
49 763 Radhakrishnan, G. V., Cook, N. M., Bueno-Sancho, V., Lewis, C. M., Persoons, A., Mitiku, A.  
50  
51 764 D., et al. 2019. MARPLE, a point-of-care, strain-level disease diagnostics and surveillance  
52  
53  
54  
55  
56  
57  
58  
59  
60



- 1  
2  
3 765 tool for complex fungal pathogens. BMC Biology. 17 Available at:  
4  
5 766 <http://dx.doi.org/10.1186/s12915-019-0684-y>.  
6  
7 767 Rees-George, J., Vanneste, J. L., Cornish, D. A., Pushparajah, I. P. S., Yu, J., Templeton, M.  
8  
9 768 D., et al. 2010. Detection of *Pseudomonas syringae* pv. *actinidiae* using polymerase chain  
10  
11 769 reaction (PCR) primers based on the 16S-23S rDNA intertranscribed spacer region and  
12  
13 770 comparison with PCR primers based on other gene regions. Plant Pathology. 59:453–464  
14  
15 771 Available at: <http://dx.doi.org/10.1111/j.1365-3059.2010.02259.x>.  
16  
17 772 Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., et al. 2009.  
18  
19 773 Database resources of the National Center for Biotechnology Information. Nucleic Acids  
20  
21 774 Res. 37:D5–15.  
22  
23 775 Schwartz, A. R., Potnis, N., Timilsina, S., Wilson, M., PatanĀ©, J., Martins, J., et al. 2015.  
24  
25 776 Phylogenomics of *Xanthomonas* field strains infecting pepper and tomato reveals diversity in  
26  
27 777 effector repertoires and identifies determinants of host specificity. Frontiers in Microbiology.  
28  
29 778 6 Available at: <http://dx.doi.org/10.3389/fmicb.2015.00535>.  
30  
31 779 Tedersoo, L., Drenkhan, R., Anslan, S., Morales-Rodriguez, C., and Cleary, M. 2019.  
32  
33 780 High-throughput identification and diagnostics of pathogens and pests: Overview and  
34  
35 781 practical recommendations. Molecular Ecology Resources. 19:47–76 Available at:  
36  
37 782 <http://dx.doi.org/10.1111/1755-0998.12959>.  
38  
39 783 Tian, L., Huang, C., Heath, L. S., and Vinatzer, B. A. 2019. LINbase: A Web service for  
40  
41 784 genome-based identification of microbes as members of crowdsourced taxa. bioRxiv.  
42  
43 785 Available at: <https://www.biorxiv.org/content/10.1101/752212v1.abstract>.  
44  
45 786 Tinivella, F., Gullino, M. L., and Stack, J. P. 2008. The Need for Diagnostic Tools and  
46  
47 787 Infrastructure. In *Crop Biosecurity*, Springer Netherlands, p. 63–71.  
48  
49 788 Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. 2017. Fast and accurate de novo genome  
50  
51 789 assembly from long uncorrected reads. Genome Research. 27(5), 737–746.  
52  
53 790 Wick, R. R., Judd, L. M., Gorrie, C. L., and Holt, K. E. 2017. Unicycler: Resolving bacterial  
54  
55  
56  
57  
58  
59  
60



- 1  
2  
3 791 genome assemblies from short and long sequencing reads. PLoS Comput. Biol.  
4  
5 792 13:e1005595.  
6  
7 793 Williamson, L., Nakaho, K., Hudelson, B., and Allen, C. 2002. *Ralstonia solanacearum* Race 3,  
8  
9 794 Biovar 2 Strains Isolated from Geranium Are Pathogenic on Potato. Plant Dis. 86:987–991.  
10  
11 795 Yan, S., Liu, H., Mohr, T. J., Jenrette, J., Chiodini, R., Zaccardelli, M., et al. 2008. Role of  
12  
13 796 Recombination in the Evolution of the Model Plant Pathogen *Pseudomonas syringae* pv.  
14  
15 797 *tomato* DC3000, a Very Atypical Tomato Strain. Applied and Environmental Microbiology.  
16  
17 798 74:3171–3181 Available at: <http://dx.doi.org/10.1128/aem.00180-08>.  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

## Tables

**Table 1.** Description of samples used in this study.

Sample Name	Short description	DNA concentration of samples (ng/ul)	Fraction of flow cell used	# reads base-called	Total length of reads base-called	% of reads classified as bacteria (based on WIMP)	Mean read length in bp	Max read length in bp	% reads >1000bp
L-K40	Tomato inoculated with <i>Pto</i> K40 in the laboratory	325.2	1	1,377,617	4.18 Gb	89%	3,037	66,015	64%
L-mix	Tomato inoculated with four <i>P. syringae</i> strains in the laboratory	450.4	1	1,006,978	4.16 Gb	95%	4,130	67,174	74%
L-mock	Non-inoculated tomato plant grown in the laboratory	33.6	1/7	82,412	103.22 Mb	8%	1,252	19,754	40%
L-culture-mix	Equal mix of 4 <i>P. syringae</i> strains grown in liquid culture	147.5	1/6	54,124	155.93 Mb	93%	2,880	76,060	39%
F1-bs	Tomato field sample with symptoms of bacterial spot	562	1/7	137,497	588.50 Mb	81%	4,280	55,436	73%
F2-bs	Tomato field sample with symptoms of bacterial spot	500.2	1/7	90,185	498.68 Mb	80%	5,529	65,598	74%
F3-bs	Tomato field sample with symptoms of bacterial spot	332.5	1/7	100,956	423.16 Mb	78%	4,191	59,405	68%
F4-bs	Tomato field sample with symptoms of bacterial spot	319.8	1/7	74,615	289.36 Mb	81%	3,878	51,268	70%

F5- Septoria a	Tomato field sample with symptoms of Septoria leaf spot	75.8	1/7	73,432	226.721 Mb	50%	3,087	43,967	59%
F6- healthy	Tomato field sample with no symptoms	29.1	1/7	35,923	66,58 Mb	31%	1,853	29,617	46%
F7-bs	Tomato field sample with symptoms of bacterial spot	331.8	1/7	118,391	432.08 Mb	75%	3,649	48,335	64%
F8-bs	Tomato field sample with symptoms of bacterial spot	154.2	1/2	106,059	371.84 Mb	70%	3,505	33,472	71%

**Table 2.** Relative abundance results (top three hits) obtained with MetaMaps and Sourmash using a custom genome database of bacterial tomato pathogens and closely related isolates.

Sample	rank	MetaMaps	%	Sourmash	%
L-K40	1	<i>Pto</i> T1 ( <i>Pto</i> strain T1)	70.94	<i>Pto</i> T1 ( <i>Pto</i> strain T1)	71.65
	2	<i>Pto</i> NCPPB1108 ( <i>Pto</i> strain T1)	15.91	<i>P. syringae</i> pv. <i>actinidiae</i>	3.67
	3	<i>Pto</i> DC3000 ( <i>Pto</i> strain DC3000)	7.81	<i>P. syringae</i>	2.44
L-mix	1	<i>Pto</i> T1 ( <i>Pto</i> strain T1)	69.48	<i>Pto</i> T1 ( <i>Pto</i> strain T1)	65.98
	2	<i>Pto</i> NCPPB 1108 ( <i>Pto</i> strain T1)	15.23	<i>Pto</i> DC3000 ( <i>Pto</i> strain DC3000)	16.01
	3	<i>Pto</i> PT23	6.90	<i>P. syringae</i> pv. <i>actinidiae</i>	2.56
L-mock	1	<i>Clavibacter michiganensis</i> <sup>1</sup>	13.30	*no matches*	
	2	<i>Xp</i>	11.39	*no matches*	
	3	<i>Ralstonia solanacearum</i>	8.86	*no matches*	
L-culture-mix	1	<i>Pto</i> DC3000 ( <i>Pto</i> strain DC3000)	38.90	<i>Pto</i> DC300 ( <i>Pto</i> strain DC3000)	75.17
	2	<i>Pto</i> T1 ( <i>Pto</i> strain T1)	27.48	<i>Pto</i> T1 ( <i>Pto</i> strain T1)	19.63
	3	<i>Pto</i> NCPPB 1108 ( <i>Pto</i> strain T1)	9.07	<i>Pto</i> PT23	1.03
F1-bs	1	<i>Xp</i> TB9 ( <i>Xp</i> group 2)	29.37	<i>Xp</i> TB15 ( <i>Xp</i> group 2)	95.18
	2	<i>Xp</i> Xp9-5 ( <i>Xp</i> group 2)	28.03	<i>Xp</i> Xp17-12	1.05
	3	<i>Xp</i> Xp7-12	14.97	<i>X. campestris</i> pv. <i>durantae</i>	0.79
F2-bs	1	<i>Xp</i> Xp9-5 ( <i>Xp</i> group 2)	15.97	<i>Xp</i> strain Xp9-5 ( <i>Xp</i> group 2)	90.72
	2	<i>Xp</i> TB9 ( <i>Xp</i> group 2)	15.14	<i>Xp</i> strain Xp17-12	4.19
	3	<i>Xp</i> Xp7-12	10.38	<i>X. arboricola</i> pv. <i>pruni</i>	1.83
F3-bs	1	<i>Xp</i> Xp17-12	50.59	<i>Xp</i> strain Xp17-12	97.76
	2	<i>Xp</i> 91-118	9.00	<i>Xp</i> strain Xp9-5 ( <i>Xp</i> group 2)	1.27
	3	<i>Xp</i> LH3	4.67	<i>X. campestris</i> pv. <i>durantae</i>	0.98
F4-bs	1	<i>Xp</i> TB9 ( <i>Xp</i> group 2)	22.38	<i>Xp</i> TB15 ( <i>Xp</i> group 2)	97.28
	2	<i>Xp</i> Xp9-5 ( <i>Xp</i> group 2)	19.30	<i>Xp</i> Xp9-5 ( <i>Xp</i> group 2)	2.11
	3	<i>Xp</i> TB15 ( <i>Xp</i> group 2)	18.80	<i>X. campestris</i> pv. <i>viticola</i>	0.61
F5-Septoria	1	<i>X. campestris</i>	30.45	<i>X. arboricola</i>	57.08
	2	<i>X. arboricola</i>	25.60	<i>X. arboricola</i>	14.76
	3	<i>X. pisi</i>	2.78	<i>Xp</i> TB9	9.59
F6-healthy	1	<i>Xp</i> Xp9-5 ( <i>Xp</i> group 2)	11.70	<i>Xp</i> TB15 ( <i>Xp</i> group 2)	98.13
	2	<i>Xp</i> TB9 ( <i>Xp</i> group 2)	11.47	<i>Xp</i> LH3	1.87
	3	<i>Xp</i> Xp7-12	10.82	*no matches	
F7-bs	1	<i>Xp</i> TB9 ( <i>Xp</i> group 2)	23.40	<i>Xp</i> TB15 ( <i>Xp</i> group 2)	89.80
	2	<i>Xp</i> Xp9-5 ( <i>Xp</i> group 2)	19.15	<i>X. arboricola</i>	5.47
	3	<i>Xp</i> TB15 ( <i>Xp</i> group 2)	17.28	<i>X. campestris</i>	1.54
F8-bs	1	<i>Xp</i> Xp9-5 ( <i>Xp</i> group 2)	26.51	<i>Xp</i> Xp9-5 ( <i>Xp</i> group 2)	94.17
	2	<i>Xp</i> TB9 ( <i>Xp</i> group 2)	17.48	<i>Xp</i> TB15 ( <i>Xp</i> group 2)	1.62
	3	<i>Xp</i> Xp17-12	15.23	<i>Xp</i> Xp17-12	1.05

<sup>1</sup> strain names are only reported for tomato pathogens

**Table 3.** Description of metagenomic assemblies.

Sample name	Total number of contigs	Total assembly length in bp	Mean contig length in bp	Longest contig in bp	2nd longest contig in bp
L-K40	24	6,777,088	282,378	6,239,052	143,110
L-mix	73	8,849,360	121,224	6,267,841	121,245
L-mock	8	117,988	14,748	64,285	12,027
L-culture-mix	20	5,896,134	294,806	777,233	631,884
F1-bs	92	12,801,147	139,142	5,084,548	898,279
F2-bs	131	8,667,146	66,161	4,444,689	280,302
F3-bs	49	12,118,489	247,316	2,320,098	1,193,622
F4-bs	18	5,216,728	289,818	1,172,667	925,913
F5-Septoria	9	122,995	13,666	38,461	25,303
F6-healthy	4	21,571	5,392	8,666	7,821
F7-bs	35	5,784,684	165,276	5,146,049	57,056
F8-bs	10	5,449,373	544,937	2,745,990	2,264,789

**Table 4.** LINbase results for two longest contigs

Sample	Longest contig (ANI %)	Taxon membership of longest contig <sup>1</sup>	Second longest contig (ANI %)	Taxon membership of 2. longest contig <sup>1</sup>	Two longest contigs merged (ANI %)	Taxon membership of merged contigs <sup>1</sup>
L-K40	<i>Pto</i> BAV1020 (99.76)	<i>Pto</i> strain T1	NA	NA	<i>Pto</i> BAV1020 (99.62)	<i>Pto</i> strain T1
L-mix	<i>Pto</i> BAV1020 (99.77)	<i>Pto</i> strain T1	NA	NA	<i>Pto</i> K40 (99.53)	<i>Pto</i> strain T1
L-culture-mix	<i>Pc</i> ICMP19117 (97.42)	<i>Pc</i>	<i>Ps</i> UB0390 (97.70)	<i>Ps</i>	<i>Pc</i> ICMP19117 (97.50)	<i>Pc</i>
F1-bs	<i>Xp</i> 10-13 (99.84)	<i>Xp</i> group 2	NA	NA	<i>Xp</i> 8-16 (99.85)	<i>Xp</i> group 2
F2-bs	<i>Xp</i> GEV2117 (99.78)	<i>Xp</i> group 2	<i>Xp</i> 7-12 (99.76)	<i>Xp</i>	<i>Xp</i> GEV1063 (99.78)	<i>Xp</i> group 2
F3-bs	<i>Pf</i> Pf0-1 (95.08)	<i>Pf</i>	<i>Pf</i> Pf0-1 (95.01)	<i>Pf</i>	<i>Pf</i> Pf0-1 (95.05)	<i>Pf</i>
F4-bs	<i>Xp</i> 91-118 (99.62)	<i>Xp</i>	<i>Xp</i> 91-118 (99.75)	<i>Xp</i>	<i>Xp</i> 91-118 (99.62)	<i>Xp</i>
F7-bs	<i>Xp</i> 8-16 (99.84)	<i>Xp</i> group 2	NA	NA	<i>Xp</i> GEV2116 (99.83)	<i>Xp</i> group 2
F8-bs	<i>Xp</i> 10-13 (99.76)	<i>Xp</i> group 2	<i>Xp</i> GEV2117 (99.78)	<i>Xp</i> group 2	<i>Xp</i> Xp10-13 (99.79)	<i>Xp</i> group 2
BAV6163	<i>Xp</i> GEV1063 (99.98)	<i>Xp</i> group 2				
BAV6164	<i>Xp</i> GEV1063 (99.98)	<i>Xp</i> group 2				

<sup>1</sup> based on taxon membership of the best hit

*Ps* = *P. syringae*, *Pf* = *Pseudomonas fluorescens*, *Pc* = *Pseudomonas congelans*, *Xp* = *X. perforans*

NA – Not available, second contig too short for identification

## Supplementary Tables

**Supplementary Table 1.** List of genomes used in the custom database.

**Supplementary Table 2.** Relative abundance values at the species level for all samples obtained with WIMP, Sourmash, and MetaMaps.

**Supplementary Table 3.** Example run times for WIMP, Sourmash, and MetaMaps.

**Supplementary Table 4.** BLASTN results for contigs of assembled metagenomes.

## Figure legends

**Figure 1.** Diseased tomato plants (A) Symptoms caused by *Pseudomonas syringae* pv *tomato* isolate K40 (strain *Pto* T1) in a laboratory-inoculation assay and (B) Bacterial spot symptoms in naturally infected plants during a disease outbreak on the Eastern Shore of Virginia.

**Figure 2.** Screenshot of the output from the ONT tool WIMP showing the taxonomy assignment for sample L-K40. A taxonomy tree is depicted on the left and the distribution of reads across domains is shown on the right.

