

Evidence of homologous recombination as a driver of diversity in *Brachyspira pilosicoli*

Anish Pandey^{1,2}, Maria Victoria Humbert¹, Alexandra Jackson¹, Jade L. Passey³, David J. Hampson⁴, David W. Cleary^{1,2}, Roberto M. La Ragione^{3,*} and Myron Christodoulides¹

Abstract

The enteric, pathogenic spirochaete *Brachyspira pilosicoli* colonizes and infects a variety of birds and mammals, including humans. However, there is a paucity of genomic data available for this organism. This study introduces 12 newly sequenced draft genome assemblies, boosting the cohort of examined isolates by fourfold and cataloguing the intraspecific genomic diversity of the organism more comprehensively. We used several *in silico* techniques to define a core genome of 1751 genes and qualitatively and quantitatively examined the intraspecific species boundary using phylogenetic analysis and average nucleotide identity, before contextualizing this diversity against other members of the genus *Brachyspira*. Our study revealed that an additional isolate that was unable to be species typed against any other *Brachyspira* lacked putative virulence factors present in all other isolates. Finally, we quantified that homologous recombination has as great an effect on the evolution of the core genome of the *B. pilosicoli* as random mutation ($r/m=1.02$). Comparative genomics has informed *Brachyspira* diversity, population structure, host specificity and virulence. The data presented here can be used to contribute to developing advanced screening methods, diagnostic assays and prophylactic vaccines against this zoonotic pathogen.

DATA SUMMARY

- (1) All samples sequenced and assembled during the course of this study have been deposited at the National Center for Biotechnology Information's (NCBI's) GenBank database under Bioproject PRJNA564276.
- (2) Metadata and GenBank accessions for the samples alongside those of other genomes used in analyses during this study are outlined in Table 1.

INTRODUCTION

Brachyspira (previously *Treponema*, *Serpula* and *Serpulina*) is the sole genus of the family *Brachyspiraceae* within the order Spirochaetales, phylum Spirochaetes [1]. The genus *Brachyspira* includes nine officially recognized species [2], several of which are pathogenic – primarily to pigs and poultry, although they can also infect other animals and humans. These species are flagellated, anaerobic, aero-tolerant Gram-negative spirochaetes that inhabit the large intestine, where they are intimately associated with the colonic or caecal mucosa. Infection with *Brachyspira hyodysenteriae* (classical agent), *Brachyspira hamptonii* or *Brachyspira suanatina* causes swine dysentery (SD), a severe

Received 23 September 2020; Accepted 16 October 2020; Published 11 November 2020

Author affiliations: ¹Molecular Microbiology, School of Clinical and Experimental Sciences, University of Southampton Faculty of Medicine, Southampton General Hospital, Southampton SO16 6YD, UK; ²Southampton NIHR Biomedical Research Centre, University Hospital Southampton NHS Trust, SO166YD, UK; ³Department of Pathology and Infectious Diseases, School of Veterinary Medicine, Faculty of Health and Medical Sciences, University of Surrey, Guildford, GU2 7AL, UK; ⁴School of Veterinary and Life Sciences, Murdoch University, Murdoch, Western Australia 6150, Australia.

*Correspondence: Roberto M. La Ragione, r.laragione@surrey.ac.uk

Keywords: *Brachyspiragenus*; *Brachyspira pilosicoli*; pangenome; recombination, microbial evolution.

Abbreviations: AIS, avian intestinal spirochaetosis; ANI, average nucleotide identity; API NIN, active pharmaceutical ingredient ninhydrin; BHI, brain heart infusion; CDS, coding sequence; CFML, clonalframe maximum-likelihood; COG, cluster of orthologous group; FABA, fastidious anaerobe blood agar; FBS, foetal bovine serum; GC, guanine-cytosine; GFF, general feature format (gene-finding format, generic feature format); HIS, human intestinal spirochaetosis; IDBA-UD, iterative de Bruijn graph de novo assembler; LOS, lipooligosaccharide; MAFFT, multiple sequence alignment program for unix-like operating systems; ML, maximum-likelihood; MLST, multi-locus sequencing typing; NADH, nicotinamide adenine dinucleotide + hydrogen; PCR, polymerase chain reaction; PHYLIP, phylogeny inference package; PIS, porcine intestinal spirochaetosis; Roary, rapid large-scale prokaryote pan genome analysis; SD, swine dysentery; SGA, string graph assembler; SNP, single-nucleotide polymorphism; SRST2, short-read sequence typing 2; VF, virulence factor; VMA, vacuolar membrane ATPase; XMFA, extended multi-fastA.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. One supplementary file is available with the online version of this article.

