

Data in Brief

Draft genome sequence of *Massilia* sp. KIM isolated from South African grassland biome soils

Surendra Vikram^{a,b}, Nerissa Govender^{a,b}, Mubanga Helen Kabwe^{a,c}, Oliver Bezuidt^{a,b}, Thulani P. Makhwanyane^{a,b,*}

^a Centre for Microbial Ecology and Genomics, University of Pretoria, Pretoria 0028, South Africa

^b Department of Genetics, University of Pretoria, Pretoria 0028, South Africa

^c Department of Microbiology and Plant Pathology, University of Pretoria, Pretoria 0028, South Africa

A B S T R A C T

Massilia sp. are aerobic, Gram-negative, rod-shaped bacteria that are found in air, water, and soils. Here we describe the draft genome sequence of *Massilia* sp. KIM, isolated from the South African grassland soils. The total length of the genome was estimated at 5.73 Mb, comprised of 17 contigs. The draft genome has been deposited in the DDBJ/EMBL/GenBank under the accession MVAD10000000 and is available for download at: <https://www.ncbi.nlm.nih.gov/nucleotide/MVAD10000000>. Additionally, the raw short reads are available in the NCBI SRA database under the accession number: SRR5469241.

Specifications

Organism/cell line/tissue	<i>Massilia</i> sp. KIM
Sex	Not applicable
Sequencer or array type	Illumina HiSeq
Data format	Analyzed
Experimental factors	<i>Massilia</i> sp. KIM isolated from an environmental soil sample
Experimental features	Isolation of bacteria, Genome sequencing, de novo assembly, annotation, and analysis
Consent	Not applicable
Sample source location	Grassland biome, South Africa and Latitude & Longitude – 28.422, 24.551

1. Direct link to deposited data

<https://www.ncbi.nlm.nih.gov/nucleotide/MVAD10000000>
[https://www.ncbi.nlm.nih.gov/sra/SRX2754567\[accn\]](https://www.ncbi.nlm.nih.gov/sra/SRX2754567[accn]).
<https://www.ncbi.nlm.nih.gov/nucleotide/MF083079>

2. Introduction

Members of genus *Massilia* have been isolated from various samples

including clinical, soil, dust, water and the phyllosphere [1]. A number of these isolates have been retrieved from plant roots and rhizosphere soils, which suggests that *Massilia* may be crucial consortia of plant growth promoting rhizobacteria (PGPR) [1]. However, very little is known regarding the ecological roles of the genus *Massilia*, especially those found in soils. To reduce this knowledge deficit, we isolated *Massilia* sp. KIM from South Africa's grassland biome. The genome of sp. KIM was sequenced and used to explore the genetic and physiological of this strain as a PGPR.

2.1. Experimental design, materials, and methods

Massilia sp. KIM was isolated from soils collected from the South African grassland biome. Briefly, 0.5 g of soil was transferred into a sterile 2 mL tube containing 1 mL of deionized water and homogenized by vortexing at maximum speed for 10 s. The solution was then centrifuged at 11000 rpm for 60 s. 100 µL of the supernatant was then plated onto R2A agar plates, supplemented with the antifungal cyclohexamine (100 mg/mL). The bacterium grew optimally at 22 °C for 3 days. Genomic DNA was extracted according to a method described by Miller et al. through a combination of chemical lysis and beat-beating [2].

The strain was morphologically identified and confirmed by PCR amplification (using the primers E9F and U1510R), followed by sequencing and phylogenetic analysis (Fig. 1). High molecular weight genomic DNA was sent to the Molecular Research LP next generation

* Corresponding author at: Room 3-14, Centre for Microbial Ecology and Genomics, Natural Sciences 2, University of Pretoria, Lynnwood Road, Pretoria 0028, South Africa.
 E-mail address: thulani.makhwanyane@up.ac.za (T.P. Makhwanyane).