**Figure 3.** Bar graph showing the comparison of results at the species level using the read-based programs WIMP, Sourmash and MetaMaps. Each barplot corresponds to individual lab samples used in the study. A = L-K40, B = L-mix, C = L-mock, and D = L-culture-mix. Relative abundance values are expressed as percentages of all sequences classified as bacteria.

**Figure 4.** Bar graph showing the comparison of results at the species level using the read-based programs WIMP, Sourmash and MetaMaps. Each barplot corresponds to individual field samples used in the study. A = F1-bs, B = F2-bs, C = F3-bs, D = F4-bs, E = F5-Septoria, F = F6-healthy, G = F7-bs and H= F8-bs. Relative abundance values are expressed as percentages of all sequences classified as bacteria.

**Figure 5.** Relative genome percentage abundance for each sample based on BLASTN using contigs as query against a custom genome database. All hits were filtered to e-values less than

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

or equal to 0.01 and the longest hit for each contig was considered to be the best hit.

For Peer Review



1  
2  
3  
4  
5  
6  
7  
8  
9  
10 **1 Strain-level identification of bacterial tomato pathogens directly from metagenomic**  
11 **2 sequences**  
12  
13

14 4 Marco E. Mechan Llontop<sup>1\*</sup>, Parul Sharma<sup>1,2\*</sup>, Marcela Aguilera Flores<sup>1,2\*</sup>, Shu Yang<sup>1</sup>, Jill Pollok<sup>1,3</sup>,  
15 Long Tian<sup>1</sup>, Chenjie Huang<sup>4</sup>, Steve Rideout<sup>1,3</sup>, Lenwood S. Heath<sup>4</sup>, Song Li<sup>1</sup>, Boris A. Vinatzer<sup>1</sup>  
16  
17

18 6  
19 7 <sup>1</sup> School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA

20 8 <sup>2</sup> Graduate program in Genetics, Bioinformatics, and Computational Biology, Virginia Tech,  
21 Blacksburg, VA  
22

23 9  
24 10 <sup>3</sup> Virginia Tech Eastern Shore Agricultural Research and Extension Center, Painter, VA, USA

25 11 <sup>4</sup> Department of Computer Sciences, Virginia Tech, Blacksburg, VA  
26  
27

28 12  
29 13 \*These authors contributed equally  
30  
31

32 15 Corresponding authors: Boris A. Vinatzer and Song Li

33 16 E-mail addresses: [vinatzer@vt.edu](mailto:vinatzer@vt.edu) [songli@vt.edu](mailto:songli@vt.edu)  
34

35 17 Phone number: +1 540 231 2126  
36

37 18 B.A. Vinatzer ORCID: 0000-0003-4612-225X  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60**Abstract**

Routine strain-level identification of plant pathogens directly from symptomatic tissue could significantly improve plant disease control and prevention. Here we tested the Oxford Nanopore Technologies (ONT) MinION™ sequencer for metagenomic sequencing of tomato plants either artificially inoculated with a known strain of the bacterial speck pathogen *Pseudomonas syringae* pv. *tomato* (*Pto*), or collected in the field and showing bacterial spot symptoms caused by either one of four *Xanthomonas* species. After species-level identification using ONT's WIMP software and the third party tools Sourmash and MetaMaps, we used Sourmash and MetaMaps with a custom database of representative genomes of bacterial tomato pathogens to attempt strain-level identification. In parallel, each metagenome was assembled and the longest contigs were used as query with the genome-based microbial identification Web service LINbase. Both the read-based and assembly-based approaches correctly identified *Pto* strain T1 in the artificially inoculated samples. The pathogen strain in most field samples was identified as a member of *Xanthomonas perforans* group 2. This result was confirmed by whole genome sequencing of colonies isolated from one of the samples. Although in our case, metagenome-based pathogen identification at the strain-level was achieved, caution still needs to be exerted when interpreting strain-level results because of the challenges inherent to assigning reads to specific strains and the error rate of nanopore sequencing.

## 39 Introduction

40 Early detection of plant disease outbreaks and accurate plant disease diagnosis are prerequisites  
41 of efficient plant disease control and prevention (Tinivella et al. 2008). In many cases, an  
42 experienced plant pathologist can quickly diagnose a disease based on symptoms. However,  
43 visual diagnosis does not identify the causative agent at the strain-level. For example, three  
44 different strains of the plant pathogen *Pseudomonas syringae* pathovar (pv.) *tomato* (*Pto*) cause  
45 indistinguishable bacterial speck disease symptoms in tomato (Cai et al. 2011). Sometimes, visual  
46 diagnosis cannot even identify a pathogen at the species level. For example, four different species  
47 of the genus *Xanthomonas* cause indistinguishable bacterial spot disease symptoms on tomato  
48 (*Solanum lycopersicum*) leaves (Jones et al. 2004). Note that in this article, we use the term  
49 "strain" as an intraspecific, monophyletic group of bacteria, which have a very recent common  
50 ancestor and are thus genotypically and phenotypically more similar to each other than to other  
51 members of the same species (Dijkshoorn et al. 2000). To avoid confusion, we use the term  
52 "isolate" instead of "strain" when referring to a pure culture of bacteria isolated on a specified date  
53 at a specified geographic location from a specific plant.

54 While most disease control measures may be the same for different pathogen strains or  
55 species, depending on the precise identity of the pathogen, additional control measures may need  
56 to be undertaken. For example, different strains of the same pathogen species may have different  
57 host ranges. Therefore, it may be necessary to avoid certain crop rotations or to eliminate certain  
58 weeds depending on the identity of the strain that causes a disease and its specific host range.  
59 In the case of *Pto*, strain T1 causes disease only in tomato while strain DC3000 causes disease  
60 in tomato and in leafy greens of the family *Brassicaceae* (Yan et al. 2008). Strain DC3000 could  
61 thus spread from tomato fields to leafy green fields, cause disease in a leafy green planted after  
62 tomato, and/or survive in weeds that belong to the *Brassicaceae* family. In other cases, identifying  
63 a pathogen to strain level could even trigger eradication procedures to stop further spread of the  
64 disease. For example, this would happen if the select agent *Ralstonia solanacearum* Race 3

1  
2  
3  
4  
5  
6  
7  
8  
9  
10 65 Biovar 2 were to be identified as the causative agent of bacterial wilt disease outbreak in the USA  
11 66 (Williamson et al. 2002). Fast strain-level plant pathogen identification would thus add significant  
12 67 value to plant disease diagnostics.

14 68 Many molecular tools have been developed over the years for pathogen identification and  
15 69 they all have their strengths and weaknesses (Fang and Ramasamy 2015). Many of them depend  
16 70 on a pure pathogen culture and thus require lengthy procedures to isolate and culture the  
17 71 pathogen from the plant tissue. Moreover, many of them cannot identify pathogens at the strain  
18 72 level. Gene sequence-based techniques, such as multilocus sequence typing/analysis (MLST/A)  
19 73 (Almeida et al. 2010), can identify a pathogen to strain-level but usually require pure cultures.  
20 74 Moreover, gene sequence-based techniques depend on previous species-level identification  
21 75 because different species require different primers to amplify the genes to be sequenced by  
22 76 polymerase chain reaction (PCR), for example see (Rees-George et al. 2010). One alternative  
23 77 gene-based method is to amplify the 16S rRNA gene directly from DNA extracted from plant tissue  
24 78 and to identify the putative pathogen based on its 16S rRNA sequence. We have recently tested  
25 79 this method but not found it to be suitable because of its low resolution (Mechan-Llontop et al.  
26 80 2019).

35 81 Whole genome sequencing (WGS) does not require PCR and strain-level identification is  
36 82 now routine practice in the surveillance of food-borne pathogen outbreaks in several countries  
37 83 (Nadon et al. 2017). With the drop ~~of~~ sequencing cost and development of genome databases  
38 84 that contain strain-level classification of plant pathogens, WGS now represents a real possibility  
39 85 in plant disease diagnostics. For example, LINbase at linbase.org (Tian et al. 2019) contains  
40 86 precise genome-based circumscriptions for many bacterial plant pathogens from the genus level  
41 87 to the strain level. Genome sequences of unknown isolates can be identified as members of  
42 88 circumscribed plant pathogens based on how similar they are at the whole genome level,  
43 89 measured as Average Nucleotide Identity (ANI) (Konstantinidis and Tiedje 2005), to the other  
44 90 members of these taxa. However, the limitation of WGS is its dependence on pure cultures.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10 91 Metagenomic sequencing consists in extracting DNA directly from plant tissue followed by  
11 92 sequencing all DNA present in the sample. Compared to WGS, the two main advantages of this  
12 93 approach are that (1) it is much faster because it does not require lengthy pathogen isolation and  
13 94 culturing procedures; and (2) it does not require much prior knowledge about the pathogen since  
14 95 any pathogen, besides RNA viruses, can be detected with this method. However, the main  
15 96 challenge of this approach is that the obtained DNA sequences also contain host plant sequences  
16 97 and microbe sequences that do not belong to the pathogen. Therefore, obtaining sufficient  
17 98 sequences of the causative agent and identifying the causative agent among all the other potential  
18 99 causative agents present in the same plant requires optimized experimental methods for DNA  
19 100 extraction and sequencing and optimized algorithms and genome databases for precise pathogen  
20 101 identification.

21 102 The sequencing method that is currently most attractive for metagenomics-based  
22 103 pathogen identification is nanopore sequencing with the Oxford Nanopore Technologies (ONT)  
23 104 MinION™ device (Jain et al. 2016). The main strengths of this method are that (1) DNA can be  
24 105 prepared for sequencing with relatively short protocols ([ranging](#) from a few hours to less than an  
25 106 hour ([protocols.io](#)); <https://community.nanoporetech.com>), (2) the MinION™ sequencer is not  
26 107 much larger than a USB stick and can be used with a desktop or a laptop computer in the lab or  
27 108 even in the field, (3) it provides the first sequencing results within minutes from the start of a  
28 109 sequencing run, and (4) the output can reach over 10 gigabases of DNA sequences (more than  
29 110 1000 times the size of an individual bacterial genomes) after 48 hours (MinION brochure 2019a).  
30 111 However, the major weaknesses are (1) the [currently](#) high sequencing error rate of approximately  
31 112 10% (Tedersoo et al. 2019; Loit et al. 2019) and (2) that the sequencing hardware only works  
32 113 once at full capacity limiting reuse (MinION brochure 2019b).

33 114 Metagenomic sequencing with the MinION™ has already been used on several crops for  
34 115 identification of various pathogens (Chalupowicz et al. 2019) using ONT's software WIMP (Juul  
35 116 et al. 2015) and on wheat to identify various fungal pathogens (Hu et al. 2019) using the sequence

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

alignment tool BLASTN (Camacho et al. 2009) in combination with custom databases. The MinION™ has also been used for plant pathogen detection and identification starting from extracted RNA or DNA in combination with general or specific primers to increase the quantity of input for the MinION™ (Loit et al. 2019; Badial et al. 2018). However, in none of these studies, was strain-level identification attempted directly from sequencing metagenomic DNA without prior amplification.

Here we tested the MinION™ with tomato plants artificially inoculated with different strains of *Pseudomonas syringae*, including isolates of the *Pto* strains T1 and DC3000 (Cai et al. 2011), and with plants from tomato fields showing symptoms of natural infection with bacterial spot for which we did not know the *Xanthomonas* species that caused the infection. We then explored the precision of identification that can be achieved when using ONT's WIMP software and the third party tools Sourmash (Brown and Irber 2016) and MetaMaps (Dilthey et al. 2019) in combination with default and custom reference databases. We also assembled metagenomic sequences into contigs that were used as input to BLASTN (Camacho et al. 2009) and the LINbase Web service for genome-based microbial identification (Tian et al. 2019).

## Materials and Methods

### Laboratory-infected tomato plants

Seeds of tomato (*Solanum lycopersicum*) 'Rio Grande' were germinated in potting mix soil (Miracle-grow, OH, USA) under laboratory conditions with a long day period (16-h photoperiod) and infected at 4 weeks of age. *Pto* isolate K40 (belonging to strain T1), *Pto* isolate DC3000 (belonging to strain DC3000) (Cai et al. 2011), *P. syringae* pv. *syringae* B728a (Feil et al. 2005), and *P. syringae* 642 (Clarke et al. 2010) were grown in King's B solid medium at 28°C for 24 hours. Isolate *Pto* K40 was suspended at a concentration corresponding to an OD600 of 0.001 in 10 mM MgSO<sub>4</sub> for single-strain inoculation. For the mixed-strain inoculation, all four isolates were suspended at an OD600 of 0.001 in 10 mM MgSO<sub>4</sub> and pooled together in equal amounts before

1  
2  
3  
4  
5  
6  
7  
8  
9  
10 143 inoculation. Silwet L-77 was added to bacterial suspensions (0.025% vol/vol) to facilitate bacterial  
11 144 infection. Plants were placed in ziplock plastic bags for high humidity conditions for 24 hours  
12  
13 145 before inoculation. After plants were spray-inoculated with 10 ml of bacterial suspensions, they  
14 146 were placed back into the plastic bags for another 24 hours. Plants were processed for DNA  
15  
16 147 extraction ~~fourth~~ three days ~~after inoculation~~ after. Inoculation with 10mM MgSO<sub>4</sub> was included as  
17  
18 148 a mock treatment.

#### 19 149

#### 20 150 Naturally infected tomato plants

21 151 Five tomato plants with bacterial spot symptoms, one plant with symptoms of Septoria leaf spot,  
22  
23 152 and one plant without symptoms were collected on August 10, 2018, on the Eastern Shore of  
24  
25 153 Virginia (Accomack and Northampton counties). The diagnosis of bacterial spot was made by  
26  
27 154 matching symptomology in the field (chlorotic haloes surrounding leaf lesions) with the presence  
28  
29 155 of bacterial streaming microscopically at 200-X. Additionally, bacteria were cultured on King's  
30  
31 156 medium B (KB) media and were found to be non-fluorescent and of deep yellow color leading to  
32  
33 157 their identification as *Xanthomonas*. The diagnosis of Septoria leaf spot was confirmed by  
34  
35 158 microscopic identification of conidia at 200X. Plants were then shipped overnight to the Virginia  
36  
37 159 Tech campus in Blacksburg, VA, where they were processed for DNA extraction. Another set of  
38  
39 160 plants with bacterial spot symptoms were collected in May, 2019. Bacteria were isolated from  
40  
41 161 symptomatic leaves on ~~King's medium B~~. Plants and plates were shipped to the Virginia Tech  
42  
43 162 campus overnight where plants and bacterial colonies were processed for DNA extraction.

Commented [SL1]: Same as KB media?

#### 43 164 DNA extraction

44  
45 165 All plant samples used for DNA extraction are listed in Table 1. DNA extraction was performed  
46  
47 166 according to (Ottesen et al. 2013) with the following modifications. Briefly, wearing gloves, the top  
48  
49 167 of each plant sample (6 to 10 leaves from the top with or without stems) was collected using  
50  
51 168 clippers. The weight of samples was between 5 to 10 grams. After removing all the dirt from the

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

169 plant surface by shaking vigorously, each sample was placed in a 6-1/2"× 5-7/8" Ziploc® bag  
170 together with 300 ml sterilized double-distilled water (DDW). Samples were sonicated for 15  
171 minutes using a Branson 1510 Ultrasonic Cleaner. DNA was extracted with DNeasy®  
172 PowerWater® Kit (QIAGEN; Catalog # 14900-50-NF). All steps for DNA extraction were  
173 performed according to the kit's specifications, except that after adding 1 mL of the kit's solution  
174 PW1, the tube was incubated at 65°C for 15 minutes and then vortexed for 20 minutes.

175 DNA from [two plant-isolated bacteria](#) was extracted [from cultures derived from single](#)  
176 [colonies](#) with the Gentra® Puregene® Cell and Tissue Kit (Gentra Systems; Catalog # D5000).  
177 All steps for DNA extraction were performed according to the Gram-negative Bacteria protocol,  
178 except that cells were collected in 1 mL of sterilized DDW in a 1.5 ml microcentrifuge tube for the  
179 lysis step. For both extraction procedures, the concentration and purity of DNA was measured  
180 using a Thermo Scientific™ NanoDrop™ One<sup>C</sup> Spectrophotometer.

