ACCESS MICROBIOLOGY
an open research platform

RESEARCH ARTICLE
Awori *et al.*, *Access Microbiology* 2023;5:000531.v4
DOI 10.1099/acmi.0.000531.v4

MICROBIOLOGY SOCIETY
OPEN DATA OPEN ACCESS

Draft genomes, phylogenomic reconstruction and comparative genome analysis of three *Xenorhabdus* strains isolated from soil-dwelling nematodes in Kenya

Ryan Musumba Awori[1,2,*], Charles N. Waturu[3], Sacha J. Pidot[4], Nelson O. Amugune[5] and Helge B. Bode[1,6,7,8]

Abstract

As a proven source of potent and selective antimicrobials, *Xenorhabdus* bacteria are important to an age plagued with difficult-to-treat microbial infections. Yet, only 27 species have been described to date. In this study, a novel *Xenorhabdus* species was discovered through genomic studies on three isolates from Kenyan soils. Soils in Western Kenya were surveyed for steinernematids and *Steinernema* isolates VH1 and BG5 were recovered from red volcanic loam soils from cultivated land in Vihiga and clay soils from riverine land in Bungoma respectively. From the two nematode isolates, *Xenorhabdus* sp. BG5 and *Xenorhabdus* sp. VH1 were isolated. The genomes of these two, plus that of *X. griffiniae* XN45 – this was previously isolated from *Steinernema* sp. scarpo that also originated from Kenyan soils – were sequenced and assembled. Nascent genome assemblies of the three isolates were of good quality with over 70% of their proteome having known functions. These three isolates formed the *X. griffiniae* clade in a phylogenomic reconstruction of the genus. Their species were delineated using three overall genome relatedness indices: an unnamed species of the genus, *Xenorhabdus* sp. BG5, *X. griffiniae* VH1 and *X. griffiniae* XN45. A pangenome analysis of this clade revealed that over 70% of species-specific genes encoded unknown functions. Transposases were linked to genomic islands in *Xenorhabdus* sp. BG5. Thus, overall genome-related indices sufficiently delineated species of two new *Xenorhabdus* isolates from Kenya, both of which were closely related to *X. griffiniae*. The functions encoded by most species-specific genes in the *X. griffiniae* clade remain unknown.

DATA SUMMARY

NCBI GenBank accession numbers of the three genome assemblies generated from this study are JACWFC000000000.1, JADEUF000000000.1 and JADEUG000000000.1. The metadata for soil samples collected in this study are listed in Table S2 (available in the online version of this article).

Accession numbers and strain names of publicly available genomes used in this study are listed in Table S3.

The supplementary workbook contains detailed raw data used for pangenome analyses. IS family transposases in the genome of strain BG5 have been deposited in the ISfinder Database under accession numbers ISXsp1, ISXsp2, ISXsp3, ISXsp4, ISXsp5, ISXsp6, ISXsp7, ISXsp8, ISXsp9, ISXsp10, ISXsp11, ISXsp12, ISXsp13, ISXsp14, ISXsp15, ISXsp16, ISXsp17, ISXsp18, ISXsp19 and ISXsp20.

Received 28 November 2022; Accepted 27 January 2023; Published 22 May 2023
Author affiliations: [1]Molecular Biotechnology, Department of Biosciences, Goethe University Frankfurt, 60438 Frankfurt am Main, Germany; [2]Elakistos Biosciences, PO Box 19301-00100, Nairobi, Kenya; [3]Horticulture Research Institute, Kenya Agricultural and Livestock Research Organisation, PO Box 220 Thika; [4]Department of Microbiology and Immunology at the Doherty Institute, University of Melbourne, Melbourne, Australia; [5]Department of Biology, University of Nairobi, PO Box 30197-00100, Nairobi, Kenya; [6]Department of Natural Products in Organismic Interactions, Max Planck Institute for Terrestrial Microbiology, 35043 Marburg, Germany; [7]Chemical Biology, Department of Chemistry, Phillips University Marburg, 35043 Marburg, Germany; [8]Senckenberg Gesellschaft für Naturforschung, 60325 Frankfurt am Main, Germany.
***Correspondence:** Ryan Musumba Awori, ryan-musumba.awori@elakistosbiosciences.com
Keywords: prokaryotic pangenomics; *Steinernema* endosymbionts; species delineation; *Xenorhabdus* bacteria.
Abbreviations: ANI, average nucleotide identity; BLAST, basic local alignment search tool; CDS, coding DNA sequence; COG, cluster of orthologous groups; dDDH, digital DNA–DNA hybridization; EPNs, entomopathogenic nematodes; GBDP, genome BLAST distance phylogeny; GC, gene clusters; GGD, genome–genome distance; IS, insertion sequence; LPS, lipopolysaccharide; MCL, Markov clustering; OGRIs, overall genome-related indices; orthoANI, orthologous average nucleotide identity; SPR, subtree pruning and regrafting.
Three supplementary figures and three supplementary tables are available with the online version of this article.

Impact Statement

Xenorhabdus bacteria are important because they produce various antimicrobials. However, not many have been isolated from their natural habitat, which is the gut of soil-dwelling *Steinernema* nematodes. Two of these bacteria were thus isolated from nematodes of soils in Western Kenya. Their genome sequences were determined and used in genomic analyses, which revealed novel species. These genomes can be used to show the location of genes encoding antimicrobial production, thus making it easier for future isolation of antimicrobials from these new bacterial strains.

Parts of the methods, results and discussion sections were previously reported in a doctoral thesis of the first author and are thus not considered a prior publication.

INTRODUCTION

Bacteria of the genus *Xenorhabdus* naturally produce specialized metabolites such as non-ribosomal peptides that have anti-protozoal, antifungal and antibacterial activities [1]. Yet, despite each *Xenorhabdus* species [2] – and sometimes even strain [3]– encoding a unique antimicrobial production profile, only 27 [4–13] *Xenorhabdus* bacteria have been isolated from their *Steinernema* nematode hosts and described as novel species. Thus, the aim of this study was to discover new *Xenorhabdus* species and strains.

Bacteria of the genus *Xenorhabdus* are naturally found in soil biota, specifically as autochthonous endosymbionts of insect-killing *Steinernema* nematodes. Once a steinernematid enters an insect prey via natural openings such as spiracles, it migrates to the haemocoel and defecates [14] its *Xenorhabdus* endosymbionts, which then secrete specialized metabolites such as insecticidal toxins [15] that result in quick death of the insect. Secreted antimicrobials function to deter soil microbial competitors from the nutrient-rich cadaver. Nematodes then utilize this nutrient-filled, cadaver enclosure to reproduce exponentially. Depletion of nutrients halts the reproductive cycle and triggers nematode bacterium re-association. This is succeeded by pre-infective juvenile steinernematids emigrating from the cadaver to the soil, where they lie in wait for insect prey. Thus, to isolate a *Xenorhabdus* bacterium one needs to first isolate its infective juvenile steinernematid host from soil, using a combination of insect larvae as bait [16] and White traps [17].

For species such as *Xenorhabdus khoisanae* (Table S1), *X. bovienii*, *X. kozodoii*, *X. poinarii* and *X. hominckii,* we see one *Xenorhabdus* species as the natural symbiont of numerous *Steinernema* species [18]. However, the reverse, one *Steinernema* species that naturally hosts, with equal fitness, two different *Xenorhabdus* species, has yet to be discovered. Thus, there is a high possibility of identifying new *Xenorhabdus* species from the over 50 described *Steinernema* species whose symbionts remain uncharacterized [19].

This research gap between steinernematid isolation and *Xenorhabdus* endosymbiont identification is also seen in Sub-Saharan Africa, where numerous steinernematids have been isolated (see Table S1 for a full list of species and location). In Kenya, apart from *X. hominickii* from *Steinernema karii* [13], and *X. griffiniae* XN45 that we isolated from *Steinernema* sp. scarpo [20], endosymbionts have yet to be isolated and described from strains that were isolated from Central and Coastal Regions including *Steinernema* sp. UH3 [21] and UH13 [22]. No steinernematid isolates have been documented from the Western region of Kenya.

Genome assemblies of >50× coverage of new isolates are not only required for the description of novel species/emendation of prokaryotic taxa [23] but also enable accurate species delineation via overall genome-related indices (OGRIs) such as orthologous average nucleotide identity (orthoANI) [24] and digital DNA–DNA hybridization (dDDH) [25], and phylogenomic reconstructions based on genome-genome distances [26]. Furthermore, comparative genome analyses of closely related strains are useful for the identification of not only genomic islands and mobile genetic elements but also the core and dispensable genes of a specific monophyletic group [27, 28].

In this study two *Xenorhabdus* strains, VH1 and BG5, were isolated from soil biota from Western Kenya. Their genomes, and that of *X. griffiniae* XN45 that we previously isolated from *Steinernema* sp. scarpo from Kenya, were sequenced and assembled. These three genomes were used for downstream species delineation and emendation, and comparative genome analyses.

METHODS

Collection of field soil samples

Fieldwork was carried out from 16 October 2018 to 4 November 2018 in the Western and Rift Valley Regions of Kenya. No access permits were required as per the exceptions of section 3(d) of the Environmental Management and Coordination (Conservation of Biological Diversity and Resources, Access to Genetic Resources and Benefit Sharing) Regulations 2006 of the Environmental Management and Coordination Act, 1999 of the Laws of Kenya. Ten localities were selected for the collection of soil samples:

Nandi Hills, Tinderet, Fort Tenan, Kakamega, Gisambai, Vihiga, Kisumu, Bungoma, Kaimosi and Mt. Elgon. Within each locality, collection points were selected from cultivated lands, fallow lands, forests, crop edges, shorelines, swamps and riverine areas. This resulted in a total of 76 soil collection points. To collect a soil sample, vegetation was first cleared from the topsoil. Then using a digging fork, soil was excavated to a depth of not more than 60 cm. Using a collection spade, the soil was scooped into a measuring cup to an amount of ca. 500 g. Twigs, branches and stones were removed before the soil sample was placed in labelled cotton bags. Dug-out soil was returned to the hole and soil samples were then transported at room temperature to the laboratories. Geographical coordinates, altitude and descriptions of soil collection points are provided in Table S2.

Isolation of nematodes from soils samples

To isolate entomopathogenic nematodes (EPNs) from soils, a soil sample was first spread out on a tray and crumbled. The soil was then redistributed into transparent polyethylene terephthalate (PET) plastic containers of 20 cm in diameter and 5 cm in depth. To bait EPNs from the soil, *Galleria mellonella* larvae were first obtained from a laboratory insect culture of KALRO-Horticulture Research Institute, where they had been reared as previously described [29]. From this culture, any healthy larvae were selected. Two/three larvae were then buried in the soil, in a hole of ca. 1 cm diameter and 5 cm depth. In total, about 15 *G. mellonella* larvae buried in five holes per container were used as bait. The container lids were replaced and set-ups were maintained at room temperature. After a maximum of 7 days, containers were checked. Dead larvae were assessed for the following characteristics that typify an EPN infection: limp cadaver, tan or red in colour, and minimal smell of putrefaction. Samples BG5 and VH1 had dead cadavers that were either light red or tan in colour. Sample BG5 was clay soil collected from fallow riverine land. Sample VH1 was collected from land cultivated with cabbages. To isolate putative EPNs from these cadavers, a modified White trap [17] was used. Briefly, clean PET containers of 20 cm diameter and 5 cm depth were filled with distilled water to a depth of ca. 4 mm. A clean Petri dish was placed upside down into the container such that the Petri dish surface was raised from the bottom of the PET container. Clean white cotton cloths of the same size as the Petri dish were placed on this raised surface. Selected cadavers were placed onto the cotton cloths. To allow putative EPNs to emigrate from the cadavers to the water, a part of the cloth was dipped in the distilled water. PET containers were covered and kept for 7 days. The distilled water was observed daily under a dissecting microscope for the presence of white motile, ca. 1 mm long nematodes. For positive samples, contaminants such as cadaver tissue debris were separated from nematodes by a series of sedimentation and decanting using clean distilled water. Nematodes were stored in contamination-free distilled water – to a depth of not more than 0.4 cm – in clear plastic containers in the dark. Stored EPN nematode cultures were named after their soil collection point.

Isolation of bacteria from nematodes

The indirect isolation of *Xenorhabdus* bacteria from haemolymph was based on a previously described method [30] with modifications [20]. First, nematode isolates from collection points BG5 and VH1 were selected. Cadavers with which BG5 and VH1 nematodes were baited were surface-sterilized and dissected under aseptic conditions. A light-yellow, viscous, heterogenous fluid was aseptically obtained and streaked onto nutrient agar supplemented with 0.0025% (w/v) bromothymol blue and 0.004% (w/v) 2,3,5 triphenyl tetrazolium chloride (NBTA). This was incubated at 30 °C for 96 h. Only colonies that had the following observed morphologies were selected for further pure culture techniques: blue/yellow pigmentation, irregular margins, umbonate shape and visible swarming patterns. On these pure cultures, a catalase test was performed, and the absence of bubble production indicated a catalase-negative isolate, and these were presumptively identified as *Xenorhabdus* species. They were named *Xenorhabdus* sp. strains BG5 and VH1.