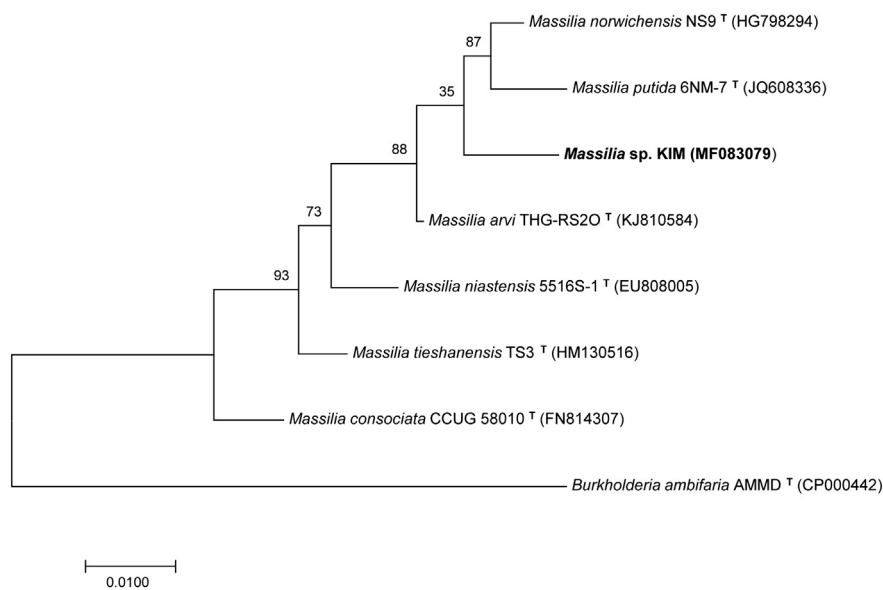


Fig. 1. A phylogenetic tree based on 16S rRNA gene sequences showing the relationship between *Massilia* sp. strain KIM (shown in bold) (NCBI accession: MF083079) and the type strains from the EZBioCloud server [3]. Mega v7.0.14 [4] was used to construct the tree based on the Maximum likelihood method and 500 bootstraps following a ClustalW alignment [5]. The 16S rRNA gene sequence of *Burkholderia ambifaria* AMMD is included as an outgroup. A total of 1375 positions were present in the final dataset. The scale bar represents 0.01 nucleotide substitutions per position.

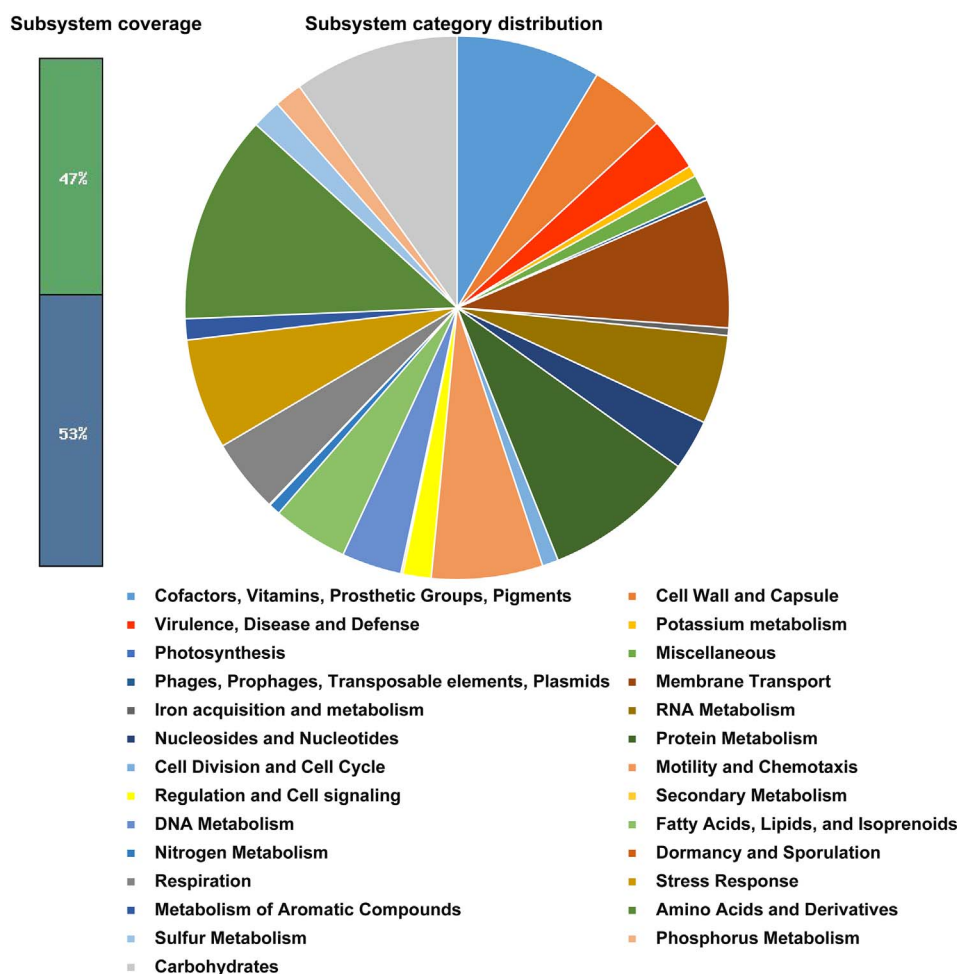


Fig. 2. An overview of the subsystem categories assigned to the genome of *Massilia* sp. KIM using RAST server. The green and blue column on the left side are showing the percentage of genes annotated to the “seed subsystem” and “not in the subsystem”, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

sequencing service (www.mrdnalab.com, Shallowater, TX, USA) for sequencing on the Illumina HiSeq platform (Illumina, Inc.) to obtain the 2 × 250 bp paired-end libraries. The genomic DNA library was prepared using the Nextera DNA Sample preparation kit (Illumina, Inc.) as detailed in the manufacturer's protocol. Sequence read quality was assessed using PRINSEQ lite version 0.20.4 [6] and the genome was assembled using SPAdes version 3.7.1 with the following parameters:

kmer *auto*, Cov_cutoff *auto* and careful options [7].