#### 182 DNA library preparation

183 Library preparation was performed according to the 1D Native barcoding genomic DNA protocols  
184 (EXP-NBD104, EXP-NBD114, and SQK-LSK108 or SQK-LSK109) provided by ONT. Sequencing  
185 libraries were prepared using the Ligation Sequencing Kit (ONT Ltd.; SQK-LSK109). For each  
186 run, NEBNext® Ultra™ II End Repair/dA-Tailing Module (New England Biolabs, Inc.; Catalog #  
187 E7546S) was used for DNA repair and end-prep for each sample. Repaired DNA was cleaned up  
188 by 1.5 volumes of AMPure XP beads, washed on a magnetic rack using freshly made 70%  
189 ethanol, and eluted with 25 µL nuclease-free water. 22.5 µL elute was used for barcoding by  
190 mixing with the Blunt/TA Ligase Master Mix (New England Biolabs, Inc.; Catalog # M0367S) and  
191 Native Barcode (Oxford Nanopore Technologies Ltd.; Native Barcoding Expansion Kit EXP-  
192 NBD104), followed by another wash step using 1.5 volumes of AMPure XP beads, and DNA was  
193 eluted in 26 µL nuclease-free water. Equimolar amounts of barcoded DNA were then pooled into  
194 a 1.5 mL microcentrifuge for ligation. Adapter ligation was performed by mixing the pooled



1  
2  
3  
4  
5  
6  
7  
8  
9  
10 195 barcoded sample with Adapter Mix (Oxford Nanopore Technologies Ltd.; SQK-LSK109),  
11 196 NEBNext® Quick Ligation Reaction Buffer (New England Biolabs, Inc.; Catalog # B6058S) and  
12 197 Quick T4 DNA Ligase (New England Biolabs, Inc.; Catalog # M2200S). Ligated DNA was cleaned  
13 198 up by one volume of AMPure XP beads, washed on a magnetic rack using Long Fragment Buffer  
14 199 (Oxford Nanopore Technologies Ltd.; SQK-LSK109), and eluted with 15 µL Elution Buffer (Oxford  
15 200 Nanopore Technologies Ltd.; SQK-LSK109).

19 201 ~~Sequencing was performed independently for each sample~~ Sequencing reactions were  
20 202 ~~performed independently for each run on a~~ ONT MinION™ flow cells (FLO-MIN106 R9 Version)  
21 203 connected to a Mk1B device (ONT Ltd.; MIN-101B) operated by the MinKNOW software ([version](#)  
22 204 [3.3.2](#) ~~latest version available~~). Each flow cell was primed with the priming buffer prepared by  
23 205 mixing 30 µL Flush Tether (ONT Ltd.; EXP-FLP001) with a tube of Flush Buffer (ONT Ltd.; EXP-  
24 206 FLP001). 12 µL of the final library mixed with Sequencing Buffer (ONT Ltd.; SQK-LSK109) and  
25 207 Library Loading Beads (ONT Ltd.; SQK-LSK109) was loaded onto the SpotON sample port of the  
26 208 flow cell in a dropwise fashion. The sequencing run was stopped after 48 hours.  
27 209

### 23 210 [Illumina genome sequencing and assembly](#)

24 211 Genomic DNA from isolated bacteria was used to prepare 350bp insert DNA libraries and  
25 212 sequence on an Illumina platform PE150 at Novogene Corporation Inc (Sacramento, CA). FastQC  
26 213 was used to assess the quality of the raw sequencing data (Andrews 2010). Adapter-trimming  
27 214 was performed using BBduk with the parameters 'k=23, mink=9, hdist=1, ktrim=r, minlength=100'  
28 215 (Bushnell 2015). Unicycler v0.4.7 with default parameters was used to *de novo* assemble the  
29 216 bacterial genomes (Wick et al. 2017).  
30 217

### 31 218 [Read-based metagenomic analysis](#)

32 219 *Guppy*

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

220 For all samples, the Fast5 files containing raw reads were base-called with the base-calling ONT  
221 software Guppy-cpu (v3.3.2) [available via the Nanopore website](#), which uses neural networks to  
222 translate raw signals into DNA sequences in fastq format, [with default parameters](#). ~~(available via~~  
223 ~~the Nanopore website~~[Nanopore websitehttps://community.nanoporetech.com](https://community.nanoporetech.com)). [All fastq files](#)  
224 [were deposited in the NCBI SRA database under PRJNA588037](#).

225 *What's in my pot? (WIMP)*

226 The ONT workflow WIMP (v2019.7.9), which uses Centrifuge (Kim et al. 2016) to assign taxonomy  
227 to reads in real-time, was used for species level identification in all samples. [The workflow uses](#)  
228 [bacterial, viral, and fungal genomes present in Refseq as the reference database](#).

229

230 *Sourmash*

231 Sourmash [is](#), a command-line tool used for k-mer based taxonomic classification ~~offer~~ genomes  
232 and metagenomes. [It uses, computes a MinHash sketching algorithm \(Ondov et al. 2016\)](#)  
233 ~~(Ondov et al. 2016)~~ to create signatures, [which are or compressed representations](#) of DNA  
234 sequences ~~that~~[which](#) are then used to assign taxonomic annotations. The *gather* function in this  
235 software was used for taxonomic classification at the species- and strain-level. For species-level  
236 classification, the default Genbank LCA ([Lowest Common Ancestor](#)) database (v.2018.03.29,  
237 k=31) containing 100,000 microbial genomes was used. For strain level-classification, a custom  
238 library with 245 microbial genomes ~~of~~ representative ~~of~~ tomato plant pathogens and close  
239 relatives was used. A complete list of genomes used in the custom reference library is provided  
240 in Supplementary Table 1. For all samples, signatures were computed at 31 k-mer size (for  
241 species level) and 51 k-mer size (for strain level) and abundance filtering was performed to  
242 exclude k-mers with an abundance of 1 (Brown and Irber 2016). Sourmash was run on Virginia  
243 Tech's High Performance Computing system, Advanced Research Computing (ARC), with [Intel](#)  
244 [Broadwell, 2.1GHz CPU](#), 32 cores and 128GB memory.

245

## 246 *MetaMaps*

247 *MetaMaps* (Dilthey et al. 2019) was used for taxonomic classification at the species-level using  
248 the miniSeq+H database, which includes more than 12,000 microbial genomes and is included  
249 with the software package. For strain-level classification, the custom library described above for  
250 Sourmash was used. However, the list of genomes was reduced to 149 to include only those  
251 genomes that had NCBI taxonomy IDs as per a prerequisite for *MetaMaps*. *MetaMaps* was also  
252 run on Virginia Tech's High Performance Computing system, Advanced Research Computing  
253 (ARC), [with Intel Broadwell, 2.1GHz CPU, with 32 cores and 128GB memory.](#)

## 254 *Metagenome-assembled genome analysis*

255 The reads of each metagenome were mapped [against each other to find overlaps](#) using minimap2  
256 (Li 2018) with the -x and ava-ont parameters. ~~and then a De novo~~ assembly was performed  
257 for each metagenome using the long reads assembler miniasm with default parameters (Li 2016).  
258 [Assembly correction was achieved by two iterations of racon \(v1.4.7\) with default parameters](#)  
259 [\(Vaser et al. 2017\)\(Vaser et al. 2017\).](#)

## 260 *BLAST*

261 The assemblies of each metagenome were used as input to the command-line version of BLASTN  
262 (Camacho et al. 2009) against the bacterial tomato pathogens custom database described above  
263 and with the parameter of e-value set to less than or equal to 0.01. The top hit was determined to  
264 be the alignment with the longest length for each contig.

## 265 *LINbase*

266 The ~~two longest two~~ contigs ~~in of~~ each metagenome [assembly](#) were used as input to LINbase at  
267 [linbase.org](#) (Tian et al. 2019) [to identify the pathogens at the strain level](#) with the function "Identify  
268 using a genome sequence" [to identify the pathogens at the strain level.](#)

## 270 **Results**

271 [Read-based pathogen identification after single-strain inoculation in the laboratory](#)

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

272 Tomato plants inoculated with *Pto* isolate K40 (strain T1) in the laboratory showed bacterial speck  
273 symptoms four days after inoculation (Figure 1A), at which time DNA was extracted [from a leaf](#)  
274 [wash fluid after sonication.](#)

275 The quantity and quality of the extracted DNA ~~is-are~~ listed in Table 2. An entire MinION™  
276 flow cell was used to sequence this sample (called L-K40). Of all the sequencing reads, 1,377,617  
277 reads (approximately 60% of the total number of reads) were base-called after the run was  
278 completed using the guppy software. The base-called reads had a total length of approximately  
279 4.2 Gigabases (Gbp) with the longest read measuring 66,000 bp (see more details about reads  
280 in Table 1).

281 The base-called reads were used as input to WIMP, which classified 89% of reads as of  
282 bacterial origin. [This result, which showed that our DNA extraction method starting from sonicated](#)  
283 [leaf washes was successful at minimizing host DNA contamination.](#) Of these reads, WIMP  
284 identified 77.47% as *P. syringae* genomospecies 3, a genome similarity group of which *Pto* is a  
285 member. This genome similarity group was never validly published as a named species and is  
286 thus referred to with the number 3 instead of a name (Gardan et al. 1999). Also NCBI's taxonomy  
287 database (Sayers et al. 2009) includes this taxon as *P. syringae* genomospecies 3. The next most  
288 abundant species were identified as *P. syringae* (9.39%), *P. cerasi* (2.09%), and *P. savastanoi*  
289 (1.60%). Figure 2 shows a screenshot of the WIMP result. The composition analysis is shown in  
290 Figure 3A (see Supplementary Table 2 for all relative abundance values for all composition  
291 analyses shown in Figure 3 and 4).

292 Next, the reads were used as input for composition analysis using Sourmash (Brown and  
293 Irber 2016) and MetaMaps (Dilthey et al. 2019) using the default reference libraries provided by  
294 these programs. Results are shown in Figure 3A. Sourmash identified 56.84% of the reads as *P.*  
295 *syringae* genomospecies 3 while MetaMaps identified over 91.53% of the reads as *P. syringae*  
296 genomospecies 3. Similarly to WIMP, both programs identified *P. syringae* as the next most

1  
2  
3  
4  
5  
6  
7  
8  
9  
10 297 abundant species (14.41% and 4.17%, respectively). All other species were found at a relative  
11 298 abundance of 2% or below. Therefore, WIMP, MetaMaps, and Sourmash all correctly identified  
12 299 the pathogen used in the inoculation as a member of *P. syringae* genomospecies 3.  
13  
14 300 Supplementary Table 3 reports the run times for the three tools for this sample.  
15

16 301 In an attempt to reach strain level resolution (since WIMP is limited to species-level  
17 302 identification), we built Sourmash and MetaMaps custom reference libraries consisting of genome  
18 303 sequences of representative bacterial tomato pathogen isolates and closely related isolates that  
19 304 do not cause disease on tomato. The libraries included multiple isolates of the *Pto* strains DC3000  
20 305 and T1 (Supplementary Table 2). When using these custom libraries, Sourmash identified 71.64%  
21 306 of the sequences in the sample as *Pto* isolate T1 (the isolate after which strain T1 is named) and  
22 307 the remaining sequences as other *P. syringae* isolates that are not pathogens of tomato (Table  
23 308 2). Only 0.9% of the sequences were misidentified as *Pto* DC3000. MetaMaps in combination  
24 309 with the same custom library identified 70.93% as *Pto* isolate T1, 15.90% as *Pto* isolate  
25 310 NCPPB1108 (another isolate belonging to strain T1), and 7.81% as *Pto* isolate DC3000.  
26 311 Therefore, both Sourmash and MetaMaps identified most of the reads correctly as an isolate  
27 312 belonging to *Pto* strain T1 but MetaMaps misidentified many more reads as *Pto* strain DC300  
28 313 compared to Sourmash.  
29  
30  
31  
32  
33  
34  
35  
36  
37

#### 38 315 [Read-based pathogen identification after multi-strain inoculation in the laboratory](#)

39  
40 316 Next, we wanted to test the bioinformatics pipelines established with the single-strain inoculation  
41 317 by using a mixed inoculum consisting of the *Pto* isolate K40 (strain T1) and the *Pto* isolate DC3000  
42 318 (strain DC3000) of *P. syringae* genomospecies 3 together with two additional isolates of the  
43 319 species *P. syringae* that do not cause disease on tomato: the bean pathogenic isolate *Psy* B728a  
44 320 and the non-pathogenic isolate *Psy* 642. DNA was again extracted on day four after inoculation  
45 321 and sequenced on an entire flow cell. All details for this sample (called L-mix) are listed in Table  
46 322 1. Approximately 1 million reads of a total length of 4.2 Gbp were obtained with the longest read  
47  
48  
49  
50

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

323 measuring 67,000 bp. Since this time 100% of reads were base-called, the number of base-called  
324 reads and the total length of reads were very similar to the single strain inoculation sample.

325 The caveat with this sample is that we did not know the relative abundance of the 4 isolates  
326 in the sample. However, since *Pto* isolates T1 and DC3000 are tomato pathogens while *Psy*  
327 isolates B728a and 642 are not, we expected that most sequences would be identified again as  
328 *P. syringae* genomospecies 3. In fact, WIMP identified 79.61% of all bacterial sequences (which  
329 constituted 95% of all reads) as *P. syringae* genomospecies 3 (Figure 3B), similar to the 77.47%  
330 identified in the single-strain inoculation sample. Compared to WIMP, Sourmash and MetaMaps  
331 showed the same trend as with the single strain inoculation sample: Sourmash found a lower  
332 relative abundance of *P. syringae* genomospecies 3 (43.24%) compared to WIMP and MetaMaps  
333 found a higher relative abundance compared to WIMP (91.09%) (Figure 3B).

334 Since both *Psy* isolates used in the inoculation belong to the species *P. syringae*, we  
335 expected a slightly higher relative abundance of *P. syringae* compared to the single strain  
336 inoculation sample. Interestingly, this expectation came true for Sourmash (36.87% versus  
337 14.4%) but for WIMP and MetaMaps the relative abundance of *P. syringae* only increased  
338 marginally from 9.38% to 10.01% and from 4.17% to 5.39%, respectively (Figure 3B).

339 We then used the custom reference libraries of representative tomato pathogens to see if  
340 Sourmash and MetaMaps could distinguish isolate K40 (of strain T1) from isolate DC3000 (of  
341 strain DC3000). Sourmash did identify isolate T1 of strain T1 at a relative abundance of 65.98%  
342 and isolate DC3000 of strain DC3000 at a relative abundance of 16.01% (Table 2) while  
343 MetaMaps identified 84.71% of the reads as isolates that belong to strain T1 and 5.61% as isolate  
344 DC3000 (not shown in Table 2 since only the top three hits are shown for each sample).

345 Since we did not know the correct relative abundances of strains in this inoculated plant  
346 sample and could thus not determine how accurate the results were, we decided to sequence an  
347 additional sample (called L-culture-mix) that consisted of DNA extracted from an equal mixture of  
348 the same four strains after they were grown separately overnight in liquid culture. Approximately

1  
2  
3  
4  
5  
6  
7  
8  
9  
10 349 54,000 reads of a total length of 150 Mbp were obtained on 1/6th of a flow cell with the longest  
11 350 read measuring 76,000 bp. WIMP classified 95% of the reads as bacterial. WIMP, MetaMaps,  
12 351 and Sourmash identified both, *P. syringae* and *P. syringae* genomospecies 3 in this sample, which  
13 352 we expected to be present at 50% each. WIMP over-estimated *P. syringae* compared to *P.*  
14 353 *syringae* genomospecies 3 (56% compared to 28%) and identified some other species at low  
15 354 relative abundance (Figure 3C). MetaMaps also overestimated *P. syringae* compared to *P.*  
16 355 *syringae* genomospecies 3: 65.58% vs 32.19%. Sourmash came the closest to the expected 1 to  
17 356 1 ratio finding 52.20% of *P. syringae* and 41.68% of *P. syringae* genomospecies 3 (Figure 3C).  
18 357 When using the custom reference libraries of tomato pathogens with MetaMaps and Sourmash,  
19 358 MetaMaps outperformed Sourmash since it identified DC3000 and T1 close to the expected 25%  
20 359 abundance: 38.89% and 27.48%, respectively (Table 2). Sourmash instead assigned a much  
21 360 higher abundance to strain DC3000 (75.1%) compared to strain T1 (19.63%) (Table 2).