Genome sequencing and assembly

Previously, we isolated *X. griffiniae* XN45 from *Steinernema* sp. scarpo, which was originally isolated from Muran'ga District in Kenya [20]. Thus, in addition to *Xenorhabdus* sp. strains VH1 and BG5, this strain was selected for genome sequencing and assembly. DNA from strain XN45 was extracted with FastDNA Spin Kit for Soil (Mp Bio) to yield a concentration of 20 ng μl^{-1} and UV absorbance ratio at $260_{nm}/280_{nm}$ (A260/280) >1.8. From this, only 0.1 ng of DNA was used to prepare a library using a Nextera XT kit (Illumina). Sequencing was done by CeGaT GmBH on a NovaSeq 6000 platform with the following parameters: short insert paired-end reads of 100 bp and targeted coverage of 100×. Output data were raw sequence reads in fastq.gz format (2.902 GB), which had Illumina standard Phred scores (offset +33) and adapter sequences already removed. In terms of quality, 91.32% of reads had a Q30 value. Genome assembly was done with Spades 3.10.1 [31] with thresholds for minimum contig length and coverage set at 1000 bp and 5× respectively.

For VH1 and BG5, DNA was isolated with Gentra Puregene DNA extraction kit (Qiagen) to yield samples of 1 µg μl^{-1} concentration and A260/280 ratios of >1.8. At the Doherty Institute, University of Melbourne, Australia, DNA libraries were created using a Nextera XT DNA preparation kit (Illumina), and whole genome sequencing was performed on a NextSeq platform (Illumina) with paired-end reads of 150 bp and targeted coverage of >50×. Genome assembly was done with Spades 3.10.1 [31]. For assembled genomes of strains VH1, BG5 and *X. griffiniae* XN45 from this study and *X. griffiniae* strain BMMCB

(doi: 10.1128/genomeA.00785–15) from Mothupi *et al.* [32], characteristics such as completeness, contamination, N50, L50, length and GC content were determined using the comprehensive genome analysis tool of the PATRIC platform [33].

Phylogenomic reconstruction and calculation of ANI values

For phylogenomic reconstruction of the genus *Xenorhabdus*, 27 fasta files (Table S3) were used as input data for a whole genome-based taxonomic analysis on the Type strain genome server platform (TYGS) [26]. On these, the MASH algorithm [34] was used to quickly calculate intergenomic relatedness and determine the strains with the smallest distances. All pairwise comparisons and inference of intergenomic distances among the set of genomes were conducted using the genome BLAST distance phylogeny (GBDP) 'trimming' algorithm and distance formula *d5* [35]. One hundred distance replicates were calculated for each. The genome–genome distance calculator (GGDC) 2.1 formula was used to calculate dDDH values and confidence intervals [35]. Confidence intervals are given in workbook S1. Intergenomic distances were then used to infer a balanced minimum evolution tree with branch support via FASTME 2.1.4 including SPR post-processing [36]. Branch support was inferred from 100 pseudo-bootstrap replicates for each. Rooting of trees was done at the midpoint whereas visualization and graphics editing were done with iTOL [37] and Inkscape [38] respectively.

Using the same workflow, a phylogenomic reconstruction with only strains VH1, BG5, BMMCB and XN45 was made. Minimum thresholds for two strains to be classified as one species and sub-species were 70% and 79% dDDH respectively [26]. To calculate ANI values among species most closely related to strains XN45, VH1 and BG5, the orthoANI algorithm was used within the OAT software package, which was also used to obtain genome–genome distance (GGD) 2.1 values [24].

Creation of pangenomes

To determine whether the genus *Xenorhabdus* had an open pangenome, genomes of *Xenorhabdus* species were first used to construct a pangenome of the genus using the anvio v7.1 pangenome workflow [28, 39] with the following parameters: use ncbi-blast, MCL inflation=10, minbit=1, exclude-partial-gene-calls. A pangenome of strains VH1, BG5, XN45 and BMMCB only was also created using the same workflow and parameters. The strain names, accession numbers and total number of gene calls of each genome used are listed in Table S3. The mean α value was determined using the P-GAP platform running on the Panweb server [40]. Briefly, RAST-k [41] annotated genomes were used as data input on the Panweb server and the following parameters were selected for clustering genes into one gene cluster: minimum 80% nucleotide similarity, and minimum 80% coverage with gene family algorithm.

Estimation of the effect of draft genomes on the determination of the core genome

To estimate how the use of draft genomes affects the determination of the core genome, two additional pangenomes were created. The first contained six genomes, each of which was composed of fewer than two contigs: *X. bovienii* CS-03 (NZ_FO818637), *X. hominickii* ANU (NZ_CP016176), *X. cabanillasii* DSMZ 19705 (NZ_QTUB01000001), *X. poinarii* G6 (FO704551), *X. nematophila* AN6/1 (FN667742) and *X. szentirmaii* US123(NIUA01000001). The second contained draft genomes of similar species: *X. bovienii* T228 (JANAIF000000000.1), *X. hominickii* DSM 17903 (NJAI00000000.1), *X. cabanillasii* JM26 (NJGH00000000.1), *X. poinarii* SK (JADLIG000000000.1) and *X. nematophila* C2-3 (JRJV00000000.1). Pangenomes were created via the aforementioned anvio workflow and the sizes of their resultant core genomes were compared.

Analysis of gain and loss of gene clusters in the *X. griffiniae* clade

Using the gene clusters (GCs) of the VH1-BG5-XN45-BMMCB pangenome, a matrix of the presence and absence of GCs among the four strains was created (workbook S1). This matrix and the GBDP phylogeny of the four strains were then used as input data for gene gain and loss analysis in the COUNT program (downloaded 17 January 2023) using Wagner parsimony (penalty=1) [42].

Characterization of core, accessory, species and strain-specific genes

By using the 'search' and 'bin' functions of the anvi´o-interactive program, GCs that were present in all genomes under analysis as single copies were obtained and binned as single-copy core GCs (SCGs). Other binned GCs were: strain BG5 specific, strain BMMCB specific, strain XN45 specific, strain VH1 specific, XN45-VH1 accessory/*X. griffiniae* species specific, XN45-VH1-BG5 accessory, XN45-VH1-BMMCB accessory and BMMCB-BG5 accessory. Using the anvi-get-sequences-for-gene-clusters program with the 'report DNA sequences' and 'concatenate' flags, sequences for the single-copy GCs of each of the bins were obtained. For those consisting of GCs that constituted genes from multiple genomes, sequences from a single genome were used to represent the GC. These sequences were annotated in PROKKA [43] to elucidate the functions encoded by predicted genes.

Clustering of gene clusters into functional groups

The functions of GCs were determined by manually querying the UniProt Knowledgebase [44] with each annotated gene symbol. Then, GC functions were assigned based on the described biological process the gene most clearly contributed to. This was supplemented by querying GCs with assigned cluster of orthologous groups (COG) functions against the respective database

Table 1. Quality and characteristics of genome assemblies

Genomes of BG5, XN45 and VH1 were assembled in this study. BMMCB was obtained from NCBI GenBank (LDNM00000000.1). Annotation statistics were determined via the PATRIC platform.

	***Xenorhabdus* sp. strain BG5**	***X. griffiniae* XN45**	***Xenorhabdus* sp. strain VH1**	***Xenorhabdus* sp. strain BMMCB**
Contigs	129	381	273	231
Guanine–cytosine content (%)	43.80	43.57	43.65	44.68
Contig L50	12	29	43	21
Contig N50 (bp)	102633	4529	29298	57901
Genome length (bp)	3933551	4215754	4224998	4183760
Fine consistency (%)	96.5	95.9	95.9	95.7
Coarse consistency (%)	97.0	96.7	96.7	96.7
Contamination (%)	0.7	0	0	0
Completeness (%)	100	100	100	100
CDS	3827	4232	4160	4318
Repeat regions	127	66	69	70
Hypothetical proteins	973	1175	1185	1193
Proteins with functional assignments	2854	3057	2975	3125

Bp, base pairs; CDS, coding DNA sequences; Contig L50, the minimum number of contigs that contain 50% of the assembly; Contig N50, the shortest contig among that minimum number of contigs, which contain 50% of the assembly.

[45]. To aid the identification of genes that encode the biosynthesis of specialized metabolites, nucleotide sequences for each bin were annotated in antiSMASH [46]. Then, GC-function lists were compiled for each bin and manually curated. GCs were grouped according to similarity of function and visually represented in column graphs.

Elucidation of genomic islands in strain BG5

To highlight putative genomic islands flanked by transposase genes, an annotated record of the BG5 genome was concatenated and used as the reference genome in BRIG [47] and compared to genomes of VH1 and XN45 by the BLAST algorithm [48] utilizing an NCBI-blast 2.4.0+ bin library. Selected rings to be visualized were for BG5 genome guanine-cytosine (G+C) content (ring 1) and skew (ring 2), VH1 genome (ring 4), XN45 genome (ring 5) and loci of coding DNA sequences (CDS) annotated as transposases on the BG5 genome (ring 6). Output visualizations were obtained as .svg files and enhanced in Inkscape [38].

RESULTS

Strains VH1, BG5, XN45 and BMMCB form a clade

Two putative *Steinernema* isolates, VH1 and BG5, were isolated from soils in Western Kenya. VH1 was isolated from red volcanic loam soils on cabbage cultivated land at a point with coordinates 0.06293, 34.72903 and altitude 1624 m, in Vihiga. BG5 was isolated from clay soils on riverine land at a crop edge at a point with coordinates 0.48044, 34.40836 and altitude 1239 m, in Bungoma. From these two, *Xenorhabdus* sp. strains VH1 and BG5 were respectively isolated. Soils from the Rift Valley Regions sampled did not yield any steinernematids. Previously, we isolated *X. griffiniae* XN45 from *Steinernema* sp. scarpo, which originated from soils in Muran'ga County, Kenya [20]. *Xenorhabdus* sp. strain BMMCB, which was designated as an *X. griffiniae* species, was previously isolated from *Steinernema* sp. BMMCB [32], whose natural habitat was red volcanic sandy-loam soils in Brits, North West Province, South Africa [49].

Thus, to investigate the phylogenomic relationships between these four strains, the quality of their genome assemblies was first determined. The draft genomes of XN45, VH1, BG5 and BMMCB were complete and consistent, with low contamination (Table 1), and of coverage >50×. They were thus of sufficient quality for species delineation via overall genome relatedness indices [23]. For the nascent BG5, XN45 and VH1 assemblies, 74.5, 72.2 and 71.5% of proteins encoded in their respective genomes had known functions (Table 1), which were slightly below the 80% average for genomes of *Gammaproteobacteria* [45].

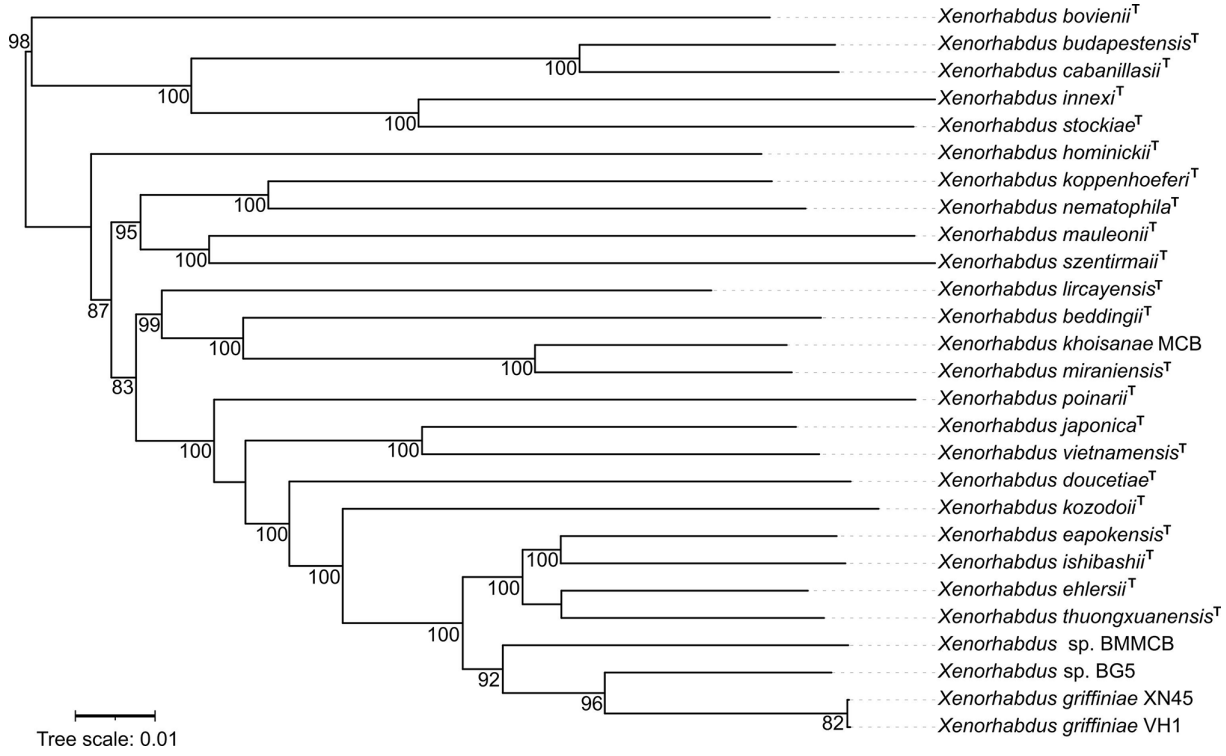

Fig. 1. Phylogenomic reconstruction of *Xenorhabdus* species using genome BLAST distance phylogeny approach (GBDP) distances calculated from genome sequences using the d_5 distance formula. Genome sequences of *Xenorhabdus* sp. strains VH1 and BG5 and *X. griffiniae* XN45 were obtained in this study. *Xenorhabdus* sp. strain VH1 and *X. griffiniae* XN45 clustered together. *Xenorhabdus* sp. strain BMMCB was previously classified as representing an *X. griffiniae* species. However, it did not cluster with *X. griffiniae* XN45. *Xenorhabdus* sp. strain BG5 did not form a clade with any extant species. GBDP pseudo-bootstrap values of above 40% are shown. The scale bar represents substitutions per site.