000470 © 2020 The Authors



This is an open-access article distributed under the terms of the Creative Commons Attribution License. This article was made open access via a Publish and Read agreement between the Microbiology Society and the corresponding author's institution.

colitis in pigs [3–5]. *Brachyspira murdochii* and *Brachyspira pilosicoli* also infect pigs, but cause milder colitis symptoms [6]. Infection of chickens with either *Brachyspira intermedia* or *Brachyspira alvinipulli* causes avian intestinal spirochaetosis (AIS) [7]. *Brachyspira innocens* is an enteric commensal of pigs, chickens and rats, and no particular disease has been associated with this species [8]. However, Burch *et al.* found that *B. innocens* infection was associated with poor performance in flocks and that this organism was significantly associated with below-target egg production in free-range flocks [9]. *Brachyspira aalborgi* was assumed to be the sole aetiological agent causing histologically identified intestinal spirochaetosis (IS) in humans (HIS) [10], but it was then reported that human gut colonization by *B. pilosicoli* is also common [11, 12].

B. pilosicoli is distinguished from the other *Brachyspira* species by its broad host range [2, 13] and zoonotic potential [13–16], and for being the sole causative agent of porcine intestinal spirochaetosis (PIS) [17]. *B. pilosicoli* can also infect chickens (causing AIS) [18, 19], wild ducks [20, 21], domesticated turkeys [22], pheasants [23, 24], rodents [25, 26], dogs [27–29] and horses [30], among several other animal species [2, 31]. Risk factors associated with human infection by *B. pilosicoli* include faecal–oral contamination and ingestion of water, living rurally and/or among animals, crowding, socioeconomic depression, travel to-and-from less economically developed countries and positive HIV status. A review collating global human clinical manifestations and prevalence of *B. pilosicoli* [2] cites that it is commonly found in faecal samples collected in the Middle East, Southeast Asia and rural Australia, but for example is less prevalent (~1.5%) in urban parts of Australia [32], the UK and Belgium.

Comparative genomics provides an opportunity to investigate the diversity and interactions of pathogens circulating among domestic animal populations and humans. There is evidence of horizontal gene transfer taking place in the gut environment shaping both the phylum Spirochaetes and genus *Brachyspira*. Horizontal gene acquisition of auxiliary and secondary metabolism genes by phylum Spirochaetes from Gram-positive Firmicutes has been noted [33]. In addition, *B. hyodysenteriae* found in the porcine large intestine possess carbohydrate metabolism genes that are more similar to those of *Clostridia* spp. and *Escherichia coli* than to those of other spirochaetes [34]. Genomic characterization is important for understanding *Brachyspira* diversity, population structure, host specificity and virulence, and could contribute to developing advanced screening, diagnostic and control measures, including vaccines. The aim of our study is to contribute a better understanding of this potentially zoonotic pathogen by providing an in-depth genomic analysis of *B. pilosicoli* isolates.

METHODS

Isolates, strains and samples

A total of 14 samples were isolated from chickens with AIS in Australia and the UK. *B. pilosicoli* isolates B04 and B06 were isolated in 1994, and B12 and B14, and B31 and B37 were

Impact Statement

The genus *Brachyspira* contains gut bacteria that can cause a disease called spirochaetosis in specific hosts. Among these organisms, *Brachyspira pilosicoli* is characterized by its unusually broad host range and its potential to cross from one species to another (zoonosis). However, there has been very little genomic data available for this organism. Our study is important for several reasons: (1) we introduced 12 newly sequenced draft genome assemblies to the literature, a 4-fold increase in the number of examined isolates; (2) we catalogued the intraspecific genomic diversity of the organism(s) comprehensively; (3) we found 1 isolate that was unable to be species typed against any other *Brachyspira*, and lacked putative virulence factors present in all other isolates, potentially suggesting a new species; (4) we quantified that homologous recombination has as great an effect on the evolution of the core genome of *B. pilosicoli* as random mutation. Overall, our in-depth characterization of the *Brachyspira* is important for understanding their genetic diversity, population structure, host specificity and virulence. All of these are important attributes that can contribute to efforts to develop new screening and diagnostic tools, and control measures such as vaccines.