22 361 Finally, we sequenced [the leaf wash from](#) a tomato plant grown in the lab that was not  
23 362 inoculated with any pathogen (called sample L-mock). Since the DNA concentration of this sample  
24 363 was very low, only approximately 82,000 base-called reads were obtained on 1/7th of a flow cell  
25 364 with a total length of 103 Mb. The longest read was only 19,000 bp long. Only 8% of the reads  
26 365 were classified as bacterial showing that this lab-grown plant was not colonized by many bacteria,  
27 366 which was probably also the reason for the low DNA concentration. WIMP, Sourmash, and  
28 367 MetaMaps provided very different results for this sample (Figure 3D). Importantly, as expected  
29 368 from a non-inoculated plant, none of the reads were identified by either of the three tools as *P.*  
30 369 *syringae* or *P. syringae* genomospecies 3.

31 370

#### 32 371 [Read-based pathogen identification in naturally infected tomato field samples](#)

33 372 After obtaining promising results in regard to strain-level identification with laboratory samples,  
34 373 we used DNA extracted from tomato field samples that were collected on the Eastern Shore of  
35 374 Virginia to test our pipelines with naturally infected plants (Table 1). The samples came from

1  
2  
3  
4  
5  
6  
7  
8  
9  
10 375 tomato plants that either showed symptoms of bacterial spot (samples F1-bs, F2-bs, F4-bs, F7-  
11 376 bs, F8-bs; see Figure 1B), symptoms of the fungal disease *Septoria* leaf spot (sample F5-  
12 377 *Septoria*) or no signs of any disease (F6-healthy). We also obtained one sample (F3-bs) with  
14 378 symptoms of bacterial spot. ~~However, but~~ colonies ~~that had been~~ obtained ~~by~~from culturing  
15 379 bacteria from this plant ~~during the initial diagnosis (not used for sequencing)~~ had been  
17 380 ~~identified~~found to be as a mixture of ~~colonies identified as either~~ *Pseudomonas* ~~and~~*ear*  
19 381 *Xanthomonas*.

20  
21 382 DNA from all tomato field samples were barcoded and sequenced together with other  
22 383 samples by multiplexing them on the same flow cell. Therefore, the number of reads (between  
23 384 35,923 for samples F6-healthy and 137,497 for F1-bs) and total read length (between 66  
25 385 megabases (Mb) for F6-healthy and 588 Mb for F1-bs) for these samples were much lower  
26 386 compared to the laboratory samples (Table 1).

27 387 Detailed results for all samples are reported in Figure 4. Similarly to the lab-inoculated  
29 388 samples, the majority of reads in the field samples that had symptoms of bacterial disease were  
30 389 classified as bacteria by WIMP (between 78 and 81%). Importantly, WIMP and Sourmash agreed  
31 390 that *X. perforans* was the species with the highest relative abundance in these samples (between  
32 391 25.82% and 56.44% for WIMP and between 18.51 and 66.01% for Sourmash) suggesting that *X.*  
33 392 *perforans* was the causative agent. Sample F3-bs, which had a mixed  
34 393 *Xanthomonas/Pseudomonas* infection based on culturing, was found by both WIMP and  
35 394 Sourmash to still be dominated by *X. perforans* (21.98% and 19.55% respectively) followed by  
36 395 either *P. oryzihabitans* (10.11%) and *P. fluorescens* (5.09%) based on WIMP or *P. putida*  
37 396 (16.98%) based on Sourmash. Therefore, the presence of a mixed infection was confirmed by  
38 397 both tools.

39  
40 398 In contrast to the results from WIMP and Sourmash, MetaMaps identified *X. euvesicatoria*  
41 399 and *X. alfalfae* instead of *X. perforans* as the two species with the highest relative abundance in  
42  
43  
44  
45  
46  
47  
48  
49



1  
2  
3  
4  
5  
6  
7  
8  
9  
10 400 all samples with bacterial spot symptoms. This is because *X. perforans* was missing from the  
11 401 MetaMaps reference library.

12  
13 402 Interestingly, even the non-symptomatic tomato sample (F6-healthy) was found to include  
14 403 *X. perforans* as the species with the highest relative abundance based on WIMP and Sourmash.  
15  
16 404 However, the relative abundance values were lower (6.89% and 18.54%, respectively). This  
17  
18 405 suggests that this plant might have been infected with *X. perforans* but was asymptomatic  
19 406 because of lower bacterial titer. This non-symptomatic sample also included a number of species  
20  
21 407 at relatively high abundance that were rarely found in the samples with bacterial spot symptoms,  
22 408 for example, *P. oleovorans*, *Sphingomonas parapaucimobilis*, *Microbacterium sp.* Leaf203, and  
23  
24 409 *Methylobacterium populi*.

25  
26 410 The sample with *Septoria* leaf spot symptoms (F5-Septoria), probably infected by the plant  
27 411 pathogenic fungus *Septoria lycopersici*, carried a diverse bacterial population consisting of  
28  
29 412 species in the genera *Pseudomonas*, *Xanthomonas*, *Pantoea*, *Curtobacterium*,  
30 413 *Methylobacterium*, and *Sphingomonas*. [The genome of \*Septoria lycopersici\* is not publicly](#)  
31  
32 414 [available and other species of the genus \*Septoria\* were not included in any of the reference](#)  
33  
34 415 [libraries. Identification of this fungal pathogen was thus not pursued any further.](#)

35 416 When we ~~analyzed our samples with~~ Sourmash and MetaMaps using our custom  
36  
37 417 database of representative bacterial tomato pathogens as reference libraries, *X. perforans*  
38 418 isolates TB9, TB15, and Xp9-5 were identified as the top hits in all plants with bacterial spot  
39  
40 419 symptoms with the exception of F3-bs, which had the mixed *Pseudomonas/Xanthomonas*  
41  
42 420 infection. In this sample, isolate Xp17-12 was identified by both Sourmash and MetaMaps as top  
43 421 hit. Interestingly, isolates TB9, TB15, and Xp9-5 are all members of the same intraspecific group,  
44  
45 422 *X. perforans* group 2, based on core genome phylogeny (Schwartz et al. 2015), suggesting that  
46 423 the *X. perforans* strain infecting the tomatoes with bacterial spot symptoms on the Eastern Shore  
47  
48 424 of Virginia was also a member of *X. perforans* group 2.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

425 For sample F8-bs, we also isolated *Xanthomonas* bacteria to compare the results from  
426 the culture-independent, read-based metagenomic approach with a culture-dependent genomic  
427 approach. DNA was extracted from two colonies and sequenced using Illumina HiSeq. The two  
428 genome sequences were assembled into 87 and 86 contigs, respectively, with a total length of  
429 5,340,265 bp and 5,339,287 bp. We used the LINbase Web service for genome-based microbial  
430 identification and found isolate GEV1063 to be the best match for both genomes with 99.98% ANI  
431 and both genomes were identified by LINbase as members of *X. perforans* group 2, which is  
432 circumscribed in LINbase as an intraspecific taxon. Therefore, the culture-dependent genome-  
433 based identification confirmed the culture-independent read-based strain-level identification of *X.*  
434 *perforans* group 2 as the causative agent in sample F8-bs.

#### 436 Metagenome assembly-based pathogen identification

437 In parallel to the read-based pipelines described above, we also assembled each metagenomic  
438 sample using all reads that had a minimum length of 1,000 bp [followed by two iterations of the](#)  
439 [error correcting tool racon \(Vaser et al. 2017\)\(add reference here\) and that were identified by](#)  
440 [WIMP as bacterial](#). The results are summarized in Table 3. The non-inoculated tomato sample  
441 from the lab (L-mock), the healthy tomato sample from the field (F6-healthy), and the sample of  
442 the tomato plant with Septoria leaf spot (F5-Septoria) had the lowest number of contigs (between  
443 4 and 9) with the shortest total length of contigs (between 21,390 bp and 122,956 bp). This was  
444 probably a result of the low number of bacterial reads in these samples (Table 1).

445 The samples with symptoms of either bacterial speck or bacterial spot had a wide range  
446 in [regard to contig number \(10 to 131 contigs\)](#) and [in the total length of contigs \(ranging from 10](#)  
447 [to 131 contigs of a total length from 5.2 to 12.85 Mbp\)](#). For our goal of identifying the causative  
448 agent in each symptomatic plant to strain level, we focused on the [two](#) longest contigs in each  
449 sample since these contigs were the most likely to be of the causative pathogenic agents. It was  
450 very promising to see that in some of the symptomatic samples the longest contig was of a size

1  
2  
3  
4  
5  
6  
7  
8  
9  
10 451 similar to an entire bacterial genome, for example, 6.08Mbp in the tomato lab sample inoculated  
11 452 with *Pto* isolate K40 (L-K40), and 5.03Mbp for the field sample F7-bs showing bacterial spot  
12 453 symptoms (Table 3). We then used the genome alignment tool MUMmer (Marçais et al. 2018) to  
13 454 determine how much of the published genome sequences these contigs covered. We found that  
14 455 in the case of sample L-K40, the longest contig aligned with ~~97.90%~~93-92% of the published  
15 456 genome sequence of isolate K40. For F7-b's, the longest contig aligned with ~~95.81%~~95-52% of  
16 457 the published *X. perforans* genome [TB15 \(the genome identified by Sourmash with the highest](#)  
17 458 [abundance in this sample\).](#)

18  
19  
20  
21  
22 459 To obtain a preliminary identification of all contigs we used BLASTN (Camacho et al. 2009)  
23 460 in combination with our custom tomato pathogen database. The results were mostly in agreement  
24 461 with the reads-based analysis at the species level (Figure 5) but *X. euvesicatoria* was identified  
25 462 as species instead of *X. perforans* in some of the samples with bacterial spot.

26  
27  
28  
29 463 To attempt identification of the longest contigs to strain level, we used these contigs as  
30 464 queries with the "Identify using a genome sequence" function in the LINbase Web service (Tian  
31 465 et al. 2019). Table 4 lists the results that were obtained for the longest two contigs (separately  
32 466 and merged) for each sample. When using the longest contig of the tomato plant inoculated with  
33 467 isolate K40 (of *Pto* strain T1 ([sample L-K40](#))), the *Pto* strain T1 isolate BAV1020 was [identified](#)  
34 468 [as the best hit but only](#) with an ANI of ~~99~~92.76% compared to the query sequence. [This very high](#)  
35 469 [ANI value shows that the error-correcting tool racon \(Vaser et al. 2017](#)~~add racon reference~~[\) was](#)  
36 470 [successful in correcting most sequencing errors in the assembly. K40 was expected to be the](#)  
37 471 [best hit for this sample since this is the isolate that was used in the inoculation. Wand-w](#)~~We do~~  
38 472 [not know why isolate BAV1020 was identified as best hit](#)~~instead of isolate K40. However, isolates~~  
39 473 [BAV1020 and K40 have a reciprocal ANI of over 99.75%,](#) ~~were both were~~ isolated from tomato  
40 474 [plants in Virginia, and both belong to \*Pto\* strain T1, making it irrelevant for strain-level identification](#)  
41 475 [which isolate was the best hit](#)~~However, based on a direct genome sequence comparison, the two~~  
42 476 [genomes are also over 99.75% identical to each other. Why the contig](#)~~Since we know that isolate~~

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

477 K40 was used as inoculum, the discrepancy between the two ANI value is necessarily a result of  
478 the high error rate of the MinION™ sequencer. Most importantly, since genome-sequenced  
479 isolates of *Pto* strain T1 have pair-wise ANI values of 99.75% or higher and the ANI between the  
480 longest contig of L-K40 and its best hit BAV1020 had an ANI of over 99.76%, we were able to  
481 identify L-K40 as member of *Pto* strain T1.

482 For the tomato plant inoculated with the four-strain mix (sample L-mix), the longest contig  
483 was again identified as *Pto* strain T1 based on the same best hit to *Pto* isolate BAV1020T1 with  
484 an ANI value of 99.773%. Interestingly, using the two longest contigs together in a single query,  
485 isolate K40 was identified as the best hit. No contig of significant length was identified as either  
486 *Pto* isolate DC3000 or the other two *Psy* isolates used in the inoculation. This may have been due  
487 to poor growth of these isolates in tomato compared to isolate K40 (as suggested by the read-  
488 based analysis above). Moreover, since the genomes of *Pto* isolates DC3000 and T1-K40 are  
489 over 98.5% identical to each other, some DC3000 reads may have been the longest contigs of  
490 this sample may have been probably assembled together with K40 reads into the same  
491 contigs. The other two isolates used for the inoculation from a combination of DC3000 and T1  
492 reads, which could not be distinguished from each other also because of the high error rate of the  
493 MinION™ sequencer.

494 For the longest contigs in the tomato field samples that showed bacterial spot symptoms,  
495 different isolates of *X. perforans* were the best hits: Xp8-16, Xp10-13, GEV1063, and GEV2116,  
496 and TB6 (Table 4). Isolates GEV1063 and TB6. These isolates both belong to *X. perforans* group  
497 2 (Schwartz et al. 2015) and this results is are thus in line with the read-based results described  
498 above. Only the second-longest contig in sample F42-bs and the two longest contigs in sample  
499 F4-bs contradicted the read-based results: *X. perforans* isolate 91-118, a member of *X. perforans*  
500 group 1B (Schwartz et al. 2015), was the best hit for this/these contigs.

501 Since for sample F8-bs we also had the genome sequences of the two cultured isolates  
502 (sequenced with Illumina and assembled with using Unicycler; see previous section), we could

1  
2  
3  
4  
5  
6  
7  
8  
9  
10 503 again directly compare the metagenomic assembly-based approach with the culture-dependent  
11 504 genomic approach. ~~Although there was no difference in the identification results themselves~~  
12 505 ~~since the~~ The best matches in LINbase for both approaches ~~was were~~ the isolate  
13 506 GEV1063 isolates of *X. perforans* group 2. ~~T~~, the ANI value of 99.76% between the longest contig  
14 507 of F8-bs and isolate GEV1063 ~~the most similar genome in LINbase was almost as high as the~~  
15 508 ANI value between the Illumina-sequenced isolates cultured from F8-bs and isolate GEV1063,  
16 509 which was was only 99.7635% while the ANI between the genome sequences of the isolated  
17 510 colonies and their most similar genome in LINbase was 99.98%. ~~As with the lab-inoculated~~  
18 511 ~~sample L-K40, this difference in ANI was probably again due to the high error rate of the MinION™~~  
19 512 ~~and was the reason we could not directly identify the causative agent as a member of *X. perforans*~~  
20 513 ~~group 2. However, since genome-sequenced isolates of *X. perforans* group 2 have pair-wise ANI~~  
21 514 values of over 99.9% and the ANI between the longest contig of F8-bs and its best hit, isolate  
22 515 GEV1063, was 99.76%, we could not identify the causative agent in sample F8-bs with high  
23 516 confidence as member of *X. perforans* group 2.  
24  
25  
26  
27  
28  
29  
30  
31  
32 517

## 518 Discussion

33  
34  
35 519 Sensitive detection and precise identification of pathogens in real time directly from symptomatic  
36 520 organisms, or even better from infected but still asymptomatic organisms, without the need for  
37 521 pathogen isolation and culturing, is the ultimate goal in control and prevention of infectious  
38 522 diseases of humans, animals, and plants.

39  
40  
41 523 As a step towards this goal in plant pathology, here we used the ONT MinION™ for precise  
42 524 identification of two bacterial tomato pathogens by sequencing metagenomic DNA directly  
43 525 extracted from symptomatic plants and analyzing the obtained sequences with a set of different  
44 526 tools and databases. However, we neither attempted to maximize sensitivity of detection nor to  
45 527 minimize the time necessary for identification.  
46  
47  
48  
49  
50

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

528 Several other reports describing the use of the MinION™ in culture-independent  
529 metagenomic DNA sequencing for plant pathogen identification have recently been published.  
530 Most of these reports either focused on species-level identification (Hu et al. 2019) and/or on  
531 accelerating the identification protocol (Loit et al. 2019). Only one report focused on strain-level  
532 identification but after polymerase chain reaction with primers specific to loci of a single pathogen  
533 species, which increased the sensitivity of detection and resolution of identification but restricts  
534 the approach to a single pathogen species at the time (Radhakrishnan et al. 2019). Our goal  
535 instead was to develop an experimental and bioinformatics pipeline that can be used for any  
536 bacterial plant pathogen, and, with modifications, possibly for fungal and oomycete pathogens as  
537 well.