Strains VH1, BG5, BMMCB and XN45 formed an exclusive clade, as demonstrated in a phylogenomic reconstruction of 24/27 described species of the genus (Fig. 1). They were more closely related to each other than to other species, and this is demonstrated by the most closely related strains to VH1, BG5, XN45 and BMMCB being XN45, XN45, VH1 and BG5 (Fig. 2) at genome–genome distances (GGD) of 0.00, 0.059, 0.00 and 0.08 respectively (Table 2).

Strains XN45 and VH1 were conspecific, as demonstrated by their GGD, dDDH and ANI values of 0.000, 99.9% and 99.9% respectively– these were all within the conspecific thresholds of <0.0361 for GGD, >70% for dDDH and >95.1% for orthoANI [24, 26, 35]. XN45 and BG5 were most closely related to each other but were not conspecific as their GGD, dDDH and ANI values of 0.059, 67.3% and 94.24% respectively were all outside conspecific thresholds. BG5 and BMMCB were most closely related to each other but were not conspecific, as seen from their GGD, dDDH and ANI values of 0.08, 57.1% and 92.25% respectively.

BMMCB and XN45 were both described as *X. griffiniae* species [20, 32]. We previously demonstrated XN45 and *X. griffiniae* ID10^T as conspecific based on percentage nucleotide similarities for their 16S rRNA, *recA* and *serC* genes as 99.595, 98.571 and 97.686% respectively. These were above the same species thresholds of 98.65, 97 and 97% respectively [11, 50]. Conversely, strains BMMCB and XN45 were not conspecific, as demonstrated by their respective percentage nucleotide similarities values of 98.545, 93.67 and 92.066% for 16S rRNA, *recA* and *serC* genes respectively [20]. Indeed, BMMCB was not an *X. griffiniae* species as its GGD, dDDH and ANI values with *X. griffiniae* XN45 were 0.08, 50.4% and 91.41% respectively – these were all outside conspecific thresholds. This was corroborated by a difference of 1.11% in G+C content between the two genomes (Table 1), which was above the 1% same species threshold [51]. Taken together, these results demonstrated that these four strains represented three species: *X. griffiniae* XN45, *X. griffiniae* VH1, and the two undescribed species *Xenorhabdus* sp. BG5, and *Xenorhabdus* sp. BMMCB.

The genus *Xenorhabdus* has an open pangenome

We hypothesized that the genus *Xenorhabdus* had a pangenome that included many strain-specific genes, due to the numerous *Xenorhabdus* strains and their respective genomes, which have not yet been isolated from under-investigated *Steinernema* species [19]. This would make it an open pangenome. The pangenome is the pool of genes from which all the taxon genomes are

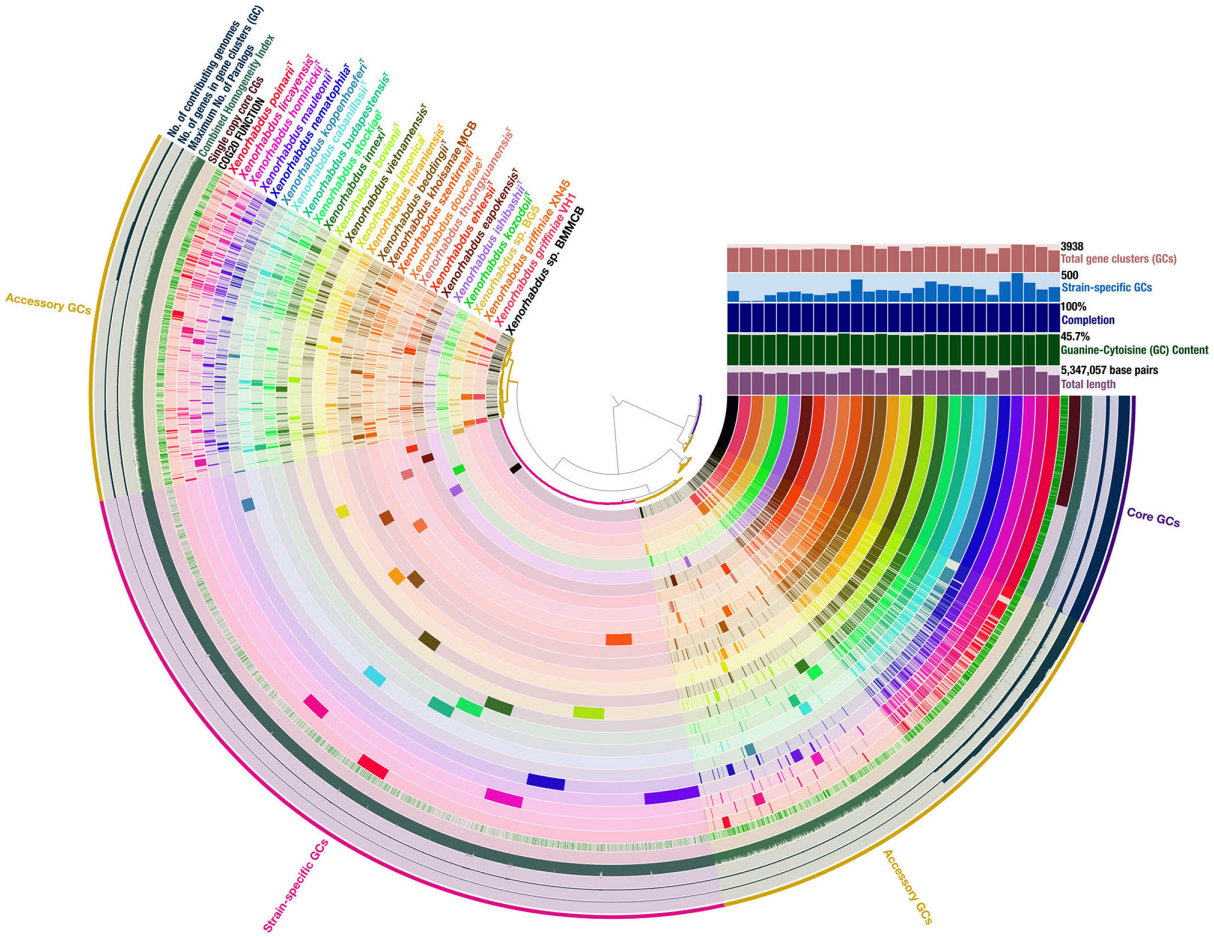

Fig. 2. Graphical representation of the pangenome of 26 species of the genus *Xenorhabdus*. The largest genome was 5347057 bp and the highest guanine–cytosine content was 45.7%. The pangenome was composed of a total of 13 469 gene clusters (GCs). Core GCs, those found in all 27 genomes, numbered 1654 in total. Accessory GCs, those found in two to 26 genomes, numbered 5992 in total. Strain-specific GCs, those found in one genome only, numbered 5820 GCs in total.

constituted. The core genes are those found in all genomes, accessory genes are those found in two or more genomes, and strain-specific genes are those found in one genome only [27].

The *Xenorhabdus* pangenome was composed of 27 genomes (from 24/27 described species) each of which had between 2771 (*X. koppenhoeferi*) and 4990 (*X. hominickii*) genes. It contained 101 832 genes, after the exclusion of 3668 partial gene calls from the analysis (Table S2). The pangenome contained a total of 13 469 GCs (Fig. 2). The core genome had 1654 GCs (12.3%). However, this likely to be an underestimate since draft genomes were mostly used in the analysis, and the use of draft as opposed to complete genomes reduced the core genome size by 5%, in our comparison of pangenomes of similar species (Fig. S3). The accessory genome had 5992 GCs (44.5%) and the strain-specific genome had 5820 GCs (43%). In total, 6834 GCs (50.7%) encoded known functions as per the Clusters of Orthologous Genes (COG) database. The highest percentage of these were in the core genome whereas the strain-specific genome had the lowest. *X. mauleonii* had the largest number of strain-specific genes (500) whereas *X. ehlersii* had the fewest (118). *X. griffiniae* VH1 and XN45 had remarkably few strain-specific GCs, 13 and 17 respectively, as they were the only two strains from the same species (Fig. 2). Strain-specific genes from newly added genomes increase the pangenome, and the rate of new strain-specific genes per newly added genome decreases to zero. If this rate of decrease is high, the result is a closed pangenome [27] whose size does not change significantly with the addition of new genomes. Conversely, open pangenomes have a slow rate of decreasing number of new strain-specific genes per newly added genome. Moreover, they are typified by a small percentage of core genes [52]. Relatedly, the mean α exponent of Heaps' Law is used to estimate this rate of decrease [53], and pangenomes with values >1 are defined as closed whereas those <1 are defined as open [27]. Thus, *Xenorhabdus* has an open pangenome, as demonstrated by its mean α exponent of Heaps' Law of 0.2752 and a core genome of 12.2% (1654/13 469). This corroborated previous global comparative genome analyses of the genus *Xenorhabdus* [2, 28, 54].

Table 2. Orthologous average nucleotide identity (orthoANI), genome-to-genome distance (GGD) and digital DNA–DNA hybridization (%) (dDDH) values for type species most closely related to *Xenorhabdus* sp. strains VH1 and BG5, *X. griffiniae* XN45 and *Xenorhabdus* sp. strain BMMCB

OrthoANI values are in the top half of the matrix (top triangle), GGD in parentheses, and values for dDDH are in the bottom half of the matrix (bottom triangle). Values that are within the threshold for two strains to be classified as one species are shaded in grey. The thresholds for conspecific strains are orthoANI values above 95.1%, dDDH values above 70% and GGD values below 0.0361. Type strains are: DSM 2270, *X. ishibashii*; DL20, *X. eapokensis*; DSM 16337, *X. ehlersii*; 30TX1, *X. thuongxuanensis*.

	DSM 22670	**XN45**	**DL20**	**DSM 16337**	**VH1**	**BG5**	**30T×1**	**BM MCB**
BM MCB	89.68 (0.103)	91.41 (0.088)	90.03 (0.010)	91.95 (0.082)	91.42 (0.088)	92.25 (0.080)	90.56 (0.094)	–
30T×1	92.55 (0.076)	91.02 (0.091)	93.42 (0.067)	93.76 (0.064)	91.01 (0.091)	91.72 (0.085)	–	49.6
BG5	90.44 (0.097)	94.24 (0.059)	90.94 (0.092)	93.08 (0.071)	94.22 (0.059)	–	57.3	57.1
VH1	89.80 (0.102)	99.99 (0.000)	90.32 (0.097)	91.95 (0.081)	–	67.4	51.1	50.5
DSM 16337	92.11 (0.080)	92.05 (0.081)	92.55 (0.077)	–	59.9	66.7	53.50	56.4
DL20	93.23 (0.069)	90.34 (0.097)	–	47.90	51.6	57.5	51.90	50.9
XN45	89.92 (0.102)	–	51.7	59.9	99.9	67.3	51.1	50.4
DSM 22670	–	50.3	51.2	46.4	50.3	55.0	48.3	48.9

Most species-specific genes in a pangenome of the *X. griffiniae* clade encode unknown proteins

A pangenome analysis of the clade containing strains XN45, VH1, BG5 and BMMCB, 'the *X. griffiniae*' clade, was conducted. It had 15411 genes that formed 4877 GCs. There were 2364 core GCs, of which 2231 were single-copy. Other groups included 766 strain BMMCB specific, 617 XN45-VH1 species specific, 377 strain BG5 specific, 297 BG5-VH1-XN45 accessory, 154 BMMCB-VH1-XN45 accessory, 150 BG5-BMMCB accessory, 54 strain XN45 specific, 26 strain VH1 specific, 14 XN45-BG5 accessory and five VH1-BMMCB accessory GCs (Fig. 3). From an analysis of gene gain and loss within this clade (Fig. S2), the species-specific genes, 79% of which encoded proteins with unknown functions, possibly resulted from a net acquisition of new genes [55].

To determine which encoded functions were enriched in the core, accessory and strain-specific genomes, single-copy GCs of each were annotated in PROKKA [43]. The functions and biological processes that the GCs encoded were then determined from the UniProt Knowledgebase [44] and COG [45] descriptions. In total, 3352 (68.4%) GCs had known functions. The core genome had the largest percentage of these while strain-specific genomes had the smallest. Specifically, 80% (1735/2182) of the core, 46% (117/252) of BG5-XN45-VH1 accessory, 36% (38/107) of BG5-BMMCB accessory, 29% (31/108) of BMMCB-XN45-VH1 accessory, 23% (75/327) of strain BG5 specific, 23% (154/655) of strain BMMCB specific and 18% (102/561) of *X. griffiniae* species specific GCs had known functions (Figs 3 and S1).