all isolated in 1998: each pair was isolated from a different farm in South East Queensland, Australia. The farms were either broiler breeder farms or egg layers. The UK isolates SAP_774, SAP_822, SAP_859, SAP_865, SAP_894, SAP_898 and SAP_772 were collected from different farms in the UK between 2007–2011. All isolates were from different free range egg-layers. Of these, 13 were species designated by Kraken v0.10.5-beta [35] with the miniKraken database v20141208 to be *B. pilosicoli* and 1 – isolate B11 isolated in Australia – was a species designated as *B. murdochii*. These isolates were assembled as described below and are presented in Table 1.

Publicly available genus *Brachyspira* and *B. pilosicoli* reference strains were acquired from the Refseq and GenBank repositories using the ncbi_ftp_download script from bacs-genomics-scripts [36]. Confirmation of genus and species designation was done using Kraken. A total of 4 *B. pilosicoli* and 8 *Brachyspira* spp. genomes were sourced and are displayed in Table 1 alongside 40 publicly available *B. hyodysenteriae* genome assemblies in Data S1 (available in the online version of this article). The most complete draft genome (highest N50 value and lowest contig number) was selected for a given *Brachyspira* species where more than one was available.

Microbiological identification prior to sequencing α -glucosidase, β -glucosidase and α -galactosidase activity tests

Suspensions of *Brachyspira* in 0.1 M sterile phosphate-buffered saline (PBS) (\geq McFarland 4.0) were prepared by transferring

Table 1. Isolates used in this study and genome assembly metrics

<i>Brachyspira</i> species	Strain	Host	Location	Accession	GC (%)	Total length	# contigs	N50
<i>B. murdochii</i>	B11	Chicken	Australia	VYIL00000000*	27.70	3156632	26	528032
<i>B. pilosicoli</i>	B04	Chicken	Australia	VYIM00000000*	27.92	2607200	11	1319479
<i>B. pilosicoli</i>	B06	Chicken	Australia	VYIN00000000*	27.81	2789651	80	165358
<i>B. pilosicoli</i>	B12	Chicken	Australia	VYIO00000000*	27.91	2593600	9	2586710
<i>B. pilosicoli</i>	B14	Chicken	Australia	VYIP00000000*	27.91	2588857	47	181709
<i>B. pilosicoli</i>	B31	Chicken	Australia	VYIQ00000000*	27.91	2594260	11	1911521
<i>B. pilosicoli</i>	B37	Chicken	Australia	VYIR00000000*	27.91	2594260	11	1911521
<i>B. pilosicoli</i>	SAP_774	Chicken	UK	VYIS00000000*	27.90	2568130	30	2542180
<i>B. pilosicoli</i>	SAP_822	Chicken	UK	VYIT00000000*	27.88	2746377	69	102423
<i>B. pilosicoli</i>	SAP_859	Chicken	UK	VYIU00000000*	27.82	2546568	20	227560
<i>B. pilosicoli</i>	SAP_865	Chicken	UK	VYIV00000000*	27.83	2547658	20	417927
<i>B. pilosicoli</i>	SAP_894	Chicken	UK	VYIW00000000*	27.91	2713772	58	195119
<i>B. pilosicoli</i>	SAP_898	Chicken	UK	VYIY00000000*	27.80	2913681	253	92127
<i>B. ?</i>	SAP_772	Chicken	UK	VYIX00000000*	27.90	2745085	21	182551
<i>B. aalborgi</i>	513	Human	Denmark	†	28.11	2558428	11	2532317
<i>B. hamptonii</i>	NSH-16	Pig	USA	CP019914	27.38	3189639	1	3189639
<i>B. hyodysenteriae</i>	ATCC27164	Pig	USA	NZ_CP015910	27.04	3074045	2	3041447
<i>B. innocens</i>	B256/ATCC29796	Pig	UK	ARQI01000129	27.73	3281611	130	52799
<i>B. intermedia</i>	PWS/A	Pig	UK	CP002874	27.21	3308048	2	3304788
<i>B. murdochii</i>	DSM 12563	Pig	Canada	CP001959	27.75	3241804	1	3241804
<i>B. pilosicoli</i>	95/1000	Pig	Australia	CP002025	27.90	2586443	1	2586443
<i>B. pilosicoli</i>	B2904	Chicken	UK	CP003490	27.79	2765477	1	2765477
<i>B. pilosicoli</i>	P43/6/78	Pig	UK	CP002873	27.92	2555556	1	2555556
<i>B. pilosicoli</i>	WesB	Human	Australia	HE793032	27.73	2889522	1	28899522
<i>B. suanatina</i>	AN4859/03	Pig	Sweden	CVLB01000001	27.00	3256103	30	2243936