538 The first critical step in metagenomic-based pathogen identification is DNA extraction.  
539 There are mainly two possibilities: extracting DNA directly from plant tissue or extracting DNA  
540 from water used to wash the plant (after sonication to help dislocate the pathogen from the tissue).  
541 The first approach has the advantage that large quantities of high-quality DNA can be extracted.  
542 The obvious disadvantage is that a large fraction of the extracted DNA is plant DNA. The second  
543 approach is the approach we decided to use since it is widely used for plant microbiome analysis,  
544 for example (Ottesen et al. 2013). Based on the results from our DNA sequence analysis, this  
545 approach allowed us to obtain DNA that was over 80% of bacterial origin for the naturally infected  
546 tomato field samples and over 90% of bacterial origin for the artificially inoculated tomato plants  
547 grown in the laboratory. This value was as high as the fraction of bacterial DNA when extracting  
548 DNA directly from a bacterial culture. Therefore, we conclude that for metagenome-based  
549 identification of bacterial foliar pathogens in symptomatic plant tissue extracting DNA from wash  
550 water after sonication is an excellent solution. Importantly, even the wash water of our healthy  
551 field sample still contained 30% of bacterial DNA, making this approach possibly still a good  
552 choice even for asymptomatic leaves with relatively low bacterial titers.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10 553 Because in this project we were not interested in speed, we used the slower, higher  
11 554 yielding DNA sequencing library preparation protocol, as suggested by ONT, without significant  
12  
13 555 modifications. Also for the sequencing protocol itself, we followed ONT's instructions without  
14 556 modifications. The first critical step after sequencing the DNA, is base-calling, which is the process  
15  
16 557 of translating the raw electrical signals measured by the MinION™ into nucleotide sequences.  
17  
18 558 Since base-calling is computationally intensive and takes longer than sequencing itself, base-  
19 559 calling needed to be completed after the sequencing runs themselves were completed. We used  
20  
21 560 the ONT Guppy base-calling tool without any polishing.

22 561 The actual assignment of sequencing reads to specific bacterial species and strains was  
23  
24 562 done using a total of five tools: 1. ONT's WIMP software with graphical user interface, which is  
25  
26 563 intuitive to use and uses the software Centrifuge (Kim et al. 2016) to rapidly identify and assign  
27 564 taxonomy to the reads coming from the sequencing base calling in real-time, 2. the command-  
28  
29 565 line tool Sourmash (Brown and Irber 2016) that computes hash sketches from DNA sequences  
30 566 and includes k-mer based taxonomic classification for genomic and metagenomic analysis, 3. the  
31  
32 567 command line tool MetaMaps (Dilthey et al. 2019) which uses approximate mapping algorithm to  
33  
34 568 map long-read metagenomic sequences to comprehensive databases, 4. the command line  
35 569 version of BLASTN (Camacho et al. 2009) was used to speed up the identification of pathogens  
36  
37 570 with a custom built database after metagenome assembly—identification of pathogens after  
38  
39 571 metagenome assembly with a custom built database, 5. after metagenome assembly performed  
40 572 with minimap 2 and miniasm (Li 2016), the two longest contigs of each metagenome assembly  
41  
42 573 assembly of metagenomes were used for obtained withby minimap2 and miniasm (Li 2016) were  
43 574 followed by taxonomy assignment with LINbase (Tian et al. 2019) of the two longest contigs  
44  
45 575 obtained by LINbase (Tian et al. 2019). Moreover, Sourmash and MetaMaps were used both with  
46  
47 576 default and custom libraries.

48 577 For species-level identification, the three read-based tools performed similarly well with  
49  
50 578 the lab samples in regard to accuracy with Sourmash coming the closest to the expected 1 : 1

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

579 ratio of *P. syringae* genomospecies 3 : *P. syringae* in the sample L-culture-mix. For the field  
580 samples, the absence of *X. perforans* in the MetaMaps default reference library did not allow  
581 MetaMaps to identify *X. perforans* while WIMP and Sourmash performed similarly well. Both  
582 identified *X. perforans* as the most abundant species in all samples with bacterial spot symptoms.

583 As for run time, only WIMP is set up to provide real-time results starting minutes after runs  
584 are initiated and results are updated as more sequencing reads are base-called. However, since  
585 base-calling cannot keep up with the amount of raw data that is being generated during a run,  
586 WIMP needs to be re-run when base-calling is completed after a run ends in order to analyze all  
587 data. This took over 36 hours for our largest sample, L-K40 (Supplementary Table 3). The  
588 advantage is that users do not need any significant local computing resources to do this since  
589 WIMP runs on ONT's cloud. For the same L-K40 sample, it took Sourmash only 35 minutes to  
590 calculate the k-mer signature and perform species-level classification while MetaMaps  
591 completed the same run in 6-8 hours. Both tools were run on Virginia Tech's ARC high-  
592 performance computing system. Therefore, Sourmash is significantly faster than MetaMaps and  
593 WIMP but still requires significant computing resources.

594 In regard to ease of use, WIMP ~~stands out cannot be beaten~~ because of its intuitive  
595 graphical user interface. Although both Sourmash and MetaMaps are command-line tools,  
596 Sourmash beats MetaMaps because of the extensive tutorials provided on the Sourmash  
597 website. The added ease of making custom reference libraries and adding genomes to existing  
598 libraries also makes Sourmash more user-friendly compared to MetaMaps, which requires NCBI  
599 taxIDs (or creation of custom taxIDs) for all genomes in custom reference libraries.

600 Assembling reads into contigs before identification did not provide any advantages for  
601 species-level identification since species-level identification was successful with read-based tools  
602 and read-based identification is generally faster since it does not require prior assembly of reads  
603 into contigs. However, this advantage of speed may diminish with an increasing number of reads



1  
2  
3  
4  
5  
6  
7  
8  
9  
10 604 since mapping of a smaller number of assembled contigs might be faster than mapping a large  
11 605 number of reads individually.

12  
13 606 For strain-level identification, WIMP cannot be used since it only reaches species-level  
14 607 resolution. When comparing MetaMaps with Sourmash, MetaMaps misidentified a larger number  
15  
16 608 of reads as strain *Pto* DC3000 compared to Sourmash in the single strain inoculation sample L-  
17 609 K40, which we knew did not contain any DNA of strain *Pto* DC3000. Instead in the sample L-  
18  
19 610 culture-mix with known equal concentrations, it was Sourmash that overestimated strain *Pto*  
20  
21 611 DC3000 compared to strain *Pto* T1. For field sample F8-bs for which we had also a culture-  
22 612 dependent result indicating *X. perforans* group 2 as causative agent, both software identified the  
23  
24 613 same best hit in the custom database that was also a member of *X. perforans* group 2. Therefore,  
25  
26 614 we conclude that Sourmash and MetaMaps did equally well in regard to strain accuracy. In regard  
27 615 to run time, Sourmash's run time increased to 1-3 hours when using a k-mer size of 51, which is  
28  
29 616 required for strain-level identification. Run time for MetaMaps decreased to 3-4 hours because of  
30  
31 617 the smaller size of the custom library in comparison to default databases. However, Sourmash  
32 618 still performed better than MetaMaps in regard to computing time.

33  
34 619 The challenge when using either Sourmash or MetaMaps for strain-level identification is  
35 620 that we had to interpret the results based on prior knowledge of which isolates in our custom  
36  
37 621 database belonged to which pathogen strain. For example, only by checking Figure 1 in (Schwartz  
38 622 et al. 2015), were we able to identify the best matches found by Sourmash and MetaMaps in our  
39  
40 623 custom database as members of *X. perforans* group 2. Moreover, a best match with an isolate  
41 624 that belongs to a certain strain, or any other group or taxon for that matter, still does not  
42  
43 625 necessarily mean that the query is a member of the same group as well. To make such a  
44  
45 626 conclusion, it is necessary to determine (1) the genomic breadth of the group, for example,  
46 627 99.75% for *X. perforans* group 2, and (2) the genomic distance of the query to a representative  
47  
48 628 member of that group with this distance needing to be smaller than the genomic breadth of the  
49  
50 629 group. Alternatively, a phylogenetic analysis could be performed to determine if the unknown is a

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

630 member of the clade that corresponds to the specific group. Because species have a standard  
631 genomic breadth of 95% ANI, WIMP, Sourmash, and MetaMaps can infer species membership  
632 from metagenomic reads relatively easily. However, strains (and any other group smaller than a  
633 species) do not have a standard ANI breadth. Therefore, Sourmash and MetaMaps would need  
634 to be given genomic circumscriptions of strains as part of the reference library information in order  
635 to precisely assign reads to strains.

636 Since the MinION™ outputs long reads, we were ~~surprisingly~~ successful in assembling  
637 reads into contigs almost as long as entire bacterial genomes, which could then be used for  
638 genome-based identification. We specifically developed the LINbase Web service for identifying  
639 microbes as members of taxa at any genomic breadth below the rank of genus (Tian et al. 2019)  
640 and we had circumscribed both *Pto* strain T1 and *X. perforans* group 2 as taxa in LINbase with  
641 genomic breadths of 99.75% and 99.9% ANI, respectively. ~~Therefore, we were able to identify~~  
642 ~~the we should have been able to avoid the problem that we had with read-based identification.~~  
643 ~~However, the challenge that arose with this approach was that because of the high error rate of~~  
644 ~~the MinION™, the ANI between all query contigs and their best matches in LINbase were below~~  
645 ~~95%. This was true even for the longest contig in sample L-K40, which had been inoculated with~~  
646 ~~strain *Pto* T1 isolate K40. Therefore, the longest contig in this sample should have had an almost~~  
647 ~~100% match in LINbase with the genome of isolate K40 and other isolates that belong to strain~~  
648 ~~T1. However, the ANI between this contig and the best match in LINbase was only 92.76%. These~~  
649 ~~ANI thresholds were chosen because genome-sequenced isolates of *Pto* strain T1 deposited in~~  
650 ~~LINbase have pair-wise ANI values of over 99.75% and genome-sequenced isolates of *X.*~~  
651 ~~*perforans* group 2 deposited in LINbase have pairwise ANI values of over 99.9%. Since ANI~~  
652 ~~between the longest contig of sample L-K40 and the isolate T1 of *Pto* strain T1 was 99.76% (and~~  
653 ~~thus above the 99.75% ANI threshold at which *Pto* strain T1 is circumscribed in LINbase), we~~  
654 ~~were able to correctly identify the causative agent in sample L-K40 as a member of *Pto* strain T1.~~  
655 ~~For the field samples, this was not possible since ANI between the longest contigs and the most~~

1  
2  
3  
4  
5  
6  
7  
8  
9  
10 656 similar isolates of *X. perforans* group 2 in LINbase was between 99.62% and 99.84% (and thus  
11 657 below the 99.9% ANI threshold at which *X. perforans* group 2 is circumscribed in LINbase). We  
12 658 expect that ~~Therefore,~~ modest reduction in the current error rate of the MinION or a small  
13 659 improvement in the error correction step would probably allow strain-level identification even in  
14 660 this case using the metagenome-assembled contigs did not allow us to identify the pathogens as  
15 661 members of the strains circumscribed in LINbase because the MinION™ error rate lowered the  
16 662 ANI between the query contig and the best match to below the genomic breadth of the  
17 663 circumscribed taxon. Being aware of the high error rate, we were still able to extrapolate from the  
18 664 best match in LINbase the identity of the correct strain. However, such a result can only be  
19 665 considered putative or preliminary.

20  
21  
22  
23  
24  
25  
26 666 In conclusion, using either the Sourmash and MetaMaps tools for metagenomic read-  
27 667 based strain identification or LINbase for assembly-based strain-level identification, putative  
28 668 strain-level identification was possible and was confirmed by culture-dependent genome-based  
29 669 identification. However, it was impossible-not yet possible to reach the same high-confidence  
30 670 strain-level identification of culture-dependent genome-based identification because of the  
31 671 absence of appropriate strain-level databases for the read-based tools and because of the  
32 672 currently still high error rate of the MinION™ (version 19.05.0) when using assembly-based  
33 673 identification. ~~Therefore, Considering the large and active user community of the MinION™~~  
34 674 ~~sequencer and the continued development of new versions of the MinION™, we expect~~  
35 675 ~~improvements in both, tool development for read-based identification, and improvements in the~~  
36 676 ~~precision at which the MinION™ can distinguish nucleotides from each other and/or base-calling~~  
37 677 ~~algorithms, which should ultimately lower the currently high error rate.~~ At this point, we consider  
38 678 culture-independent metagenomic sequencing with the MinION™ an excellent approach to obtain  
39 679 results when high confidence strain-level identification is not required or when a culture-  
40 680 dependent genome-based identification is used as a follow-up. ~~However, considering the large~~  
41 681 ~~and active user community of the MinION™ sequencer and the continued development of new~~

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

682 [versions of the MinION™, we expect improvements in the precision at which the MinION™ can](#)  
683 [distinguish nucleotides from each other, in base-calling algorithms, in error correction, and in tool](#)  
684 [development for read-based identification. Together, these improvements can be expected to](#)  
685 [take us to high-confidence strain-level identification of bacterial plant pathogens from](#)  
686 [metagenomic sequences in the near future.](#)

#### 688 **Author contributions**

689 BAV and SL developed the project. MEML performed most of the wet-lab experiments. MAF and  
690 PS did most of the bioinformatics analyses. SY contributed to the wet-lab experiments. LT and  
691 CH, under supervision from BAV and LSH, developed LINbase. BAV, with contributions from  
692 MEML, MAF, PS, and SL wrote the manuscript. All authors read and approved the final version  
693 of the manuscript.

#### 695 **Conflict of Interest**

696 LINbase uses the trademarks Life Identification Number® and LIN®, which are registered by This  
697 Genomic Life, Inc. LSH and BAV report in accordance with Virginia Tech policies and procedures  
698 and their ethical obligation as researchers that they have a financial interest in This Genomic Life,  
699 Inc. Therefore, their financial interests may be affected by the research reported in this  
700 manuscript. They have disclosed those interests fully to Virginia Tech, and they have in place an  
701 approved plan for managing any potential conflicts arising from this relationship.