The most enriched functions in the core were those of housekeeping, such as translation, ribosome structure and biogenesis, amino acid transport and metabolism, carbohydrate transport and metabolism, and cell wall/membrane/envelope (Fig. 3). For the last, the Gram-negative nature of the bacteria [30] was demonstrated by the presence of genes encoding lipopolysaccharide (LPS) biosynthesis such as *lpt, rfa, lpx* operons, *lapA-B*, *msbA*, *waaA*, *galU*, *wbgU* and *yhjD. Xenorhabdus* are characterized as motile, peritrichously flagellated rods [30] and this was demonstrated by the presence of genes encoding flagellum biogenesis and motility such as *flgA*, *flgD, flgJ-L, flhA-D, fliC-T* and *fliZ*. Strains XN45, VH1 and BG5 exhibited swarming motility, and this was supported by the presence of the following core genes: *flgB-C*, *flgE-G* and *motA-B*. Other core genes included those encoding antibiotic resistance such as *acrABRZ, mdtABCK macAB fsr bsr* and *lmrA*. The biosynthesis of a few specialized metabolites was also a core function (Fig. 3), and this corroborated a pangenome analysis of genes encoding specialized metabolites from both *Xenorhabdus* and *Photorhabdus* bacteria [28].

X. griffinae genomes (XN45 and VH1) were enriched with a wide selection of genes that encoded carbohydrate metabolism and transport. This was demonstrated by enrichment of genes encoding metabolism and transport of apiose, fuculose, tagatose, galactose and sorbose in the XN45-VH1-BG5 accessory GCs, ribose, galactose, *myo*-inositol, D-malate and galactonate in the XN45-VH1-BMMCB GCs, and glucoside, glycolate and glycerate in the *X. griffiniae* species-specific GCs (Figs 3 and S1).

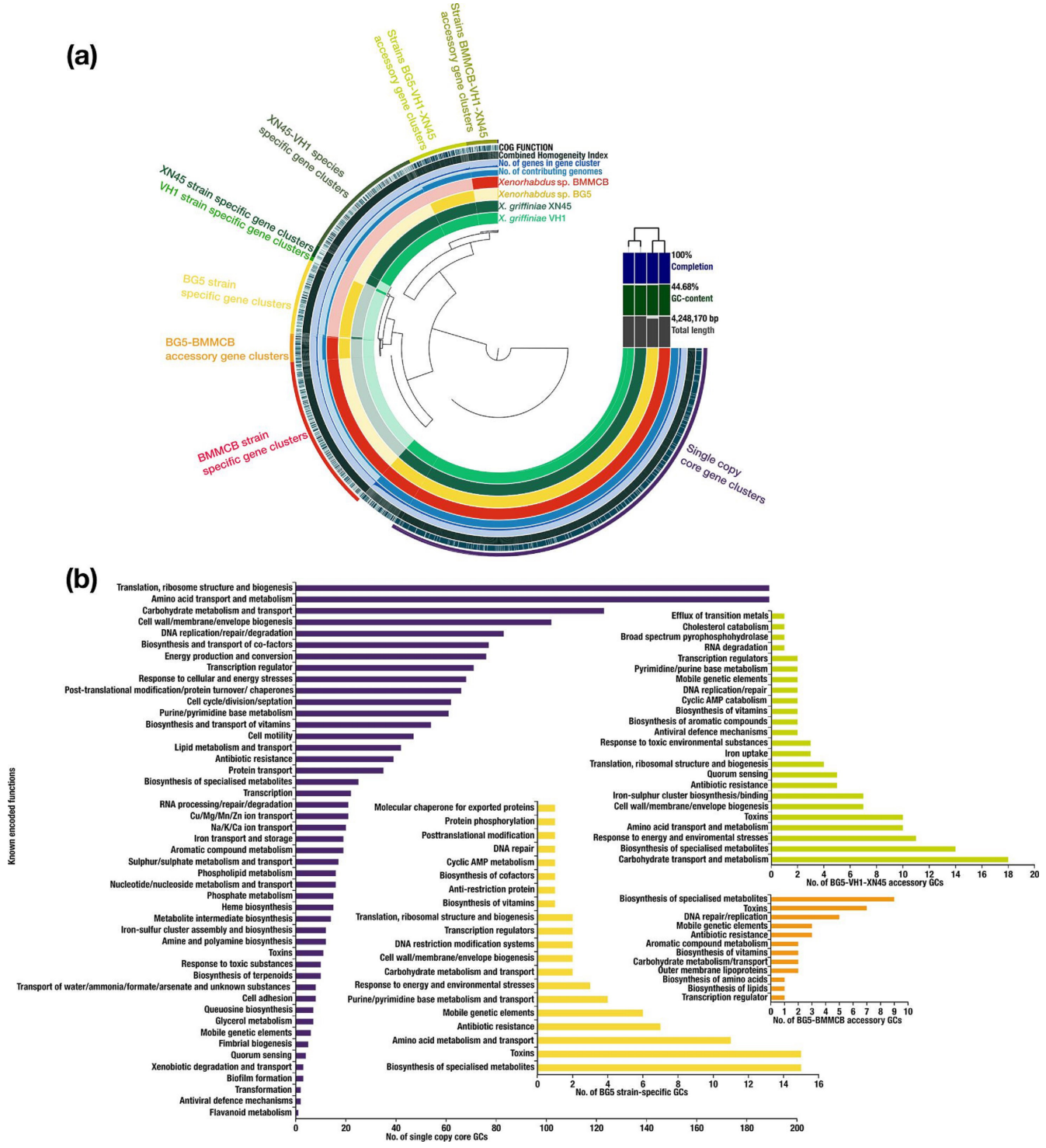

Fig. 3. (a) Graphical representation of a pangenome of a monophyletic group of three *Xenorhabdus* species. Core genes clusters (GCs) were those found in all four genomes, and numbered 2364. Accessory GCs were those found in two or three genomes. BG5-BMMCB, BG5-VH1-XN45 and BMMCB-VH1-XN45 accessory genomes had 150, 297 and 154 GCs respectively. Strain-specific GCs were those found in one genome only, and BMMCB, BG5, XN45 and VH1 had 766, 377, 54 and 26 strain-specific GCs respectively. (b) Bar charts of known functions encoded by *Xenorhabdus* sp. strain BG5 (strain BG5) GCs. Commensurate with their lifestyle as endosymbionts of entomopathogenic nematodes, these bacteria encoded the following non-canonical core functions: antibiotic resistance, and biosynthesis of specialized metabolites and toxins. For XN45-VH1-BG5 accessory GCs, those that encoded the metabolism and transport of carbohydrates such as apiose, fuculose, tagatose, galactose and sorbose, and those that encoded biosynthesis of specialized metabolites such as antibiotics, polyketides, non-ribosomal peptides and siderophores were enriched. For BG5-BMMCB accessory GCs, those that encoded the biosynthesis of specialized metabolites were enriched. For BG5 strain-specific GCs, those that encoded biosynthesis of specialized metabolites and toxins such as type II, III and IV secretion system toxins were enriched among those with known protein functions.

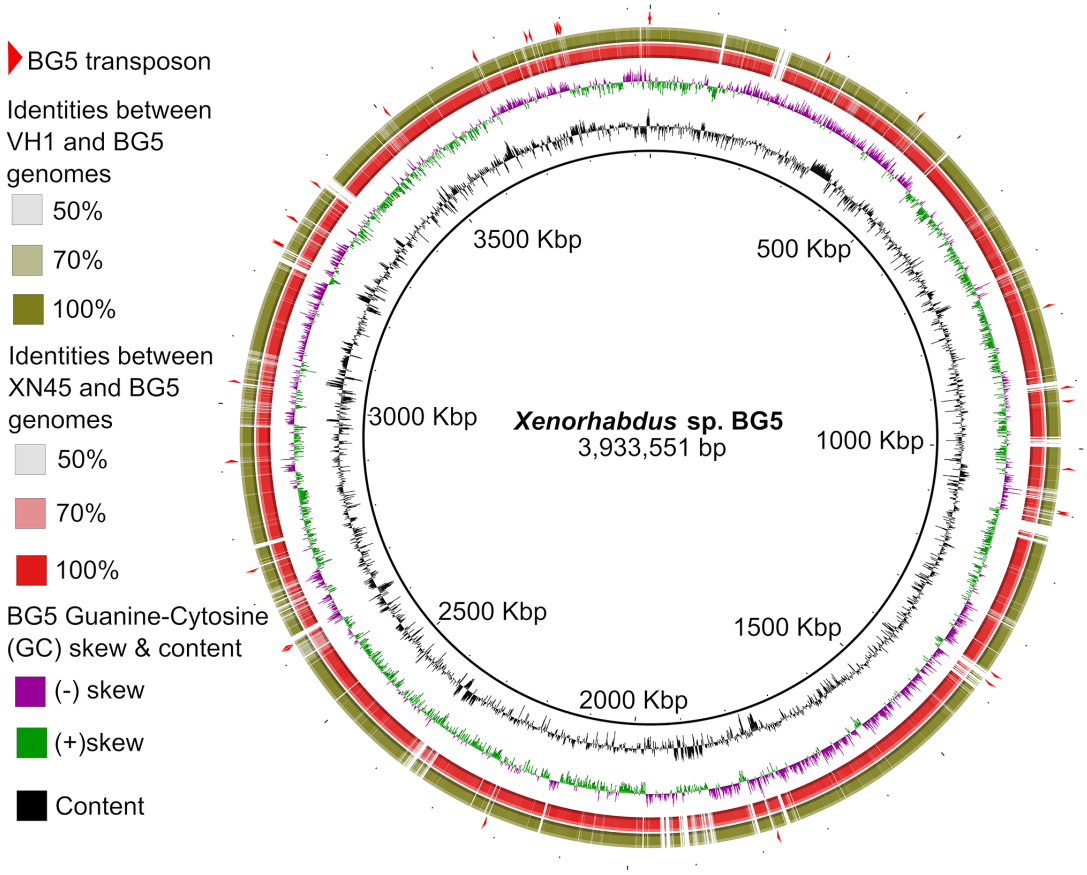

Fig. 4. Genome visualizations of *Xenorhabdus* sp. BG5 when compared to *X. griffiniae* XN45 (red) and *X. griffiniae* VH1 (green). Genomic islands of *Xenorhabdus* sp. BG5 are denoted by white breaks in *X. griffiniae* genomes, and these represented cognate nucleotide sequences that were less than 50% identical. Red triangles in the outermost ring denote loci of IS family transposase genes on the BG5 genome. All transposase genes that were not on contig edges flanked genomic islands. Genomic islands not associated with transposase genes are also shown. The visualization was created in Blast Ring Image Generator (BRIG).

Biosynthesis of specialized metabolites – such as non-ribosomal peptides, polyketides siderophores and antibiotics – was highly species dependent, as demonstrated by its enrichment in species-specific GCs. This was similarly observed with the production of type II, III and IV secretion system toxins. However, these enrichments were from a small fraction (21%) of species-specific GCs; most genes specific to these three *Xenorhabdus* species encoded proteins with unknown functions.

Xenorhabdus sp. BG5 genomic islands are flanked by transposase genes

To investigate whether genes specific to *Xenorhabdus* sp. BG5 formed genomic islands, its genome was compared to those of *X. griffiniae* XN45 and VH1. Regions of less than 50% nucleotide similarity were determined and visualized in BRIG [47]. These were verified as genomic islands by genome alignments in Mauve. Most CDS in these genomic islands encoded hypothetical proteins. Genes encoding transposases in BG5 either flanked genomic islands/were the genomic island (Fig. 4). Insertion sequence (IS) elements predominated the type of transposases predicted to be encoded by these genes, as shown here: IS110 family transposase ISSfl8, IS3 family transposase ISKpn37, IS481 family transposase ISVvu4, IS5 family transposase ISSod6, IS630 family transposase ISPlu10, IS1 family transposase ISEhe5, IS1 family transposase ISPda1, IS110 family transposase ISPlu13, IS3 family transposase ISAlg, IS630 family transposase ISEc40 and IS982 family transposase ISNsp1 (workbook S1). IS elements contribute to genome reshuffling [56] and thus may be implicated in the creation of genomic islands in *Xenorhabdus* sp. BG5.

DISCUSSION

This study aimed to discover new *Xenorhabdus* strains because different species [2, 57] and even strains [3] from this genus have different antimicrobial production profiles. Soils of Western Kenya were selected as they had not hitherto been

investigated for steinernematids, unlike those from Central Kenya [16, 21]. Soils were collected from Bungoma County and sites with the occurrence of steinernematids were clay soils on riverine land, corroborating previous studies on similar soils [58]. From Vihiga County, soils with the occurrence of steinernematids were red volcanic loam soils on cabbage cultivated land, corroborating previous studies on similar lands [21, 59]. From nematodes isolated from these soils, *Xenorhabdus* sp. strains VH1 and BG5 were isolated.

Previously we isolated *X. griffiniae* XN45 from *Steinernema* sp. scarpo which also originated from Kenyan soils. Using draft genome assemblies of strains XN45, VH1 and BG5 (Table 1) which were all of suitable quality for species delineation as per the standards of Chun *et al.* [23], the phylogenomic reconstruction of the genus (Fig. 1) demonstrated that these Kenyan strains formed a monophyletic group that could be enlarged to include strain BMMCB. This strain was previously designated as representing an *X. griffiniae* species. The draft assembly of XN45 was used for species delineation via analysis of orthoANI, DDH and G+C content thresholds for conspecific strains [24, 25, 51]. From these, strain BMMCB was identified as an undescribed species whereas strain VH1 was designated as *X. griffiniae* VH1. Strain BG5 was most closely related to XN45. However, ANI, dDDH and GGD values for the two were not consistent with those of conspecific strains. It was thus designated as an undescribed species of the genus *Xenorhabdus*. This demonstrated the importance of genome assemblies for accurate species delineation via overall genome relatedness indices, corroborating previous pivotal studies [23, 26, 60].