3. Data description

The draft genome sequence of *Massilia* sp. KIM constituted a total of 17 contigs (> 500 bp) with 5,734,388 bp, and a G + C content of 67.1% with N50 contigs size of 3,374,390 bp. The genome was

Table 1
Count of 79 tRNAs predicted in the genome of *Massilia* sp. KIM using ARAGORN webserver.

tRNA	Count
tRNA-Gln	1
tRNA-Thr	4
tRNA-Cys	1
tRNA-Asn	3
tRNA-Val	5
tRNA-His	2
tRNA-Asp	3
tRNA-Pro	5
tRNA-Tyr	1
tRNA-Met	5
tRNA-Leu	10
tRNA-Ile	1
tRNA-Arg	7
tRNA-Glu	6
tRNA-Ala	4
tRNA-Trp	2
tRNA-Phe	2
tRNA-Ser	6
tRNA-Lys	3
tRNA-Gly	8
Total	79

annotated using the PGAP pipeline [8] and Rapid Annotations using Subsystems Technology (RAST) server [9].

The RAST server predicted 5022 CDS's where 2344 CDS's (47%) were annotated as seed subsystem features (459 subsystems) and 2678 CDS's (53%) annotated as outside of the seed subsystem (Fig. 2). In total 3632 and 1390 CDSs were assigned as non-hypothetical and hypothetical, respectively. The tRNA analysis using ARAGORN v1.2.38 revealed 79 tRNA genes in the genome [10], the occurrence of different tRNAs in the draft genome is shown in Table 1. We also analyzed antibiotics and secondary metabolite production using the antiSMASH 3.0 server [11]. The genome of strain KIM harboured four major putative gene clusters linked to one unknown terpene, one complete carotenoid biosynthesis and two gene clusters for the bacteriocin (ribosomally synthesised antibacterial peptide) biosynthesis. Our analysis suggest that *Massilia* sp. KIM has extensive capacity for antibiotic and terpenoids production, which may form a basis for further characterization.

Acknowledgments

We wish to thank the Centre for High Performance Computing (CHPC), an initiative supported by the Department of Science and Technology of South Africa for access computational resources. The project was funded by the National Research Foundation of South Africa (Grant ID: 98117) and the University of Pretoria (Genomics Research Institute). SV was supported by the postdoctoral research fellowship from Claude Leon Foundation, South Africa.

References

- [1] M. Ofek, Y. Hadar, D. Minz, Ecology of root colonizing *Massilia* (Oxalobacteraceae), *PLoS One* 7 (2012) e40117 [doi.org/10.1371/journal.pone.0040117](http://dx.doi.org/10.1371/journal.pone.0040117).
- [2] D.N. Miller, J.E. Bryant, E.L. Madsen, W.C. Ghiorse, Evaluation and optimization of DNA extraction and purification procedures for soil and sediment samples, *Appl. Environ. Microbiol.* 65 (1999) 4715–4724.
- [3] S.H. Yoon, S.M. Ha, S. Kwon, J. Lim, Y. Kim, H. Seo, et al., Introducing EzBioCloud: a taxonomically united database of 16S rRNA and whole genome assemblies, *Int. J. Syst. Evol. Microbiol.* (2016), <http://dx.doi.org/10.1099/ijsem.0.001755>.
- [4] S. Kumar, G. Stecher, K. Tamura, MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets, *Mol. Biol. Evol.* 33 (2016) 1870–1874, <http://dx.doi.org/10.1093/molbev/msw054>.
- [5] J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res.* 22 (1994) 4673–4680.
- [6] R. Schmieder, R. Edwards, Quality control and preprocessing of metagenomic datasets, *Bioinformatics* 27 (2011) 863–864, <http://dx.doi.org/10.1093/bioinformatics/btr026>.
- [7] A. Bankevich, S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin, A.S. Kulikov, et al., SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, *J. Comput. Biol.* 19 (2012) 455–477, <http://dx.doi.org/10.1089/cmb.2012.0021>.
- [8] T. Tatusova, M. DiCuccio, A. Badretdin, V. Chetvernin, E.P. Nawrocki, L. Zaslavsky, et al., NCBI prokaryotic genome annotation pipeline, *Nucleic Acids Res.* 44 (14) (2016) 6614–6624, <http://dx.doi.org/10.1093/nar/gkw569>.
- [9] R.K. Aziz, D. Bartels, A.A. Best, M. DeJongh, T. Disz, R.A. Edwards, et al., The RAST server: rapid annotations using subsystems technology, *BMC Genomics* 9 (2008) 75, <http://dx.doi.org/10.1186/1471-2164-9-75>.
- [10] D. Laslett, B. Canback, ARAGORN, a program for the detection of transfer RNA and transfer-messenger RNA genes, *Nucleic Acids Res.* 32 (2004) 11–16, <http://dx.doi.org/10.1093/nar/gkh152>.
- [11] T. Weber, K. Blin, S. Duddela, D. Krug, H.U. Kim, R. Brucoleri, et al., antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters, *Nucleic Acids Res.* 43 (2015) W237–W243, <http://dx.doi.org/10.1093/nar/gkv437>.