#### 703 **Funding**

704 This study was supported by the College of Agriculture and Life Sciences at Virginia Polytechnic  
705 Institute and State University and by the National Science Foundation (IOS-1754721). Funding to  
706 BAV and SL was also provided in part by the Virginia Agricultural Experiment Station and the  
707 Hatch Program of the National Institute of Food and Agriculture, US Department of Agriculture.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10 708  
11 709 **Acknowledgements**  
12  
13 710 The authors acknowledge Advanced Research Computing (ARC) at Virginia Tech for providing  
14 711 computational resources and technical support that have contributed to the results reported within  
15 712 this paper. URL: <http://www.arc.vt.edu>  
16 713

#### 19 714 **Literature cited**

20  
21 715 Almeida, N. F., Yan, S., Cai, R., Clarke, C. R., Morris, C. E., Schaad, N. W., et al. 2010.  
22 716 PAMDB, a multilocus sequence typing and analysis database and website for plant-  
23 717 associated microbes. *Phytopathology*. 100:208–215.  
24 718  
25 718 Andrews, S. 2010. Babraham bioinformatics-FastQC a quality control tool for high throughput  
26 719 sequence data. URL: [https://www.bioinformatics.babraham.ac.](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)  
27 719 [uk/projects/fastqc/](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)(accessed 06. 12. 2018). Available at:  
28 720  
29 720 <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.  
30 721  
31 722 Badial, A. B., Sherman, D., Stone, A., Gopakumar, A., Wilson, V., Schneider, W., et al. 2018.  
32 723 Nanopore Sequencing as a Surveillance Tool for Plant Pathogens in Plant and Insect  
33 724 Tissues. *Plant Disease*. 102:1648–1652 Available at: [http://dx.doi.org/10.1094/pdis-04-17-](http://dx.doi.org/10.1094/pdis-04-17-0488-re)  
34 724  
35 724 0488-re.  
36 725  
37 725  
38 726 Brown, C. T., and Irber, L. 2016. sourmash: a library for MinHash sketching of DNA. *J. Open*  
39 727 *Source Software*. 1:27.  
40 727  
41 728 Bushnell, B. 2015. BBMap. Available at: <https://sourceforge.net/projects/bbmap/>.  
42 728  
43 729 Cai, R., Lewis, J., Yan, S., Liu, H., Clarke, C. R., Campanile, F., et al. 2011. The plant pathogen  
44 730 *Pseudomonas syringae* pv. *tomato* is genetically monomorphic and under strong selection  
45 730  
46 731 to evade tomato immunity. *PLoS Pathog*. 7:e1002130.  
47 731  
48 732 Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. 2009.  
49 733 BLAST+: architecture and applications. *BMC Bioinformatics*. 10:421.  
50 733

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 734 Chalupowicz, L., Dombrovsky, A., Gaba, V., Luria, N., Reuven, M., Beerman, A., et al. 2019.  
735 Diagnosis of plant diseases using the Nanopore sequencing platform. *Plant Pathol.* 68:229–  
736 238.
- 737 Clarke, C. R., Cai, R., Studholme, D. J., Guttman, D. S., and Vinatzer, B. A. 2010.  
738 *Pseudomonas syringae* strains naturally lacking the classical *P. syringae* *hrp/hrc* Locus are  
739 common leaf colonizers equipped with an atypical type III secretion system. *Mol. Plant.*  
740 *Microbe. Interact.* 23:198–210.
- 741 Dijkshoorn, L., Ursing, B. M., and Ursing, J. B. 2000. Strain, clone and species: comments on  
742 three basic concepts of bacteriology. *J. Med. Microbiol.* 49:397–401.
- 743 Dilthey, A. T., Jain, C., Koren, S., and Phillippy, A. M. 2019. Strain-level metagenomic  
744 assignment and compositional estimation for long reads with MetaMaps. *Nat. Commun.*  
745 10:3066.
- 746 Fang, Y., and Ramasamy, R. P. 2015. Current and Prospective Methods for Plant Disease  
747 Detection. *Biosensors.* 5:537–561.
- 748 Feil, H., Feil, W. S., Chain, P., Larimer, F., DiBartolo, G., Copeland, A., et al. 2005. Comparison  
749 of the complete genome sequences of *Pseudomonas syringae* pv. *syringae* B728a and pv.  
750 *tomato* DC3000. *Proc. Natl. Acad. Sci. U. S. A.* 102:11064–11069.
- 751 Gardan, L., Shafik, H., Belouin, S., Broch, R., Grimont, F., and Grimont, P. A. 1999. DNA  
752 relatedness among the pathovars of *Pseudomonas syringae* and description of  
753 *Pseudomonas tremae* sp. nov. and *Pseudomonas cannabina* sp. nov. (ex Sutic and Dowson  
754 1959). *Int. J. Syst. Bacteriol.* 49 Pt 2:469–478.
- 755 Hu, Y., Green, G. S., Milgate, A. W., Stone, E. A., Rathjen, J. P., and Schwessinger, B. 2019.  
756 Pathogen Detection and Microbiome Analysis of Infected Wheat Using a Portable DNA  
757 Sequencer. *Phytobiomes Journal.* 3:92–101.
- 758 Jain, M., Olsen, H. E., Paten, B., and Akeson, M. 2016. The Oxford Nanopore MinION: delivery  
759 of nanopore sequencing to the genomics community. *Genome Biol.* 17:239.

- 1  
2  
3  
4  
5  
6  
7  
8  
9  
10 760 Jones, J. B., Lacy, G. H., Bouzar, H., Stall, R. E., and Schaad, N. W. 2004. Reclassification of  
11 761 the Xanthomonads associated with bacterial spot disease of tomato and pepper. *Syst. Appl.*  
12 762 *Microbiol.* 27:755–762.
- 14 763 Juul, S., Izquierdo, F., Hurst, A., Dai, X., Wright, A., Kulesha, E., et al. 2015. What's in my pot?  
15 764 Real-time species identification on the MinION. *bioRxiv.* :030742.
- 17 765 Kim, D., Song, L., Breitwieser, F. P., and Salzberg, S. L. 2016. Centrifuge: rapid and sensitive  
18 766 classification of metagenomic sequences. *Genome Res.* 26:1721–1729.
- 20 767 Konstantinidis, K. T., and Tiedje, J. M. 2005. Genomic insights that advance the species  
21 768 definition for prokaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 102:2567–2572.
- 24 769 Li, H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 34:3094–  
25 770 3100.
- 27 771 Li, H. 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long  
28 772 sequences. *Bioinformatics.* 32:2103–2110.
- 30 773 Loit, K., Adamson, K., Bahram, M., Puusepp, R., Anslan, S., Kiiker, R., et al. 2019. Relative  
31 774 performance of Oxford Nanopore MinION vs. Pacific Biosciences Sequel third-generation  
32 775 sequencing platforms in identification of agricultural and forest pathogens. *bioRxiv.* :592972  
33 776 Available at: <https://www.biorxiv.org/content/10.1101/592972v1.abstract> [Accessed  
34 777 September 8, 2019].
- 38 778 Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., and Zimin, A. 2018.  
39 779 MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.*  
40 780 14:e1005944.
- 43 781 Mechan-Llontop, M. E., Tian, L., Bernal-Galeano, V., Reeves, E., Hansen, M. A., Bush, E., et al.  
44 782 2019. Assessing the potential of culture-independent 16S rRNA microbiome analysis in  
45 783 disease diagnostics: the example of *Dianthus gratianopolitanus* and *Robbsia andropogonis*.  
46 784 *European Journal of Plant Pathology.* Available at: [http://dx.doi.org/10.1007/s10658-019-](http://dx.doi.org/10.1007/s10658-019-01850-8)  
47 785 01850-8 [Accessed September 16, 2019].

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 786 MinION brochure. 2019a. Oxford Nanopore Technologies. Available at:  
787 <http://nanoporetech.com/resource-centre/minion-brochure> [Accessed September 14, 2019].
- 788 MinION brochure. 2019b. Oxford Nanopore Technologies. Available at:  
789 <http://nanoporetech.com/resource-centre/minion-brochure> [Accessed September 14, 2019].
- 790 Nadon, C., Van Walle, I., Gerner-Smidt, P., Campos, J., Chinen, I., Concepcion-Acevedo, J., et  
791 al. 2017. PulseNet International: Vision for the implementation of whole genome sequencing  
792 (WGS) for global food-borne disease surveillance. *Euro Surveill.* 22 Available at:  
793 <http://dx.doi.org/10.2807/1560-7917.ES.2017.22.23.30544>.
- 794 [Ondov, B.D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S. and](#)  
795 [Phillippy, A. M., 2016. Mash: fast genome and metagenome distance estimation using](#)  
796 [MinHash. \*Genome biology.\* 17\(1\):132. Available at doi:10.1186/s13059-016-0997-x.](#)
- 797
- 798 Ottesen, A. R., González Peña, A., White, J. R., Pettengill, J. B., Li, C., Allard, S., et al. 2013.  
799 Baseline survey of the anatomical microbial ecology of an important food plant: *Solanum*  
800 *lycopersicum* (tomato). *BMC Microbiol.* 13:114.
- 801 [protocols.io. High molecular weight DNA extraction from all kingdoms. Available at:](#)  
802 <https://www.protocols.io/groups/high-molecular-weight-dna-extraction-from-all-kingdoms>  
803 [\[Accessed November 13, 2019\].](#)
- 804 Radhakrishnan, G. V., Cook, N. M., Bueno-Sancho, V., Lewis, C. M., Persoons, A., Mitiku, A.  
805 D., et al. 2019. MARPLE, a point-of-care, strain-level disease diagnostics and surveillance  
806 tool for complex fungal pathogens. *BMC Biology.* 17 Available at:  
807 <http://dx.doi.org/10.1186/s12915-019-0684-y>.
- 808 Rees-George, J., Vanneste, J. L., Cornish, D. A., Pushparajah, I. P. S., Yu, J., Templeton, M.  
809 D., et al. 2010. Detection of *Pseudomonas syringae* pv. *actinidiae* using polymerase chain  
810 reaction (PCR) primers based on the 16S-23S rDNA intertranscribed spacer region and



- 1  
2  
3  
4  
5  
6  
7  
8  
9  
10 811 comparison with PCR primers based on other gene regions. *Plant Pathology*. 59:453–464  
11 812 Available at: <http://dx.doi.org/10.1111/j.1365-3059.2010.02259.x>.
- 12  
13 813 Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., et al. 2009.  
14 814 Database resources of the National Center for Biotechnology Information. *Nucleic Acids*  
15 815 *Res.* 37:D5–15.
- 16  
17 816 Schwartz, A. R., Potnis, N., Timilsina, S., Wilson, M., PatanĀ©, J., Martins, J., et al. 2015.  
18 817 Phylogenomics of *Xanthomonas* field strains infecting pepper and tomato reveals diversity in  
19 818 effector repertoires and identifies determinants of host specificity. *Frontiers in Microbiology*.  
20 819 6 Available at: <http://dx.doi.org/10.3389/fmicb.2015.00535>.
- 21  
22 820 Tedersoo, L., Drenkhan, R., Anslan, S., Morales-Rodriguez, C., and Cleary, M. 2019.  
23 821 High-throughput identification and diagnostics of pathogens and pests: Overview and  
24 822 practical recommendations. *Molecular Ecology Resources*. 19:47–76 Available at:  
25 823 <http://dx.doi.org/10.1111/1755-0998.12959>.
- 26  
27 824 Tian, L., Huang, C., Heath, L. S., and Vinatzer, B. A. 2019. LINbase: A Web service for  
28 825 genome-based identification of microbes as members of crowdsourced taxa. *bioRxiv*.  
29 826 Available at: <https://www.biorxiv.org/content/10.1101/752212v1.abstract>.
- 30  
31 827 Tinivella, F., Gullino, M. L., and Stack, J. P. 2008. The Need for Diagnostic Tools and  
32 828 Infrastructure. In *Crop Biosecurity*, Springer Netherlands, p. 63–71.
- 33  
34 829 [Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. 2017. Fast and accurate de novo genome](#)  
35 830 [assembly from long uncorrected reads. \*Genome Research\*. 27\(5\), 737–746. Available at](#)  
36 831 [doi: 10.1101/gr.214270.116](https://doi.org/10.1101/gr.214270.116)
- 37  
38 832 Wick, R. R., Judd, L. M., Gorrie, C. L., and Holt, K. E. 2017. Unicycler: Resolving bacterial  
39 833 genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.*  
40 834 13:e1005595.
- 41  
42 835 Williamson, L., Nakaho, K., Hudelson, B., and Allen, C. 2002. *Ralstonia solanacearum* Race 3,  
43 836 Biovar 2 Strains Isolated from Geranium Are Pathogenic on Potato. *Plant Dis*. 86:987–991.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

837 Yan, S., Liu, H., Mohr, T. J., Jenrette, J., Chiodini, R., Zaccardelli, M., et al. 2008. Role of  
838 Recombination in the Evolution of the Model Plant Pathogen *Pseudomonas syringae* pv.  
839 *tomato* DC3000, a Very Atypical Tomato Strain. *Applied and Environmental Microbiology*.  
840 74:3171–3181 Available at: <http://dx.doi.org/10.1128/aem.00180-08>.

841 ~~[Ondov, B.D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S. and](#)~~  
842 ~~[Phillippy, A. M., 2016. Mash: fast genome and metagenome distance estimation using](#)~~  
843 ~~[MinHash. \*Genome biology\*, 17\(1\):132. Available at doi:10.1186/s13059-016-0997-x.](#)~~

## Tables

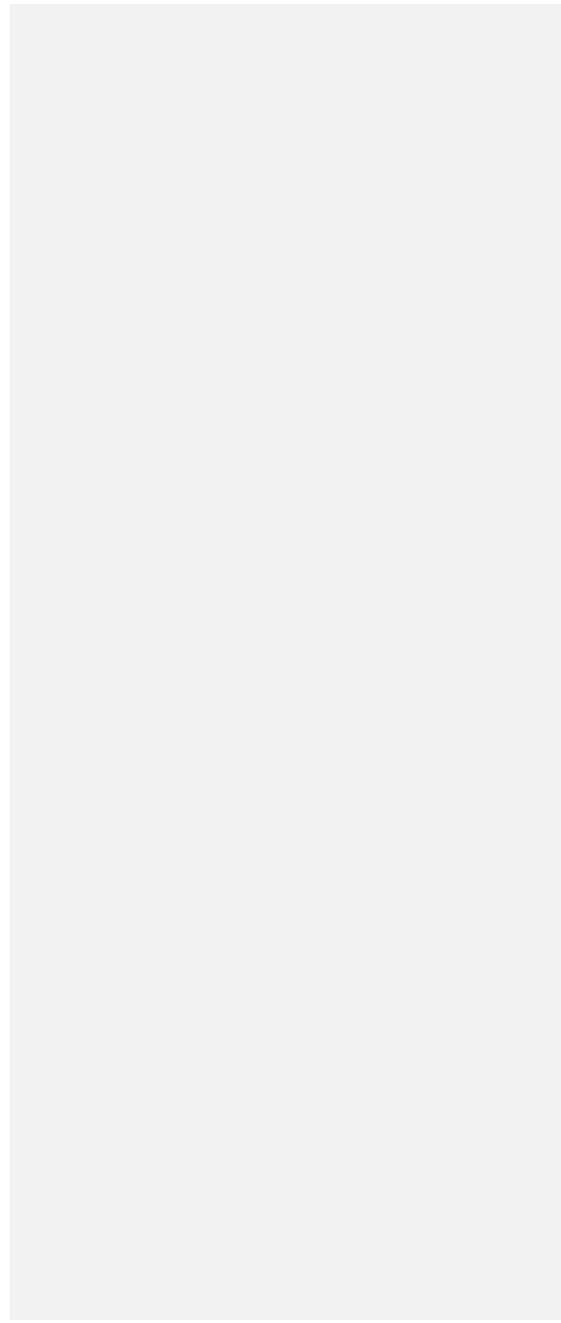
**Table 1.** Description of samples used in this study.