All *Brachyspira* genomes used during the course of this study are listed in this table. Isolates assembled during the course of this study are separated from assemblies sourced from public repositories by a dividing line and have the "*" symbol following their accession number. The species of every assembly has been confirmed by Kraken and the accession numbers provided were obtained from NCBI GenBank. Genome statistics, including GC percentage, total length of the isolate genome, number of contigs constituting the assembly and the largest contig and the N50 value are given. N50 is a weighted median statistic that comments on the distribution of contig lengths and overall genome assembly quality. Fifty per cent of the entire assembly is contained in contigs equal to or larger than the N50 value. 'B' refers to the genus *Brachyspira*. †The *B. aalborgi* strain was obtained from the Sanger METAHIT consortium (<https://www.sanger.ac.uk/resources/downloads/bacteria/metahit/>).

surface growth from FABA agar using a sterile swab. Aliquots of the suspensions were transferred into separate universal tubes and an α -glucosidase (p-nitrophenyl- α -D-glucopyranoside), β -glucosidase (p-nitrophenyl- β -D-glucopyranoside) or α -galactosidase (p-nitrophenyl- α -D-galactopyranoside) diatab (Rosco Diagnostics) was added to each aliquot. The suspensions were incubated anaerobically at 37°C for 16h. Results were recorded whereby a yellow colour change was regarded as positive and absence of colour change was regarded as negative for the respective enzyme activity.

Hippurate test

Suspensions of *Brachyspira* in 1% (w/v) sodium hippurate solution (Sigma-Aldrich) (\geq McFarland 2.0) were prepared by transferring surface growth from FABA agar using a sterile swab. The suspensions were incubated anaerobically at 37°C for 24h, after which 150 μ l of API NIN (ninhydrin) reagent (BioMérieux) was added. Following 10 min incubation at room temperature, results were recorded whereby a blue-purple colour change was regarded as positive and a clear-orange colour change was regarded as negative for the

Table 2. Primer sequences for genus *Brachyspira*- and species-specific PCR

Target species	Target gene	Primer name	Primer sequence (5'–3')	Size (bp)	Reference
Genus <i>Brachyspira</i>	16S rRNA	Br16S-F	TGAGTAACACGTAGGTAATC	1309	[37]
		Br16S-R	GCTAACGACTTCAGGTA AAAAC		
<i>Brachyspira pilosicoli</i>	16S rRNA	Acoli-F	AGAGGAAAGTTTTTCGCTT	439	[38]
		Acoli-R	CCCCTACAATATCCAAGACT		

ability to hydrolyse sodium hippurate to glycine and sodium benzoate.

Indole test

Suspensions of *Brachyspira* in brain heart infusion (BHI) medium supplemented with 10% (v/v) foetal bovine serum (FBS) (\geq McFarland 4.0) were prepared by transferring surface growth from FABA agar using a sterile swab. The inoculated BHI broth was incubated anaerobically at 37°C for 24h, after which 150 μ l of API JAMES (Kovac's) reagent (BioMérieux) was added. Following a 10 min incubation at room temperature, results were recorded whereby the formation of a pink-red pellicle was regarded as positive and a yellow pellicle was regarded as negative for the ability to cleave indole from tryptophan.