The open pangenome of the genus *Xenorhabdus* corroborated not only a larger pangenome analysis of 40 *Xenorhabdus* strains [54] but also the large number of *Xenorhabdus* strains –hosted by over 50 described *Steinernema* species [19] – that have yet to be isolated, identified and their genome sequences determined.

Using a pangenome analysis of the clade, the four closely related strains XN45, VH1, BG5 and BMMCB were further distinguished as representing three species based on the large numbers of species-specific genes – these were 377, 617 and 766 for *Xenorhabdus* sp. BG5, *X. griffiniae* and *Xenorhabdus* sp. BMMCB respectively. Conversely, strain VH1 was of the same species as XN45 as its genome only had 26 unique genes when compared to that of XN45. Similar pangenome analyses of other clades may elucidate a minimum number of strain-specific genes that delineate species in this genus. Notably, 79% of the functions of proteins encoded by species-specific genes were unknown, compared to 20% for proteins encoded by the core genome. It has long been established that most species-specific prokaryotic genes encode unknown functions [61]. Indeed, in 613 prokaryotic species, over 50 % of a subset of species-specific protein-coding genes encoded unknown functions [62]. These genes lead to speciation when they encode environmentally important traits [61]. For all three species, species-specific genes – only the subset that had known functions – were enriched for the biosynthesis of specialized metabolites. However, this enrichment was probably overestimated because genes encoding secondary metabolites of *Xenorhabdus* species are often long and clustered in genome loci that span thousands of base pairs, which leads to their frequent fragmentation in draft genomes, resulting in inflated counts [63]. *Xenorhabdus* sp. BG5 had genomic islands when its genome was compared to those of *X. griffiniae* strains. Some of these islands were flanked by genes encoding transposases, the vast majority of which were IS elements. IS elements are known to contribute to genome reshuffling [56] suggesting that IS transposases contributed to the creation of genomic islands in *Xenorhabdus* sp. BG5. However, the majority of the islands were not associated with genes encoding transposases, implicating other factors such as phages, as drivers of these differences. In conclusion, two *Xenorhabdus* bacteria isolated from steinernematids from soils in Western Kenya were identified as a novel species and strain. Within the *X. griffiniae* clade, most species-specific genes encoded unknown functions. These genomes, species delineations and genome analyses are useful for *in silico*-based discovery of antimicrobials from the genus *Xenorhabdus*.

Funding information
This study was supported by the Kenya National Research Fund grant NRF first CALL/MULTIDISCIPLINARY RESEARCH/127 'Drug Development of Antibiotics: *Xenorhabdus* bacteria from Kenya' to N.O.A. This research was also supported by the German Academic Exchange Service (DAAD) under programme number 57 299 294 and reference number 91 653 288 and GRADE completion scholarship to R.M.A. Research in the Bode lab was supported by LOEWE Translational Biodiversity Genomics funded by the State of Hesse, Germany. Tim Stinear provided genome sequencing laboratory support for this study.

Author contributions
R.M.A. – conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, validation, visualization, writing-first draft, and writing-editing and reviewing. C.N.W. – methodology, resources, validation, and writing-editing and reviewing. S.J.P. – data curation, formal analysis, investigation, and writing-editing and reviewing. N.O.A. – funding acquisition, methodology, project administration, resources, supervision, and writing-editing and reviewing. H.B.B. – funding acquisition, methodology, project administration, resources, supervision, validation, and writing-editing and reviewing.

Conflicts of interest
The authors declare no competing financial interest. R.M.A. is a proprietor of Elakistos Biosciences.

Ethical statement
No humans or vertebrate organisms were investigated in this study.

References

1. Booysen E, Dicks LMT. Does the future of antibiotics lie in secondary metabolites produced by *Xenorhabdu*s spp.? A review. *Probiotics Antimicrob Proteins* 2020;12:1310–1320.
2. Tobias NJ, Wolff H, Djahanschiri B, Grundmann F, Kronenwerth M, *et al*. Natural product diversity associated with the nematode symbionts *Photorhabdus* and *Xenorhabdus*. *Nat Microbiol* 2017;2:1676–1685.
3. Muangpat P, Suwannaroj M, Yimthin T, Fukruksa C, Sitthisak S, *et al*. Antibacterial activity of *Xenorhabdus* and *Photorhabdus* isolated from entomopathogenic nematodes against antibiotic-resistant bacteria. *PLoS One* 2020;15:e0234129.
4. Akhurst RJ, Boemare NE. A numerical taxonomic study of the genus *Xenorhabdus* (Enterobacteriaceae) and proposed elevation of the subspecies of *X. nematophilus* to species. *J Gen Microbiol* 1988;134:1835–1845.
5. Castaneda-Alvarez C, Prodan S, Zamorano A, San-Blas E, Aballay E. *Xenorhabdus lircayensis* sp. nov., the symbiotic bacterium associated with the entomopathogenic nematode *Steinernema unicornum*. *Int J Syst Evol Microbiol* 2021;71.
6. Ferreira T, van Reenen CA, Endo A, Spröer C, Malan AP, *et al*. Description of *Xenorhabdus khoisanae* sp. nov., the symbiont of the entomopathogenic nematode *Steinernema khoisanae*. *Int J Syst Evol Microbiol* 2013;63:3220–3224.
7. Kämpfer P, Tobias NJ, Ke LP, Bode HB, Glaeser SP. *Xenorhabdus thuongxuanensis* sp. nov. and *Xenorhabdus eapokensis* sp. nov., isolated from *Steinernema* species. *Int J Syst Evol Microbiol* 2017;67:1107–1114.
8. Kuwata R, Qiu L-H, Wang W, Harada Y, Yoshida M, *et al*. *Xenorhabdus ishibashii* sp. nov., isolated from the entomopathogenic nematode *Steinernema aciari*. *Int J Syst Evol Microbiol* 2013;63:1690–1695.
9. Lengyel K, Lang E, Fodor A, Szállás E, Schumann P, *et al*. Description of four novel species of *Xenorhabdus*, family *Enterobacteriaceae*: *Xenorhabdus budapestensis* sp. nov., *Xenorhabdus ehlersii* sp. nov., *Xenorhabdus innexi* sp. nov., and *Xenorhabdus szentirmaii* sp. nov. *Syst Appl Microbiol* 2005;28:115–122.
10. Nishimura Y, Hagiwara A, Suzuki T, Yamanaka S. *Xenorhabdus japonicus* sp. nov. associated with the nematode *Steinernema kushidai*. *World J Microbiol Biotechnol* 1994;10:207–210.
11. Tailliez P*et al*. Phylogeny of *Photorhabdus* and *Xenorhabdus* based on universally conserved protein-coding sequences and implications for the taxonomy of these two genera. proposal of new taxa: *X. vietnamensis* sp. nov., *P. luminescens* subsp. *caribbeanensis* subsp. nov., *P.luminescens* subsp. *hainanensis* subsp. nov., *P. temperata* subsp. *khanii* subsp. nov., *P. temperata* subsp. *tasmaniensis* subsp. nov., and the reclassification of *P. luminescens* subsp. *thracensis* as *P. temperata* subsp. *thracensis* comb. nov. *Int J Syst Evol Microbiol* 2010;60:1921–1937.
12. Tailliez P, Pagès S, Edgington S, Tymo LM, Buddie AG. Description of *Xenorhabdus magdalenensis* sp. nov., the symbiotic bacterium associated with *Steinernema australe*. *Int J Syst Evol Microbiol* 2012;62:1761–1765.
13. Tailliez P, Pagès S, Ginibre N, Boemare N. New insight into diversity in the genus *Xenorhabdus*, including the description of ten novel species. *Int J Syst Evol Microbiol* 2006;56:2805–2818.
14. Heryanto C, Eleftherianos I. Nematode endosymbiont competition: fortune favors the fittest. *Mol Biochem Parasitol* 2020;238:111298.
15. Vigneux F, Zumbihl R, Jubelin G, Ribeiro C, Poncet J, *et al*. The xaxAB genes encoding a new apoptotic toxin from the insect pathogen *Xenorhabdus nematophila* are present in plant and human pathogens. *J Biol Chem* 2007;282:9571–9580.
16. Waturu CN, Hunt DJ, Reid AP. *Steinernema karii* sp. n.(*Nematoda*: *Steinernematidae*), a new entomopathogenic nematode from Kenya. *Int J Nematol* 1997;7:68–75.
17. White GF. A method for obtaining infective nematode larvae from cultures. *Science* 1927;66:302–303.
18. Awori RM. Nematophilic bacteria associated with entomopathogenic nematodes and drug development of their biomolecules. *Front Microbiol* 2022;13:993688.
19. Bhat AH, Chaubey AK, Askary TH. Global distribution of entomopathogenic nematodes, *Steinernema* and *Heterorhabditis*. *Egypt J Biol Pest Control* 2020;30:31.
20. Awori RM, Ng'ang'a PN, Nyongesa LN, Amugune NO, Masiga D. Mursamacin: a novel class of antibiotics from soil-dwelling roundworms of Central Kenya that inhibits methicillin-resistant *Staphylococcus aureus*. *F1000Res* 2017;5:2431.
21. Mwaniki SW, Nderitu JH, Olubayo F, Kimenju JW, Nguyen K. Factors influencing the occurrence of entomopathogenic nematodes in the Central Rift Valley Region of Kenya. *African J Ecol* 2008;46:79–84.
22. Spiridonov SE, Reid AP, Podrucka K, Subbotin SA, Moens M. Phylogenetic relationships within the genus *Steinernema* (Nematoda: Rhabditida) as inferred from analyses of sequences of the ITS1-5.8S-ITS2 region of rDNA and morphological features. *Nematol* 2004;6:547–566.
23. Chun J, Oren A, Ventosa A, Christensen H, Arahal DR, *et al*. Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int J Syst Evol Microbiol* 2018;68:461–466.
24. Lee I, Ouk Kim Y, Park S-C, Chun J. OrthoANI: an improved algorithm and software for calculating average nucleotide identity. *Int J Syst Evol Microbiol* 2016;66:1100–1103.
25. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, *et al*. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* 2007;57:81–91.
26. Meier-Kolthoff JP, Göker M. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. *Nat Commun* 2019;10:2182.
27. Medini D, Donati C, Rappuoli R, Tettelin H. The pangenome: a data-driven discovery in biology. In: *The Pangenome*. Cham: Springer, 2020. pp. 3–20.
28. Shi Y-M, Hirschmann M, Shi Y-N, Ahmed S, Abebew D, *et al*. Global analysis of biosynthetic gene clusters reveals conserved and unique natural products in entomopathogenic nematode-symbiotic bacteria. *Nat Chem* 2022;14:701–712.
29. Ngugi CN. Characterization and Evaluation of Entomopathogenic Nematodes for the Management of Tomato Leafminer (Tuta absoluta Meyrick). PhD Dissertation. Nairobi, Kenya: Faculty of Science and Technology, University of Nairobi, 2021.
30. Boemare N, Akhurst R, Stackebrandt E. The genera *Photorhabdus* and *Xenorhabdus*. In: Dworkin M, Falkow S, Rosenberg E and Schleifer KH (eds). *The Prokaryotes: Volume 6: Proteobacteria: Gamma Subclass*. New York, NY: Springer New York; 2006. pp. 451–494.
31. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, *et al*. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.
32. Mothupi B, Featherston J, Gray V. Draft whole-genome sequence and annotation of *Xenorhabdus griffiniae* strain BMMCB associated with the South African entomopathogenic nematode *Steinernema khoisanae* strain BMMCB. *Genome Announc* 2015;3:e00785-15.
33. Wattam AR, Brettin T, Davis JJ, Gerdes S, Kenyon R, *et al*. Assembly, annotation, and comparative genomics in PATRIC, the all bacterial bioinformatics resource center. *Methods Mol Biol* 2018;1704:79–101.
34. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, *et al*. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* 2016;17:132.
35. Meier-Kolthoff JP, Auch AF, Klenk H-P, Göker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 2013;14:60.
36. Lefort V, Desper R, Gascuel O. FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol Biol Evol* 2015;32:2798–2800.
37. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 2019;47:W256–W259.
38. Bah T. *Inkscape: Guide to a Vector Drawing Program*. Prentice Hall Press, 2007.

39. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, *et al*. Anvi'o: an advanced analysis and visualization platform for 'omics' data. *PeerJ* 2015;3:e1319.
40. Pantoja Y, Pinheiro K, Veras A, Araújo F, Lopes de Sousa A, *et al*. PanWeb: a web interface for pan-genomic analysis. *PLoS One* 2017;12:e0178154.
41. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, *et al*. RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci Rep* 2015;5:8365.
42. Csurös M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* 2010;26:1910–1912.
43. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.
44. UniProt Consortium. The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res* 2009;37:D169–74.
45. Galperin MY, Kristensen DM, Makarova KS, Wolf YI, Koonin EV. Microbial genome analysis: the COG approach. *Brief Bioinform* 2019;20:1063–1070.
46. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, *et al*. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res* 2019;47:W81–W87.
47. Alikhan N-F, Petty NK, Ben Zakour NL, Beatson SA. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 2011;12:402.
48. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–3402.
49. Mothupi B. Genomics of entomopathogenic bacterial endosymbiont species associated with desiccation tolerant entomopathogenic nematode. South Africa: MSc Dissertation, Faculty of Science, University of the Witwatersrand, Johannesburg, 2016.
50. Kim M, Oh H-S, Park S-C, Chun J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int J Syst Evol Microbiol* 2014;64:346–351.
51. Meier-Kolthoff JP, Klenk HP, Göker M. Taxonomic use of DNA G+C content and DNA-DNA hybridization in the genomic age. *Int J Syst Evol Microbiol* 2014;64:352–356.
52. Delmont TO, Eren AM. Linking pangenomes and metagenomes: the *Prochlorococcus* metapangenome. *PeerJ* 2018;6:e4320.
53. Heaps HS. *Information Retrieval, Computational and Theoretical Aspects*. Academic Press, 1978.
54. Rivera-Ramírez A, Salgado-Morales R, Jiménez-Pérez A, Pérez-Martínez R, García-Gómez BI, *et al*. Comparative genomics and pathogenicity analysis of two bacterial symbionts of entomopathogenic nematodes: the role of the GroEL protein in virulence. *Microorganisms* 2022;10:486.
55. Iranzo J, Wolf YI, Koonin EV, Sela I. Gene gain and loss push prokaryotes beyond the homologous recombination barrier and accelerate genome sequence divergence. *Nat Commun* 2019;10:5376.
56. Siguier P, Gourbeyre E, Chandler M. Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiol Rev* 2014;38:865–891.
57. Fodor A, Fodor AM, Forst S. Comparative analysis of antibacterial activities of xenorhabdus species on related and non-related bacteria in vivo. *J Microbiol Antimicrob* 2010;2:36–46.
58. Brodie G. Natural occurrence and distribution of entomopathogenic nematodes (*Steinernematidae, Heterorhabditidae*) in Viti Levu, Fiji Islands. *J Nematol* 2020;52:1–17.
59. Zepeda-Jazo I, Molina-Ochoa J, Lezama-Gutiérrez R, Skoda SR, Foster JE. Survey of entomopathogenic nematodes from the families *Steinernematidae and Heterorhabditidae* (*Nematoda*: *Rhabditida*) in Colima, México. *Int J Trop Insect Sci* 2014;34:53–57.
60. Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, *et al*. Microbial species delineation using whole genome sequences. *Nucleic Acids Res* 2015;43:6761–6771.
61. Konstantinidis KT, Tiedje JM. Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci* 2005;102:2567–2572.
62. Koressaar T, Remm M. Characterization of species-specific repeats in 613 prokaryotic species. *DNA Res* 2012;19:219–230.
63. Klassen JL, Currie CR. Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation. *BMC Genomics* 2012;13:14.

Peer review history

VERSION 3

Editor recommendation and comments

https://doi.org/10.1099/acmi.0.000531.v3.1

Rubén de Dios; Brunel University London, Life Sciences, UNITED KINGDOM

Date report received: 27 January 2023
Recommendation: Accept

Comments: In this manuscript version, Awori et al. have applied the minor changes suggested for acceptance. I would like to thank the authors for minding he comments of the reviewers and myself. Congratulations!

Author response to reviewers to Version 2

Response to Editor

Note: The Editor's comments are in boldface.

After comment 4 of Reviewer 2, the authors corrected the terminology from Xenorhabdus sp. nov. BG5 to Xenorhabdus sp. BG5. However, there is still one reference to Xenorhabdus sp. nov. BG5 in the Figure 4 legend (L 459-464) that the authors might have missed. Please consider a correction.

Dear Editor,

Sorry for this oversight. This has been corrected (L 459-464).

Figure 4. Genome visualisations of *Xenorhabdus*sp. BG5 when compared to *X. griffiniae*XN45 (red) and *X. griffiniae*VH1 (green). Genomic islands of *Xenorhabdus*sp. BG5 are denoted by white breaks in *X. griffiniae*genomes, and these represented cognate nucleotide sequences that were less than 50% identical. Red triangles in the outermost ring denote loci of IS family transposase genes on the BG5 genome. All transposase genes that were not on contig edges flanked genomic islands. Genomic islands not associated with transposase genes are also shown. The visualisation was created in Blast Ring Image Generator (BRIG).

In the text modification introduced after comment 19 of Reviewer 2, the authors mention that this phenomenon of inflated counts due to enrichment in secondary metabolite related genes at the boundaries of contigs is frequent. It would be ideal if you could support this statement with at least one reference.

Dear Editor, this has been corrected (L 507-511)

However, this enrichment was likely overestimated because genes encoding secondary metabolites of *Xenorhabdus*species are often long and clustered in genome loci that span thousands of base pairs, which leads to their frequent fragmentation in draft genomes, resulting in inflated counts (**Klassen & Currie, 2012)**.

VERSION 2

Editor recommendation and comments

https://doi.org/10.1099/acmi.0.000531.v2.3

Rubén de Dios; Brunel University London, Life Sciences, UNITED KINGDOM

Date report received: 25 January 2023
Recommendation: Minor Amendment

Comments: In this revised manuscript, Awori et al. present the analysis of various Xenorhabdus genomes introducing the modifications suggested by the reviewers. After these changes, the manuscript has improved its quality and is now in the position to be accepted for publication. Congratulations to the authors for this piece of work. Nevertheless, there are two minor aspects that the authors might have overlooked during the corrections: • After comment 4 of Reviewer 2, the authors corrected the terminology from Xenorhabdus sp. nov. BG5 to Xenorhabdus sp. BG5. However, there is still one reference to Xenorhabdus sp. nov. BG5 in the Figure 4 legend (L 459-464) that the authors might have missed. Please consider a correction. • In the text modification introduced after comment 19 of Reviewer 2, the authors mention that this phenomenon of inflated counts due to enrichment in secondary metabolite related genes at the boundaries of contigs is frequent. It would be ideal if you could support this statement with at least one reference. Please consider these minor issues in a second revised version before progressing the manuscript to accepted for publication. Again, I would like to congratulate the authors for this final piece of work.

SciScore report

https://doi.org/10.1099/acmi.0.000531.v2.1

iThenticate report

https://doi.org/10.1099/acmi.0.000531.v2.2

Author response to reviewers to Version 1

Note: The reviewers' comments are in boldface.

Response to reviewer 1

Methodological rigour, reproducibility and availability of underlying data

1. In the section of materials and methods "Isolation of nematodes from soils samples", the authors described the use of Galleria mellonella larvae as bait. They should indicate from where they got the Galleria larvae and how they can ensure that the isolated nematodes were not already infecting the larvae.

Thank you for your correction. The following change was made (144-147):

"To bait EPNs from the soil, *Galleria mellonella*larvae were first obtained from a laboratory insect culture of KALRO-Horticulture Research Institute, where they had been reared as previously described (Ngugi, 2021). From this culture, any healthy larvae were selected."

2. Any Galleria larvae was used or the author selected them within a range of weight. This should be specified in the text.

Thank you for your correction. The following change was made (144-147):

"To bait EPNs from the soil, *Galleria mellonella*larvae were first obtained from a laboratory insect culture of KALRO-Horticulture Research Institute, where they had been reared as previously described (Ngugi, 2021). From this culture, any healthy larvae were selected."

3. In the section of materials and methods "Isolation of bacteria from nematodes", there are not any reference. How were Xenorhabdus sp. isolated previously? The author should clarify that in this section.

Thank you for your correction. The following change was made (lines 168-169):

"The indirect isolation of *Xenorhabdus*bacteria from haemolymph was based on a previously described method(Boemare & Akhurst, 2006) with modifications (Awori *et al.,*2017)."

Presentation of results

4. There are 27 genomes in the Table S2, which is cited in the text in the section "Creation of pangenomes". However, only 26 genomes (25 species) are used in the construction of the pangenome (Figure 2).

Thank you for your correction.

The pangenome analysis has been redone to include the genome of *X. lircayensis*(lines 357-362)

Figure 2. Graphical representation of the pangenome of 26 species of the *Xenorhabdus*genus. The largest genome was 5,347,057 bp and the highest guanine-cytosine (GC) content was 45.7%. The pangenome was composed of a total of 13,469 gene clusters (GCs). Core GCs, those found in all 27 genomes, were 1654 in total. Accessory GCs, those found in two to twenty-six genomes, were 5992 in total. Strain-specific GCs, those found in one genome only, were 5820 GCs in total.

5. In the description of Table S2 is written "…genomes used in phylogenomic and pangenomic analyses". Therefore, Table S2 should be also cited in section "Phylogenomic reconstruction and calculation of ANI values".

Thank you very much for your correction. The following change was made (lines 204-206):

"For the *Xenorhabdus*genus phylogenomic reconstruction, twenty-seven fasta files (Supplementary Table S3) were used as input data for a whole genome-based taxonomic analysis on the Type strain genome server platform."

In this section, is stated that "26 fasta files were used as input…" (Line 215), however, in the Figure 1 there are 27 genomes.

Thank you very much for your correction. The following change was made (lines 204-206):

"For the *Xenorhabdus*genus phylogenomic reconstruction, twenty-seven fasta files (Supplementary Table S3) were used as input data for a whole genome-based taxonomic analysis on the Type strain genome server platform."

6. Please, clarify the number of genomes used for each section and why X. lircayensis is included in the phylogenomic analyses but not in the pangenome.

Thank you for your question. When we did our pangenome analysis, *X. lircayensis*had not yet been published. After it was published, we were only able to redo the phylogenomic but not pangenomic analysis. However, we have now succeeded in redoing the pangenomic analysis to include *X. lircayensis*. The manuscript has been changed accordingly. Now, the same number of genomes were used for both analyses.

Any other relevant comments

7. It is difficult to read the page 5 (Introduction, from line 103 to 123) where multiple Steinernema sp. and Xenothabdus isolates are named. I suggest to the authors to include this part in results (supplementary table) and add the Steinernema sp., the corresponding Xenorhabdus isolate, if any, the location, and the reference.

This has been edited in lines 102-104 as follows:

"This research gap between steinernematid isolation and *Xenorhabdus*endosymbiont identification is also seen in Sub-Saharan Africa, where numerous steinernematids have been isolated from this region (refer to Supplementary Table S1 for a full list of species and location).The new table is found in lines 45-46 of the supplementary section."

Supplementary Table S1.*Steinernema*isolates from locations in Sub-Saharan Africa.

Isolate	Location	*Xenorhabdus*symbiont	Reference
Steinernema pwaniensis	Mwasembe, Tanzania	No published isolate	Půža *et al.*,2017
*Steinernema*sp. Thui	Akoutaossé, Benin	*X. indica*	Godjo *et al.*,2018
S. cameroonense	Obala, Cameroon	*Xenorhabdus* sp. A71	Kanga *et al.*,2012; Kanga *et al.*,2014
S. nyetense	Nyété, Cameroon	No published isolate	Kanga *et al.*,2012
S. ethiopiense	*Mendi area, Ethiopia*	No published isolate	*Tamiru et al., 2012*
S. yirgalemense	*Yirgalem, Ethiopia*	*X. indica*	Ferreira *et al.*,2016;*Nguyen et al., 2004*
S. nguyeni	Clanwilliam S. Africa	*X. bovienii*	Antoinette P. Malan *et al.*,2016; Dreyer *et al.*,2017
S. tophus	Clanwilliam S. Africa	No published isolate	Cimen *et al.*,2014
S. sacchari	Gingindlovu, S. Africa	*X. khoisanae*	Dreyer *et al.*,2017;Nthenga *et al.*,2014
S. fabii	Piet Retief, S. Africa	No published isolate	Abate *et al.*,2016

S. bertusi	Port Edward, S. Africa	No published isolate	Katumanyane *et al.*,2020
S. citrae	Piketberg, S. Africa	No published isolate	Nomakholwa *et al.*,2011
S. innovationi	Free State, S. Africa	No published isolate	Çimen *et al.*,2015
S. jeffreyense	Jefferson Bay, S. Africa	*X. khoisanae*	A.P. Malan *et al.*,2016; Dreyer *et al.*,2017
S. beitlechemi	Bethlehem,S. Africa	*X. khoisanae*	Cimen *et al.*,2016
S. khoisanae	Villiersdorp, S. Africa	*X. khoisanae*	Ferreira *et al.*,2013;Malan *et al.*,2006
Steinernema sp. WS9	Fridenheim, S. Africa	*X. griffiniae*	Jonike Dreyer *et al.*,2018
S. australe TEL	Walkerville, South Africa	No published isolate	Lephoto & Gray 2019
Steinernema sp. HBG28	Guateng, S. Africa	*X. khoisanae*	Naidoo *et al.*,2015
*Steinernema*sp. LAOS	S. Africa	*Xenorhabdus* sp. strain GDc328	Soobramoney *et al.*,2015
Steinernema sp. BMMCB	Brits, S. Africa	*Xenorhabdus* sp. BMMCB	Mothupi *et al.*,2015

8. Line 37 in supplementary Table S2: (Typo) Replace phlyogenomic by phylogenomic.

Thank you. This was corrected (Lines 49-50).

Supplementary Table S3. Accessions numbers, strains and predicted gene counts of genomes used in phylogenomic and pangenomic analyses.

Response to reviewer 2

1) Several times, the authors describe genes that are unique to each lineage as "causing speciation" (e.g., L. 49, L. 378, L. 507). It is true that gene gain and loss can underpin the creation of species boundaries, although how this works in bacteria is still subject to debate. What is clear is that not all differences in gene content or variation are causal in the speciation process, e.g., much likely accumulates via drift or is genetically linked to causal variants. Because this study only identifies strain-specific variation and does not directly test whether that variation causes speciation, such causal language should be removed throughout the manuscript.

Thank you for your correction.

Such casual language has been removed from the manuscript.