Sample Name	Short description	DNA concentration of samples (ng/ul)	Fraction of flow cell used	# reads base-called	Total length of reads base-called	% of reads classified as bacteria (based on WIMP)	Mean read length in bp	Max read length in bp	% reads >1000bp
L-K40	Tomato inoculated with <i>Pto</i> K40 in the laboratory	325.2	1	1,377,617	4.18 Gb	89%	3,037	66,015	64%
L-mix	Tomato inoculated with four <i>P. syringae</i> strains in the laboratory	450.4	1	1,006,978	4.16 Gb	95%	4,130	67,174	74%
L-mock	Non-inoculated tomato plant grown in the laboratory	33.6	1/7	82,412	103.22 Mb	8%	1,252	19,754	40%
L-culture-mix	Equal mix of 4 <i>P. syringae</i> strains grown in liquid culture	147.5	1/6	54,124	155.93 Mb	93%	2,880	76,060	39%
F1-bs	Tomato field sample with symptoms of bacterial spot	562	1/7	137,497	588.50 Mb	81%	4,280	55,436	73%
F2-bs	Tomato field sample with symptoms of bacterial spot	500.2	1/7	90,185	498.68 Mb	80%	5,529	65,598	74%
F3-bs	Tomato field sample with symptoms of bacterial spot	332.5	1/7	100,956	423.16 Mb	78%	4,191	59,405	68%
F4-bs	Tomato field sample with symptoms of bacterial spot	319.8	1/7	74,615	289.36 Mb	81%	3,878	51,268	70%

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47

F5-Septoria	Tomato field sample with symptoms of Septoria leaf spot	75.8	1/7	73,432	226.721 Mb	50%	3,087	43,967	59%
F6-healthy	Tomato field sample with no symptoms	29.1	1/7	35,923	66.58 Mb	31%	1,853	29,617	46%
F7-bs	Tomato field sample with symptoms of bacterial spot	331.8	1/7	118,391	432.08 Mb	75%	3,649	48,335	64%
F8-bs	Tomato field sample with symptoms of bacterial spot	154.2	1/2	106,059	371.84 Mb	70%	3,505	33,472	71%

For Peer Review



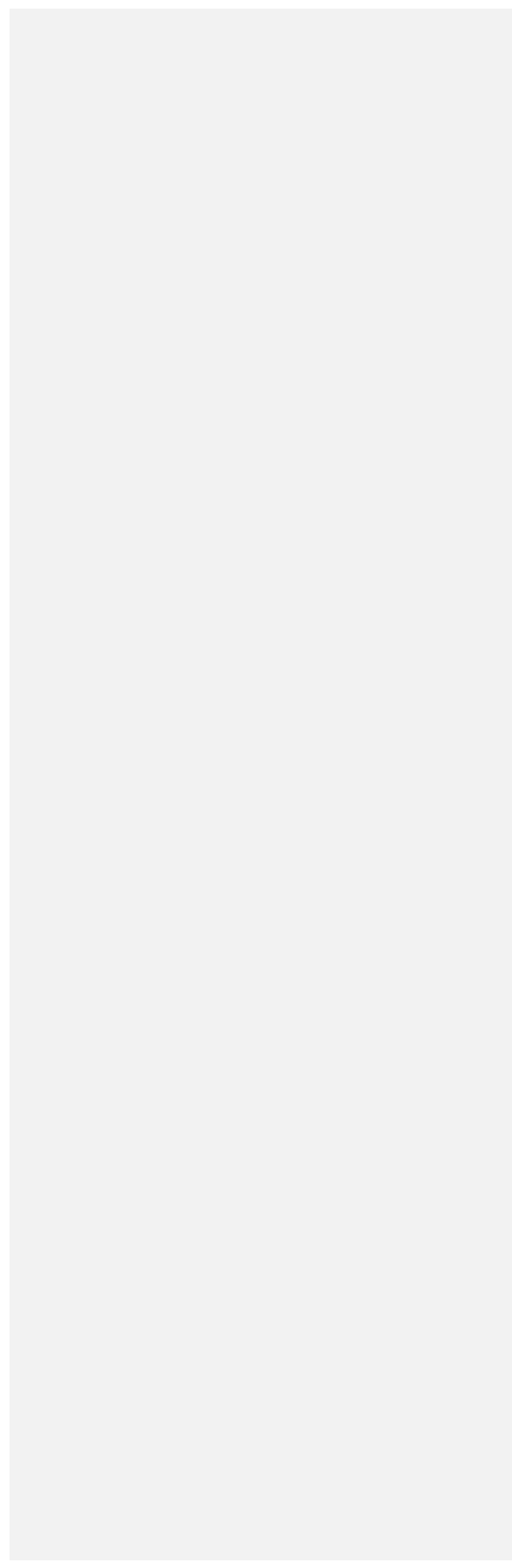
**Table 2.** Relative abundance results (top three hits) obtained with MetaMaps and Sourmash using a custom genome database of bacterial tomato pathogens and closely related isolates.

Sample	rank	MetaMaps	%	Sourmash	%
L-K40	1	<i>Pto</i> T1 ( <i>Pto</i> strain T1)	70.94	<i>Pto</i> T1 ( <i>Pto</i> strain T1)	71.65
	2	<i>Pto</i> NCPPB1108 ( <i>Pto</i> strain T1)	15.91	<i>P. syringae</i> pv. <i>actinidiae</i>	3.67
	3	<i>Pto</i> DC3000 ( <i>Pto</i> strain DC3000)	7.81	<i>P. syringae</i>	2.44
L-mix	1	<i>Pto</i> T1 ( <i>Pto</i> strain T1)	69.48	<i>Pto</i> T1 ( <i>Pto</i> strain T1)	65.98
	2	<i>Pto</i> NCPPB 1108 ( <i>Pto</i> strain T1)	15.23	<i>Pto</i> DC3000 ( <i>Pto</i> strain DC3000)	16.01
	3	<i>Pto</i> PT23	6.90	<i>P. syringae</i> pv. <i>actinidiae</i>	2.56
L-mock	1	<i>Clavibacter michiganensis</i> <sup>1</sup>	13.30	*no matches*	
	2	<i>Xp</i>	11.39	*no matches*	
	3	<i>Ralstonia solanacearum</i>	8.86	*no matches*	
L-culture-mix	1	<i>Pto</i> DC3000 ( <i>Pto</i> strain DC3000)	38.90	<i>Pto</i> DC300 ( <i>Pto</i> strain DC3000)	75.17
	2	<i>Pto</i> T1 ( <i>Pto</i> strain T1)	27.48	<i>Pto</i> T1 ( <i>Pto</i> strain T1)	19.63
	3	<i>Pto</i> NCPPB 1108 ( <i>Pto</i> strain T1)	9.07	<i>Pto</i> PT23	1.03
F1-bs	1	<i>Xp</i> TB9 ( <i>Xp</i> group 2)	29.37	<i>Xp</i> TB15 ( <i>Xp</i> group 2)	95.18
	2	<i>Xp</i> Xp9-5 ( <i>Xp</i> group 2)	28.03	<i>Xp</i> Xp17-12	1.05
	3	<i>Xp</i> Xp7-12	14.97	<i>X. campestris</i> pv. <i>durantae</i>	0.79
F2-bs	1	<i>Xp</i> Xp9-5 ( <i>Xp</i> group 2)	15.97	<i>Xp</i> strain Xp9-5 ( <i>Xp</i> group 2)	90.72
	2	<i>Xp</i> TB9 ( <i>Xp</i> group 2)	15.14	<i>Xp</i> strain Xp17-12	4.19
	3	<i>Xp</i> Xp7-12	10.38	<i>X. arboricola</i> pv. <i>pruni</i>	1.83
F3-bs	1	<i>Xp</i> Xp17-12	50.59	<i>Xp</i> strain Xp17-12	97.76
	2	<i>Xp</i> 91-118	9.00	<i>Xp</i> strain Xp9-5 ( <i>Xp</i> group 2)	1.27
	3	<i>Xp</i> LH3	4.67	<i>X. campestris</i> pv. <i>durantae</i>	0.98
F4-bs	1	<i>Xp</i> TB9 ( <i>Xp</i> group 2)	22.38	<i>Xp</i> TB15 ( <i>Xp</i> group 2)	97.28
	2	<i>Xp</i> Xp9-5 ( <i>Xp</i> group 2)	19.30	<i>Xp</i> Xp9-5 ( <i>Xp</i> group 2)	2.11
	3	<i>Xp</i> TB15 ( <i>Xp</i> group 2)	18.80	<i>X. campestris</i> pv. <i>viticola</i>	0.61
F5-Septoria	1	<i>X. campestris</i>	30.45	<i>X. arboricola</i>	57.08
	2	<i>X. arboricola</i>	25.60	<i>X. arboricola</i>	14.76
	3	<i>X. pisi</i>	2.78	<i>Xp</i> TB9	9.59
F6-healthy	1	<i>Xp</i> Xp9-5 ( <i>Xp</i> group 2)	11.70	<i>Xp</i> TB15 ( <i>Xp</i> group 2)	98.13
	2	<i>Xp</i> TB9 ( <i>Xp</i> group 2)	11.47	<i>Xp</i> LH3	1.87
	3	<i>Xp</i> Xp7-12	10.82	*no matches	
F7-bs	1	<i>Xp</i> TB9 ( <i>Xp</i> group 2)	23.40	<i>Xp</i> TB15 ( <i>Xp</i> group 2)	89.80
	2	<i>Xp</i> Xp9-5 ( <i>Xp</i> group 2)	19.15	<i>X. arboricola</i>	5.47
	3	<i>Xp</i> TB15 ( <i>Xp</i> group 2)	17.28	<i>X. campestris</i>	1.54
F8-bs	1	<i>Xp</i> Xp9-5 ( <i>Xp</i> group 2)	26.51	<i>Xp</i> Xp9-5 ( <i>Xp</i> group 2)	94.17
	2	<i>Xp</i> TB9 ( <i>Xp</i> group 2)	17.48	<i>Xp</i> TB15 ( <i>Xp</i> group 2)	1.62
	3	<i>Xp</i> Xp17-12	15.23	<i>Xp</i> Xp17-12	1.05

<sup>1</sup> for non-tomato pathogens only the species is reported strain names are only reported for tomato pathogens

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For Peer Review



**Table 3.** Description of metagenomic assemblies.

Sample name	Total number of contigs	Total assembly length in bp	Mean contig length in bp	Longest contig in bp	2nd longest contig in bp
L-K40	24	6,777,088	282,378	6,239,052	143,110
L-mix	73	8,849,360	121,224	6,267,841	121,245
L-mock	8	117,988	14,748	64,285	12,027
L-culture-mix	20	5,896,134	294,806	777,233	631,884
F1-bs	92	12,801,147	139,142	5,084,548	898,279
F2-bs	131	8,667,146	66,161	4,444,689	280,302
F3-bs	49	12,118,489	247,316	2,320,098	1,193,622
F4-bs	18	5,216,728	289,818	1,172,667	925,913
F5-Septoria	9	122,995	13,666	38,461	25,303
F6-healthy	4	21,571	5,392	8,666	7,821
F7-bs	35	5,784,684	165,276	5,146,049	57,056
F8-bs	10	5,449,373	544,937	2,745,990	2,264,789

**Table 4.** LINbase results for two longest contigs

Sample	Longest contig (ANI %)	Taxon membership of longest contig <sup>1</sup>	Second longest contig (ANI %)	Taxon membership of second <sup>2</sup> longest contig <sup>1</sup>	Two longest contigs merged (ANI %)	Taxon membership of merged contigs <sup>1</sup>
L-K40	<i>Pto</i> BAV1020 (99.76)	<i>Pto</i> strain T1	NA	NA	<i>Pto</i> BAV1020 (99.62)	<i>Pto</i> strain T1
L-mix	<i>Pto</i> BAV1020 (99.77)	<i>Pto</i> strain T1	NA	NA	<i>Pto</i> K40 (99.53)	<i>Pto</i> strain T1
L-culture-mix	<i>Pc</i> ICMP19117 (97.42)	<i>Pc</i>	<i>Ps</i> UB0390 (97.70)	<i>Ps</i>	<i>Pc</i> ICMP19117 (97.50)	<i>Pc</i>
F1-bs	<i>Xp</i> 10-13 (99.84)	<i>Xp</i> group 2	NA	NA	<i>Xp</i> 8-16 (99.85)	<i>Xp</i> group 2
F2-bs	<i>Xp</i> GEV2117 (99.78)	<i>Xp</i> group 2	<i>Xp</i> 7-12 (99.76)	<i>Xp</i>	<i>Xp</i> GEV1063 (99.78)	<i>Xp</i> group 2
F3-bs	<i>Pf</i> Pf0-1 (95.08)	<i>Pf</i>	<i>Pf</i> Pf0-1 (95.01)	<i>Pf</i>	<i>Pf</i> Pf0-1 (95.05)	<i>Pf</i>
F4-bs	<i>Xp</i> 91-118 (99.62)	<i>Xp</i>	<i>Xp</i> 91-118 (99.75)	<i>Xp</i>	<i>Xp</i> 91-118 (99.62)	<i>Xp</i>
F7-bs	<i>Xp</i> 8-16 (99.84)	<i>Xp</i> group 2	NA	NA	<i>Xp</i> GEV2116 (99.83)	<i>Xp</i> group 2
F8-bs	<i>Xp</i> 10-13 (99.76)	<i>Xp</i> group 2	<i>Xp</i> GEV2117 (99.78)	<i>Xp</i> group 2	<i>Xp</i> Xp10-13 (99.79)	<i>Xp</i> group 2
BAV6163	<i>Xp</i> GEV1063 (99.98)	<i>Xp</i> group 2				
BAV6164	<i>Xp</i> GEV1063 (99.98)	<i>Xp</i> group 2				

<sup>1</sup> based on taxon membership of the best hit

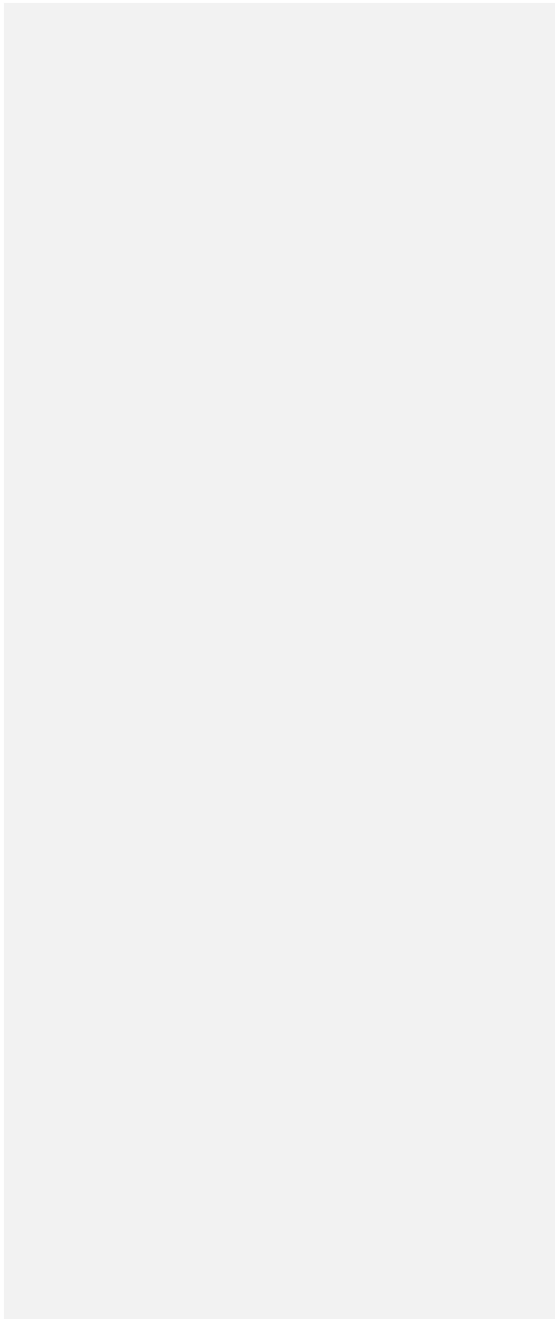
*Ps* = *P. syringae*, *Pf* = *Pseudomonas fluorescens*, *Pc* = *Pseudomonas congelans*, *Xp* = *X. perforans*



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47

NA – Not available, second contig too short for identification

For Peer Review



## Supplementary Tables

**Supplementary Table 1.** List of genomes used in the custom database.

**Supplementary Table 2.** Relative abundance values at the species level for all samples obtained with WIMP, Sourmash, and MetaMaps.

**Supplementary Table 3.** Example run times for WIMP, Sourmash, and MetaMaps.

**Supplementary Table 4.** BLASTN results for contigs of assembled metagenomes.

## Figure legends

**Figure 1.** Diseased tomato plants (A) Symptoms caused by *Pseudomonas syringae* pv *tomato* isolate K40 (strain *Pto* T1) in a laboratory-inoculation assay and (B) Bacterial spot symptoms in naturally infected plants during a disease outbreak on the Eastern Shore of Virginia.

**Figure 2.** [Screenshot of the output from the ONT tool WIMP showing the taxonomy assignment for sample L-K40. A taxonomy tree is depicted on the left and the distribution of reads across domains is shown on the right.](#)

**Figure 3.** Bar graph showing the comparison of results at the species level using the read-based programs WIMP, Sourmash and MetaMaps. Each barplot corresponds to individual lab samples used in the study. A = L-K40, B = L-mix, C = L-mock, and D = L-culture-mix. Relative abundance values are expressed as percentages of all sequences classified as bacteria.

**Figure 4.** Bar graph showing the comparison of results at the species level using the read-based programs WIMP, Sourmash and MetaMaps. Each barplot corresponds to individual field samples used in the study. A = F1-bs, B = F2-bs, C = F3-bs, D = F4-bs, E = F5-Septoria, F = F6-healthy, G = F7-bs and H= F8-bs. Relative abundance values are expressed as percentages of all sequences classified as bacteria.

**Figure 5.** Relative genome percentage abundance for each sample based on BLASTN using contigs as query against a custom genome database. All hits were filtered to e-values less than

1  
2  
3  
4  
5  
6  
7  
8  
9 or equal to 0.01 and the longest hit for each contig was considered to be the best hit.  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For Peer Review

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

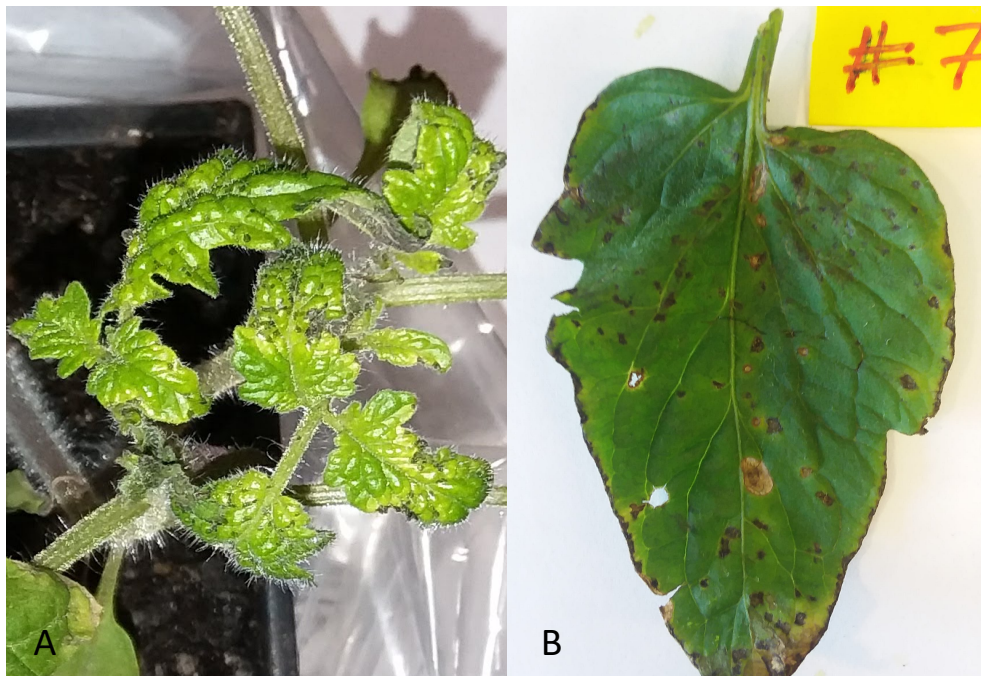


Figure 1

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

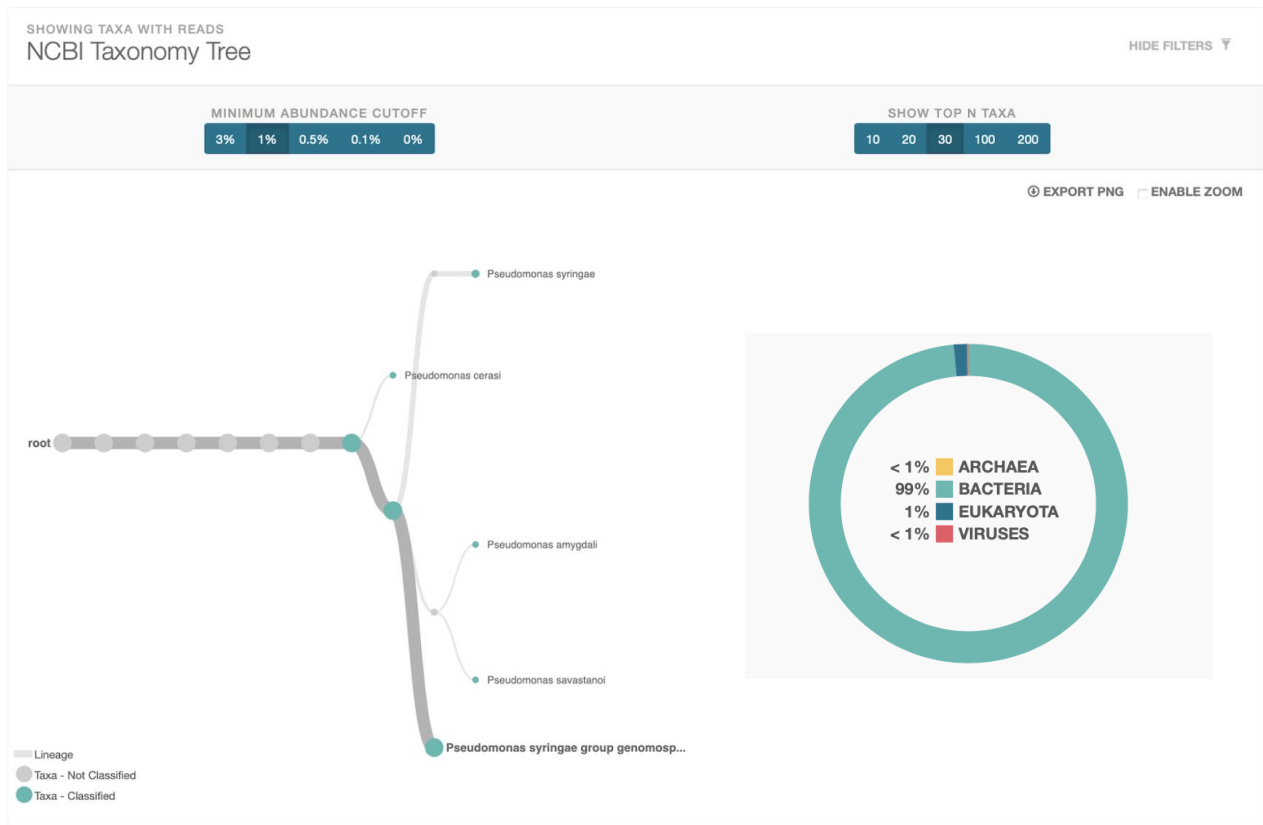


Figure 2

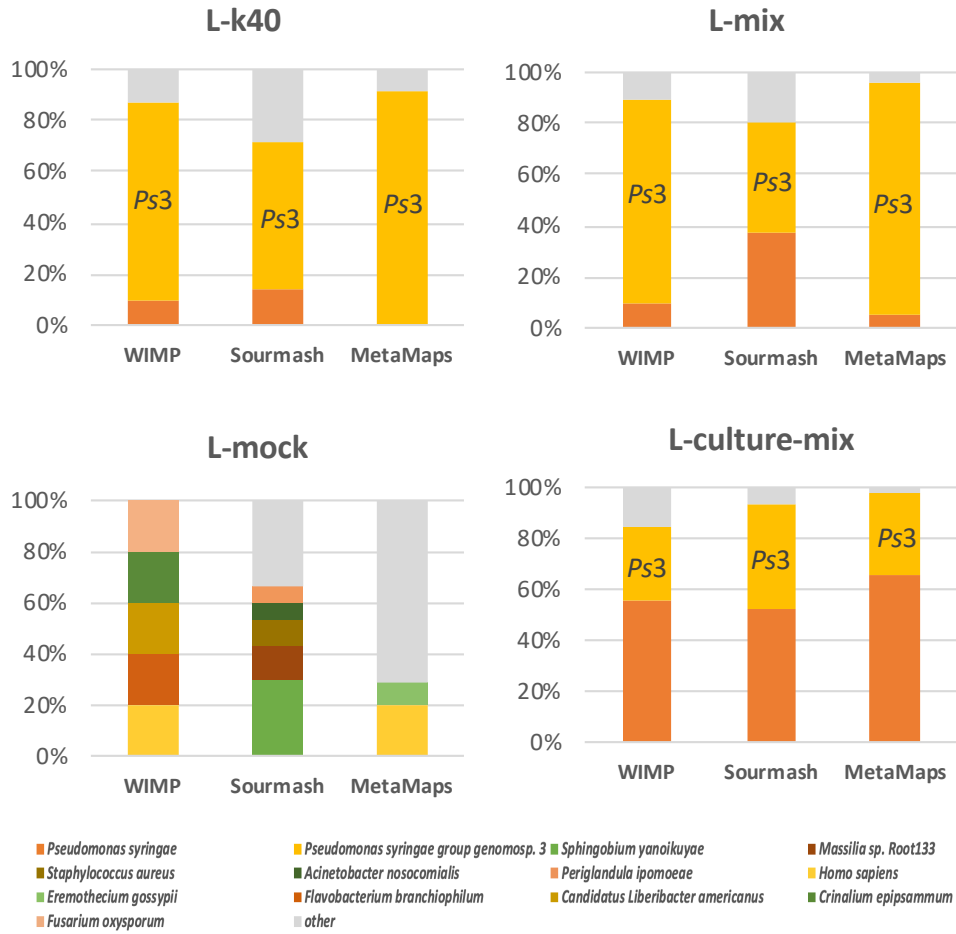


Figure 3

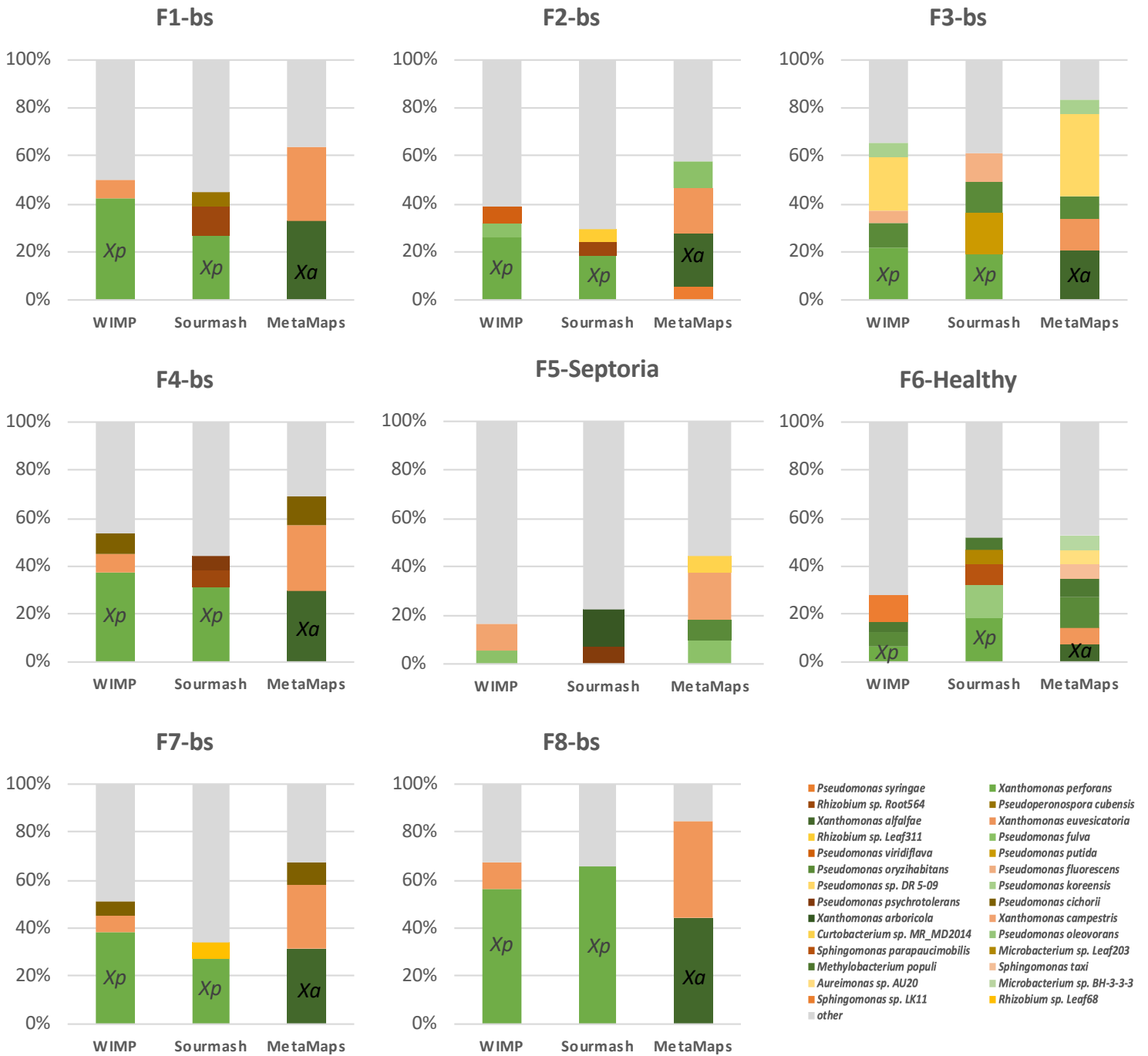


Figure 4

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

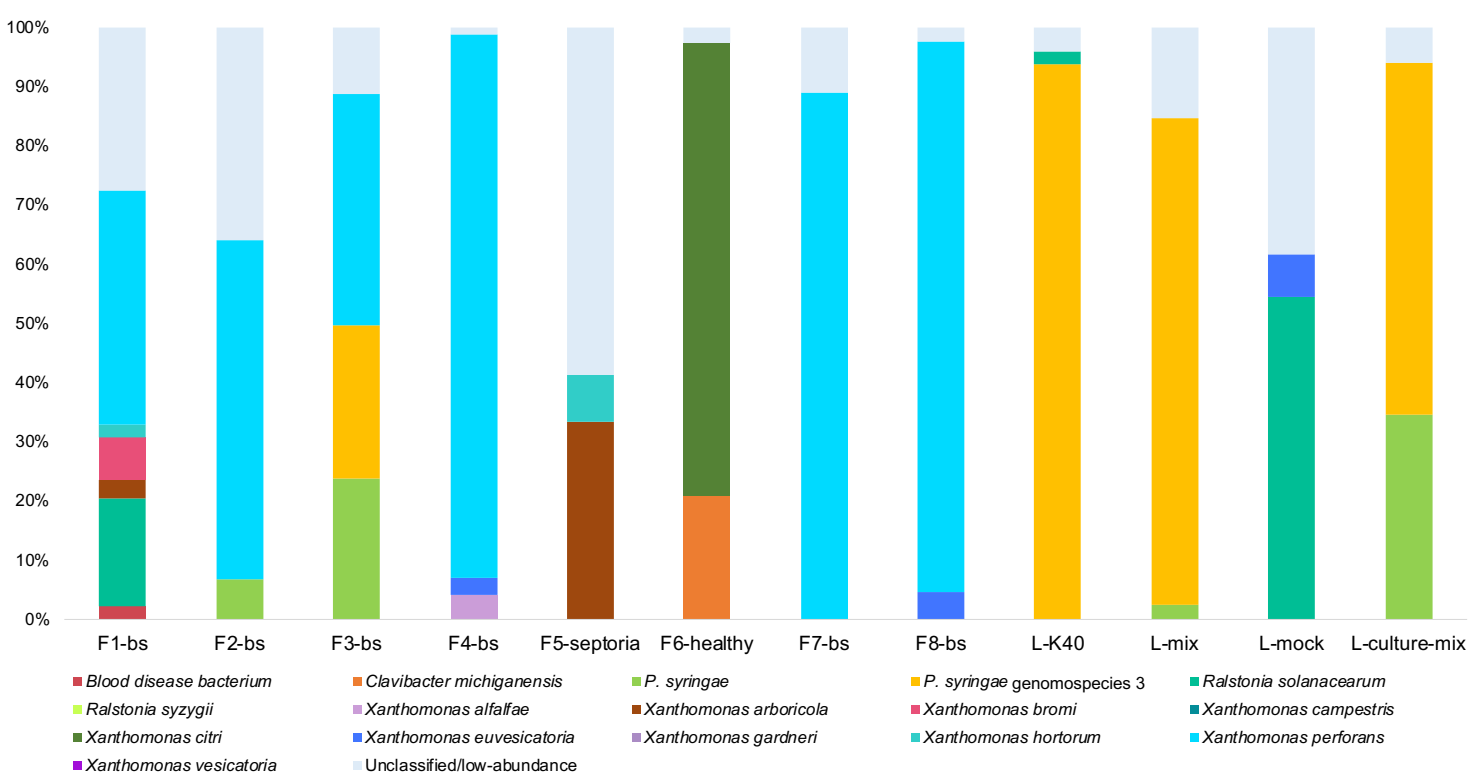


Figure 5