(2) The authors repeatedly imply that each Steinernema hosts a unique species of bacterial symbiont (e.g., LL. 79-83; LL. 99-100). However, this is untrue, e.g., the association with X. bovienii with multiple species of Steinernema hosts has been studied in some detail - Murfin et al. 2015 mBio 6:e00076-15, Murfin et al. 2015 BMC Genomics 16:889. That said, those studies did show that there was co-adaptation of X. bovienii strains with the Steinernema hosts from which they had been isolated. Thus, unique variation also exists in Xenorhabdus below the level of species in this case. These nuances should be better reflected in the manuscript.

Thank you for your correction. We poorly communicated our intended message.

The intended implication was not that each *Steinernema*species hosts a unique bacterial symbiont; the reference to host switching and demonstration that *X. khoisane*is hosted by several different *Steinernema*species in the introduction dispels this notion. The intended implication was that each *Steinernema*species naturally only associates with one *Xenorhabdus*species, but the reverse is not true. In fact, Murfin et al. 2015 mBio 6:e00076-15 stated this in their introduction.

"*Xenorhabdus bovienii*bacterial strains are broad-host-range symbionts that associate with at least nine *Steinernema*nematode species from two phylogenetic subclades.Conversely, each of the(nine different)nematode host species harbors only *X. bovienii*."

Thus, if one has 50 *Steinernema*species, one can expect to isolate at most 50 *Xenorhabdus*species from them. This number will most likely be lower because some *Xenorhabdus*species will associate with more than one *Steinernema*species.

To correct our poor communication, we have deleted associated sections and clarified our message inLL. 96-101 as follows.

"For species such as *Xenorhabdus khoisanae*(Supplementary Table S1), *X. bovienii*, *X. kozodoii*, *X. poinarii,*and *X. hominckii,*we see one *Xenorhabdus*species as the natural symbiont of numerous *Steinernema*species (Awori,2022). However, the reverse, one *Steinernema*species that naturally hosts, with equal fitness, two different *Xenorhabdus*species is yet to be discovered. Thus, there is a high possibility of identifying new *Xenorhabdus*species from the over fifty described *Steinernema*species whose symbionts remain uncharacterised (Bhat *et al.*,2020)."

(3) It would be useful to know more about the Steinernema nematodes from which the strains in this study were isolated, if that information is available. Given the introductory material regarding co-adaptation between distinct bacteria-host pairs, it would be especially useful to know if the sampled nematodes represented unique species, especially the hosts for the two X. griffiniae strains.

More information on the *Steinernema*nematodes is currently not available. However, one ofthecurrent proposed projects of one of the co-authors entails determining the ITS sequences of these two *Steinernema*isolates. Should it be successful,he plans to avail this information to the scientific community.

4) I recommend that the authors moderate the language about defining new species somewhat (e.g., LL. 334-336), given that their analysis is not sufficient to describe new species following the International Code of Nomenclature of Prokaryotes, which would require publication of novel species names in the International Journal of Systematic and Evolutionary Microbiology. I agree that the author's genomic analyses are consistent with BG5 and BMMCB representing currently unnamed species within the genus Xenorhabdus, but this is distinct from and more general than the phrasing used here.

Thank you very much for the correction. Lines 346-348 were changed as shown below.Moreover, throughout the manuscript, *Xenorhabdus*sp. nov. BG5 and *Xenorhabdus*sp. nov. BMMCB were changed to *Xenorhabdus*sp. BG5 and *Xenorhabdus*sp. BMMCB respectively.

"Taken together, these results demonstrated that these four strains were three species: *X. griffiniae*XN45, *X. griffiniae*VH1, and the two undescribed species *Xenorhabdus*sp. BG5, and *Xenorhabdus*sp. BMMCB."

5) Given that the genomes used in the pangenome analysis are draft quality, meaning that they lack certain genes due to assembly artifacts, it seems to me that using 100% presence in all analyzed genomes to define the set of "core" genes is too strict a threshold, even if this is the default used by anvi'o. Instead, I suggest defining the core by using the dendrogram at the center of Figure 2 that clusters gene families by the conservation between the analyzed genomes, which seems to have two main branches that seem to separate conserved and variable genes, even if they are not 100% conserved. Finer nodes might alternatively used, but regardless the issue of how genome quality affects genome content analyses should be explicitly addressed. Similar caveats apply to Figure 3, although this issue seems more difficult to mitigate the with fewer genomes presented here.

Thank you for your correction.

This has been addressed by estimating how much the use of draft genomes as opposed to complete genomes affects the size of the core genome. The following has been included in the manuscript.

(Lines 235-244)

Estimation of the effect of draft genomes on the determination of the core genome

To estimate how the use of draft genomes affects the determination of the core genome, two additional pangenomes were created. The first contained six genomes, each of which was composed of less than two contigs: *X. bovienii*CS-03(NZ_FO818637), *X. hominickii*ANU (NZ_CP016176), *X. cabanillasii*DSMZ 19705 (NZ_QTUB01000001), *X. poinarii*G6 (FO704551), *X. nematophila*AN6/1 (FN667742), and *X. szentirmaii*US123(NIUA01000001). The second contained draft genomes of similar species: *X. bovienii*T228 (JANAIF000000000.1), *X. hominickii*DSM 17903 (NJAI00000000.1), *X. cabanillasii*JM26 (NJGH00000000.1), *X. poinarii*SK (JADLIG000000000.1), and *X. nematophila*C2-3 (JRJV00000000.1). Pangenomes were created via the aforementioned anvio workflow and the size of their resultant core genomes were compared.

Lines 39-44 Supplementary section

Supplementary Figure S3. Comparison of core genomes from two pangenomes of similar *Xenorhabdus*species created from draft and complete (composed of less than two contigs) genomes. For the pangenome made from A) draft genomes, the number of core, and single copy core GCs were 1818 and 1365 respectively. For the pangenome made from B) complete genomes, the number of core and single copy core gene clusters (GCs) were 1917 and 1728 respectively. Thus, use of draft genomes underestimated the number of core and single copy core GCs by 5% and 21% respectively.

6) LL. 358-359: Because two X. griffiniae strains were used in the analysis vs. only a single strain of all other species, it is inappropriate to directly compare the unique genes in each strain to those in the other species. Instead, it would be stronger to compare the genes that are both unique to each X. griffiniae plus those that are shared between the two X. griffiniae strains but absent from all other sampled Xenorhabdus species.

Thank you for your correction. This has been changed, as shown below (lines 373-375).

*X. mauleonii*had the largest number of strain-specific genes (500) whereas *X. ehlersii*had the least (118). *X. griffiniae*VH1 and XN45 had remarkably few strain specific GCs, 13 and 17 respectively, as they were the only two strains from the same species (Figure 2).

(7) The analysis suggesting that gene gain or loss arose to differences in gene content between the four focal strains analyzed here (LL. 380-385) is incomplete and needs to be modified because it does not include a reconstruction of gene content of the common ancestor of the entire BMMCB/BG5/XN45/VH1 clade. This would require discussing the gene content of the phylogenetic neighbors of this clade, i.e., X. bozodoii and the ancestor of X. thuongxuanensis/X. ehlersii/X. ishibashii/X. eapokensis. Methods exist to do this (e.g., ANGST; David and Alm 2011 Nature 469:93) but these were not applied here.

Thank you for your correction. This has been addressed by including an analysis gene gain and loss as shown below (lines 246-249).

Analysis of gain and loss of gene clusters in the *X. griffiniae*clade

Using the GCs of the VH1-BG5-XN45-BMMCB pangenome, a matrix of the presence and absence of GCs among the four strains was created(Supplementary workbook 1). This matrix and the GDBP phylogeny of the four strains were then used as input data for gene gain and loss analysis in the COUNT program (downloaded 17/01/2023) using Wagner parsimony (penalty=1) (Csűös, 2010).

Lines 35-38Supplementary section

Supplementary Figure S2. Evolution of the number of gene clusters in the *Xenorhabdus griffiniae*clade. GC denotes total gene clusters present in either an extant or extinct genome. Green and red triangles denote gene gains and losses respectively.

LL. 393-396 of the main section

"From an analysis of gene gain and loss within this clade (Figure S2), the species specific genes, seventy-nine percent of which encoded proteins with unknown functions, possibly resulted from a net acquisition of new genes (Iranzo *et al.*,2019)."

(8)I must admit that I have difficulty reconciling the authors' transposon analysis with the BRIG analysis in Figure 4. As I understand it, genomic islands in BG5 are indicated by gaps in the outer two rings, i.e., these genes are lacking in XN45 and VH1. Given the statements about transposons flanking gene islands, I expected the arrows representing transposons to align with those gaps, but they do not consistently do so, i.e., there are many gaps without transposons associated with them and places where the transposons do not obviously align with gaps. Is this because the gaps are relatively small? If most of the gaps are, in fact, not associated with transposons, then I think that this needs to be more explicit in the text, because it implies that transposons are only associated with a minority of genome differences and therefore not the main driver of these differences.

"It implies that transposons are only associated with a minority of genome differences and therefore not the main driver of these differences." This is the intended implication.

Hence why we did not state them as the main drivers and as only contributing to genome reshuffling. From results of an ongoing unpublished study of one of the authors, we see a major driver of such differences are phage-related genes. However, this is beyond the scope of this manuscript. Taking these insights together, the following statement was added to the discussion section (lines 514-515).

"However, majority of the islands were not associated with genes encoding transposases, implicating other factors such as phages, as drivers of these differences."

Minor comments:

(9) Permission for collecting the samples used in this study or that such permission is not necessary must be indicated.

Thank you for the correction. The following was included in the manuscript(lines 127-130).

"No access permits were required as per the exceptions of section 3 (d) of the Environmental Management and Coordination (Conservation of Biological Diversity and Resources, Access to Genetic Resources and Benefit Sharing) Regulations 2006 of the Environmental Management and Coordination Act, 1999 of the Laws of Kenya."

10) Table 1 caption: the genome quality statistics listed in this table are said to be derived from PGAP, but I am unaware of this pipeline performing such an analysis. The methods state that PATRIC was used for this analysis instead, which seems the more likely citation here.

Yes, it was PATRIC. Thanks for the correction.

(11) LL. 298-300: The terms "monophyletic" and "paraphyletic" are used incorrectly here. "Paraphyletic" refers to the situation where members of two different taxa are interdigitated amongst each other within the same phylogenetic clade. This is not the case here, where BG5 and BMMCB, which are both clearly not X. griffiniae, are clear outgroups to XN45 and VH1, which are; thus, X. griffiniae is monophyletic. All four strains do form a single clade that contains three putative species (contra L. 299 - "BG5 did not form a clade") that would be paraphyletic if BG5 was classified as a different species but BMMCB retained its X. griffiniae classification. However, this paper clearly removes that former classification.

Thank you for your correction. The terms have been deleted.

(12) Table 2 caption: I suggest adding "of the matrix" to "the top half" and "the bottom half", it took me a while to understand what exactly was going on here. Also, are the ANI and dDDH values given averages of the bidirectional tests? These values sometimes vary slightly based on which genome is used as a query, depending on the implementation.

Thank you for your correction. This was changed, as shown below(320-324):

"Table 2. Orthologous average nucleotide identities (orthoANI), genome-to-genome distances (GGD), and digital DNA-DNA hybridization (in %) (dDDH) values for type species most closely related to*Xenorhabdus*sp. strains VH1& BG5, *X. griffiniae*XN45, and *Xenorhabdus*sp. strain BMMCB. OrthoANI values are in the top half of the matrix (top triangle), GGD in brackets, and values for dDDH are in the bottom half of the matrix (bottom triangle).Values that are within the threshold for two strains to be classified as one species are shaded in grey."

(13) LL. 338-339: "...a pangenome that lacked many strain-specific genes" - I think that "lacked" should actually be "included" here, especially given that the following text describes those genes.

Yes. This was corrected in lines 350-351.

"We hypothesized that the *Xenorhabdus*genus had a pangenome that included many strain-specific genes, due to the numerous *Xenorhabdus*strains and their respective genomes…"

(14) As a suggestion, the authors might consider reordering the strains in Figure 2. At the very least, BMMCB and BG5 should be next to XN45 and VH1 because this is the central comparison made in this study. Using an ordering that matches Figure 1 might also make comparisons more logical (i.e., ordering by phylogeny instead of alphabetically).

Thank you for your correction.

They have been ordered according to phylogeny and BMMCB and BG5 are now flanking VH1 & XN45 (lines 357-362).

Figure 2. Graphical representation of the pangenome of 26 species of the *Xenorhabdus*genus. The largest genome was 5,347,057 bp and the highest guanine-cytosine (GC) content was 45.7%. The pangenome was composed of a total of 13,469 gene clusters(GCs). Core GCs, those found in all 27 genomes, were 1654 in total. Accessory GCs, those found in two to twenty-six genomes, were 5992 in total. Strain-specific GCs, those found in one genome only, were 5820 GCs in total.

(15) Similarly, I also suggest using the same general order (i.e., from largest to smallest or vice versa) in all of the bar charts in Figure 3b and Supplementary Figure S1 so that they are more directly comparable to each other. Otherwise it gives the impression of differences (largest bars at the top vs. at the bottom) that do not actually exist in the data.

Thank you for your suggestion. The current order of the bar charts was chosen as it enabled the placement of the various figures without making the final image looked cramped. Different placements were tried, including your suggestion, and this was the most fitting.

(16) L. 466 and L. 470: replace "nascent" with "draft"

Thank you for your correction. This was rectified.

(17) L. 474: replace "failed to pass conspecific thresholds" with more specific language that describes the actual result, i.e., that the dDDH, ANI, and GGD values were not consistent with these strains belonging to the same species.

Thank you for your correction. This was rectified as shown below (lines 485-6).

"However, ANI, dDDH and GGD values for the two were not consistent with those of conspecific strains. It was thus designated as an undescribed species of the genus *Xenorhabdus*."

18) L. 474: the "sp. nov." designation should only be used when describing a new binomial name, and thus is inappropriate here. It indicates the first use of a species name, not that the strains being analyzed represent a new species that has yet to be formally described with such a name.

Thank you for your correction. Throughout the manuscript, the terms*Xenorhabdus*sp. nov. BG5 and *Xenorhabdus*sp. nov. BMMCB were changed to *Xenorhabdus*sp. BG5 and *Xenorhabdus*sp. BMMCB respectively.

(19) LL. 497-498: "leads to their overestimation when they occur on contig edges" - is the point here that genes that secondary metabolite genes disproportionately split by contig breaks due to their length and repetitive nature, and therefore have inflated counts? If so this might be stated more explicitly. Miller et al. 2017 Mar. Drugs 15:165 and Klassen and Currie 2012 BMC Genomics 13:14 are both references that make this point explicitly.

Thank you for your correction. Yes, this is the point.However, we did not get this point from either of the two references but from our own experiences in *Xenorhabdus*genome studies.Lines 507-510 were rephrased as follows to make the point more explicit.

"However, this enrichment was likely overestimated because genes encoding secondary metabolites of *Xenorhabdus*species are often long and clustered in genome loci that span thousands of base pairs, which leads to their frequent fragmentation in draft genomes, resulting in inflated counts"

VERSION 1

Editor recommendation and comments

https://doi.org/10.1099/acmi.0.000531.v1.5

Rubén de Dios; Brunel University London, Life Sciences, UNITED KINGDOM

Date report received: 06 January 2023
Recommendation: Major Revision

Comments: In this manuscript, Awori et al. present the analysis of various Xenorhabdus genomes. As this bacterial genus can be the source of potential antimicrobial compounds, the description of new species does have an urgent interest for the scientific community. The topic is introduced well and the overall methodological approach and result interpretation seems fairly well performed. The manuscript has been reviewed by two experts in the field and their reviews are enclosed. However, as spotted by the reviewers, several nuances must be introduced. Moreover, various points for the methodology and the results would need further clarification, and a number of statements need to be softened to properly reflect the results. Please consider the reviewers' comments thoroughly and address their concerns point by point in a separate document. A revised manuscript should include appropriate revisions.

Reviewer 2 recommendation and comments

https://doi.org/10.1099/acmi.0.000531.v1.4

Anonymous.

Date report received: 22 December 2022
Recommendation: Major Revision

Comments: In this manuscript, Awori et al. describe the genomes of several novel Xenorhabdus strain isolated from Kenya, adding to the diversity of this genus and setting the stage for future biotechnological applications. Overall, I agree with the author's conclusions regarding the novelty of these strains, although there are some places where I think that further clarification is needed, especially to avoid overstating what has actually been discovered in this work. Major comments: (1)Several times, the authors describe genes that are unique to each lineage as "causing speciation" (e.g., L. 49, L. 378, L. 507). It is true that gene gain and loss can underpin the creation of species boundaries, although how this works in bacteria is still subject to debate. What is clear is that not all differences in gene content or variation are causal in the speciation process, e.g., much likely accumulates via drift or is genetically linked to causal variants. Because this study only identifies strain-specific variation and does not directly test whether that variation causes speciation, such causal language should be removed throughout the manuscript. (2) Theauthorsrepeatedly imply that each Steinernema hosts a unique species of bacterial symbiont (e.g., LL. 79-83; LL. 99-100). However, this is untrue, e.g., the association with X. bovienii with multiple species of Steinernema hosts has been studied in some detail - Murfin et al. 2015 mBio 6:e00076-15, Murfin et al. 2015 BMC Genomics 16:889. That said, those studies did show that there was co-adaptation of X. bovienii strains with the Steinernema hosts from which they had been isolated. Thus, unique variation also exists in Xenorhabdus below the level of species in this case. These nuances should be better reflected in the manuscript. (3) It would be useful to know more about the Steinernema nematodes from which the strains in this study were isolated, if that

information is available. Given the introductory material regarding co-adaptation between distinct bacteria-host pairs, it would be especially useful to know if the sampled nematodes represented unique species, especially the hosts for the two X. griffiniae strains. (4) I recommend that the authors moderate the language about defining new species somewhat (e.g., LL. 334-336), given that their analysis is not sufficient to describe new species following the International Code of Nomenclature of Prokaryotes, which would require publication of novel species names in the International Journal of Systematic and Evolutionary Microbiology. I agree that the author's genomic analyses are consistent with BG5 and BMMCB representing currently unnamed species within the genus Xenorhabdus, but this is distinct from and more general than the phrasing used here. (5) Given that the genomes used in the pangenome analysis are draft quality, meaning that they lack certain genes due to assembly artifacts, it seems to me that using 100% presence in all analyzed genomes to define the set of "core" genes is too strict a threshold, even if this is the default used by anvi'o. Instead, I suggest defining the core by using the dendrogram at the center of Figure 2 that clusters gene families by the conservation between the analyzed genomes, which seems to have two main branches that seem to separate conserved and variable genes, even if they are not 100% conserved. Finer nodes might alternatively used, but regardless the issue of how genome quality affects genome content analyses should be explicitly addressed. Similar caveats apply to Figure 3, although this issue seems more difficult to mitigate the with fewer genomes presented here. (6) LL. 358-359: Because two X. griffiniae strains were used in the analysis vs. only a single strain of all other species, it is inappropriate to directly compare the unique genes in each strain to those in the other species. Instead, it would be stronger to compare the genes that are both unique to each X. griffiniae plus those that are shared between the two X. griffiniae strains but absent from all other sampled Xenorhabdus species. (7) The analysis suggesting that gene gain or loss arose to differences in gene content between the four focal strains analyzed here (LL. 380-385) is incomplete and needs to be modified because it does not include a reconstruction of gene content of the common ancestor of the entire BMMCB/BG5/XN45/VH1 clade. This would require discussing the gene content of the phylogenetic neighbors of this clade, i.e., X. bozodoii and the ancestor of X. thuongxuanensis/X. ehlersii/X. ishibashii/X. eapokensis. Methods exist to do this (e.g., ANGST; David and Alm 2011 Nature 469:93) but these were not applied here. (8) I must admit that I have difficulty reconciling the authors' transposon analysis with the BRIG analysis in Figure 4. As I understand it, genomic islands in BG5 are indicated by gaps in the outer two rings, i.e., these genes are lacking in XN45 and VH1. Given the statements about transposons flanking gene islands, I expected the arrows representing transposons to align with those gaps, but they do not consistently do so, i.e., there are many gaps without transposons associated with them and places where the transposons do not obviously align with gaps. Is this because the gaps are relatively small? If most of the gaps are, in fact, not associated with transposons, then I think that this needs to be more explicit in the text, because it implies that transposons are only associated with a minority of genome differences and therefore not the main driver of these differences. Minor comments: (9) Permission for collecting the samples used in this study or that such permission is not necessary must be indicated. (10) Table 1 caption: the genome quality statistics listed in this table are said to be derived from PGAP, but I am unaware of this pipeline performing such an analysis. The methods state that PATRIC was used for this analysis instead, which seems the more likely citation here. (11) LL. 298-300: The terms "monophyletic" and "paraphyletic" are used incorrectly here. "Paraphyletic" refers to the situation where members of two different taxa are interdigitated amongst each other within the same phylogenetic clade. This is not the case here, where BG5 and BMMCB, which are both clearly not X. griffiniae, are clear outgroups to XN45 and VH1, which are; thus, X. griffiniae is monophyletic. All four strains do form a single clade that contains three putative species (contra L. 299 - "BG5 did not form a clade") that would be paraphyletic if BG5 was classified as a different species but BMMCB retained its X. griffiniae classification. However, this paper clearly removes that former classification. (12) Table 2 caption: I suggest adding "of the matrix" to "the top half" and "the bottom half", it took me a while to understand what exactly was going on here. Also, are the ANI and dDDH values given averages of the bidirectional tests? These values sometimes vary slightly based on which genome is used as a query, depending on the implementation. (13) LL. 338-339: "...a pangenome that lacked many strain-specific genes" - I think that "lacked" should actually be "included" here, especially given that the following text describes those genes. (14) As a suggestion, the authors might consider reordering the strains in Figure 2. At the very least, BMMCB and BG5 should be next to XN45 and VH1 because this is the central comparison made in this study. Using an ordering that matches Figure 1 might also make comparisons more logical (i.e., ordering by phylogeny instead of alphabetically). (15) Similarly, I also suggest using the same general order (i.e., from largest to smallest or vice versa) in all of the bar charts in Figure 3b and Supplementary Figure S1 so that they are more directly comparable to each other. Otherwise it gives the impression of differences (largest bars at the top vs. at the bottom) that do not actually exist in the data. (16) L. 466 and L. 470: replace "nascent" with "draft" (17) L. 474: replace "failed to pass conspecific thresholds" with more specific language that describes the actual result, i.e., that the dDDH, ANI, and GGD values were not consistent with these strains belonging to the same species. (18) L. 474: the "sp. nov." designation should only be used when describing a new binomial name, and thus is inappropriate here. It indicates the first use of a species name, not that the strains being analyzed represent a new species that has yet to be formally described with such a name. (19) LL. 497-498: "leads to their overestimation when they occur on contig edges" - is the point here that genes that secondary metabolite genes disproportionately split by contig breaks due to their length and repetitive nature, and therefore have inflated counts? If so this might be stated more explicitly. Miller et al. 2017 Mar. Drugs 15:165 and Klassen and Currie 2012 BMC Genomics 13:14 are both references that make this point explicitly.

Please rate the manuscript for methodological rigour
Satisfactory

Please rate the quality of the presentation and structure of the manuscript
Satisfactory

To what extent are the conclusions supported by the data?
Partially support

Do you have any concerns of possible image manipulation, plagiarism or any other unethical practices?
No

Is there a potential financial or other conflict of interest between yourself and the author(s)?
No

If this manuscript involves human and/or animal work, have the subjects been treated in an ethical manner and the authors complied with the appropriate guidelines?
Yes

Reviewer 1 recommendation and comments

https://doi.org/10.1099/acmi.0.000531.v1.3

Anonymous.

Date report received: 16 December 2022
Recommendation: Minor Amendment

Comments: In this study, the authors isolated new Xenorhabdus strains from nematodes collected from different soils in Kenya. Xenorhabdus species are endosymbiont of Steinernema nematodes which infect and kill insects. The isolation of new Xenorhabdus strains could be of interest to identify new antimicrobials that are produced for this genus of bacteria to eliminate competitors from the soil once the insect is dead. They sequenced, annotated, analysed and compared the genomes of the isolates to classify them phylogenetically and elucidate the pangenome of the Xenorhabdus genus, using for that other Xenorhabdus genomes already published. The pangenome was used to extract the core gene cluster (gene cluster present in all the analysed genomes), accessory gene cluster (present only in some genomes) and strain specific gene cluster (found only in one strain), which were functionally categorizes. The work is well conducted and the results are presented clearly. However, there are some minor issues that should be addressed: Methodological rigour, reproducibility and availability of underlying data 1. In the section of materials and methods "Isolation of nematodes from soils samples", the authors described the use of Galleria mellonella larvae as bait. They should indicate from where they got the Galleria larvae and how they can ensure that the isolated nematodes were not already infecting the larvae. 2. Any Galleria larvae was used or the author selected them within a range of weight. This should be specified in the text. 3. In the section of materials and methods "Isolation of bacteria from nematodes", there are not any reference. How were Xenorhabdus sp. isolated previously? The author should clarify that in this section. Presentation of results 4. There are 27 genomes in the Table S2, which is cited in the text in the section "Creation of pangenomes". However, only 26 genomes (25 species) are used in the construction of the pangenome (Figure 2). 5. In the description of Table S2 is written "…genomes used in phylogenomic and pangenomic analyses". Therefore, Table S2 should be also cited in section "Phylogenomic reconstruction and calculation of ANI values". In this section, is stated that "26 fasta files were used as input…" (Line 215), however, in the Figure 1 there are 27 genomes. 6. Please, clarify the number of genomes used for each section and why X. lircayensis is included in the phylogenomic analyses but not in the pangenome. Any other relevant comments 7. It is difficult to read the page 5 (Introduction, from line 103 to 123) where multiple Steinernema sp. and Xenothabdus isolates are named. I suggest to the authors to include this part in results (supplementary table) and add the Steinernema sp., the corresponding Xenothabdus isolate, if any, the location, and the reference. 8. Line 37 in supplementary Table S2: (Typo) Replace phlyogenomic by phylogenomic.

Please rate the manuscript for methodological rigour
Good

Please rate the quality of the presentation and structure of the manuscript
Good

To what extent are the conclusions supported by the data?
Strongly support

Do you have any concerns of possible image manipulation, plagiarism or any other unethical practices?
No

Is there a potential financial or other conflict of interest between yourself and the author(s)?
No

If this manuscript involves human and/or animal work, have the subjects been treated in an ethical manner and the authors complied with the appropriate guidelines?
Yes

SciScore report

https://doi.org/10.1099/acmi.0.000531.v1.1

iThenticate report

https://doi.org/10.1099/acmi.0.000531.v1.2