Brachyspira polymerase chain reaction (PCR)

Established genus *Brachyspira*- and species-specific PCRs were used for the initial identification of *Brachyspira* isolates. The genus-specific PCRs were based on the 16S rRNA gene and used to confirm that all the isolates were from the genus *Brachyspira* [37]. PCRs designed for the identification of *B. pilosicoli* were based on two well-conserved genes in the genus, the NADH oxidase (nox) and 16S rRNA genes [38] (Table 2). For PCR, a 20 μ l reaction mixture was prepared consisting of GoTaq Master Mix (Promega), the forward and reverse primers (20 pmol μ l⁻¹) (Sigma-Aldrich), DNA template (20–50 ng μ l⁻¹) and DNase-free water. PCR amplifications were performed using a Techne thermocycler as follows: 95°C for 5 min to denature the DNA, followed by 30 cycles of denaturation at 95°C for 60 s, annealing at 55°C for 60 s and an extension at 72°C for 60 s, followed by a final DNA extension at 72°C for 7 min. Samples were then cooled to 4°C.

Genus *Brachyspira*-specific primers targeted the amplification of the 16S rRNA gene, which produced an amplicon of 1309 bp [37] (Table 2). *B. pilosicoli* species-specific primers targeting the 16S rRNA gene in *B. pilosicoli* produced an amplicon of 439 bp [38]. Gel electrophoresis of the PCR products was performed, and the species of each isolate was inferred using the amplicon sizes stated above and summarized in Table 2.

Growth, DNA extraction and sequencing

B. pilosicoli strains (and one accidental isolate of *B. murdochii*) were grown for 2 to 5 days on fastidious anaerobe agar (LAB090) supplemented with defibrinated sheep blood (5% v/v) at 37°C in an anaerobic gas jar with the Anaerogen gas-generating system (Oxoid, UK). Genomic DNA (gDNA)

was extracted with the Wizard Genomic DNA Purification kit (Promega, USA) following the manufacturer's protocol. Purified gDNA was quantified by absorbance with a NanoDrop UV-Vis Spectrophotometer 1000 (Thermo Fisher Scientific, USA) and with a Qubit 2.0 Fluorometer (Thermo Fisher Scientific, USA). The quality of all gDNA sample preparations was assessed by 0.7% (w/v) agarose gel electrophoresis before sequencing. Isolates were sequenced at the Oxford Genomics Centre (<https://www.well.ox.ac.uk/ogc/microbial-dna-sequencing/>) on a MiSeq to generate 2×300 bp paired-end data.

Read quality control and genome assembly

Paired-end reads were assembled using the automated, A5-miseq pipeline v20160825 [39]. Briefly, adapter sequence removal was done using Trimmomatic v0.32 [40], followed by k-mer-based error correction using SGA v0.9.9 [41]. Contigs were assembled using the IDBA-UD algorithm v1.0.6 [42]. Misassemblies were detected and removed where read pairs were not mapping within an expected distance. A final, stringent round of scaffolding repaired broken contigs. Lastly, assembly metrics were determined using Quast v4.0 [43].

Multi-locus sequencing typing (MLST)

Designation of sequence types was done using Short-Read Sequence Typing 2 (SRST2) [44] and the *B. pilosicoli* MLST scheme [45].

Annotation, pangenome construction and phylogenetic inference

Genome assemblies were annotated with Prokka v1.12 [46], using the -use_genus flag and a list of proteins derived from three, previously annotated *B. pilosicoli* genomes with the -proteins flag. GFF annotations were used in conjunction with MAFFT v7.3.1 [47] as part of the Roary pipeline v3.8.0 [48] to generate the core genome alignment (gene presence based on BLASTP seq-ID: 0.95 and presence among 99% strains) alongside the pangenome. The core alignment was used in conjunction with FastTree v2.1.11 [49], recompiled with -duse_double to resolve shorter branch lengths found between isolates with little variation. The generalized time reversible (-gtr) and nucleotide (-nt) model was used. Phylogenetic tree visualization and editing was done using Figtree v1.4.2 [50]. The pangenome was visualized alongside the phylogenetic tree using Phandango [51] (<https://jameshadfield.github.io/phandango>; date of accession: 19 July 2020).

Table 3. MLST allele data for 12 *B. pilosicoli* isolates

Isolate	<i>adh</i>	<i>alp</i>	<i>est</i>	<i>gdh</i>	<i>glp</i>	<i>pgm</i>	<i>thi</i>	ST
B04	5	111	116	12	87	131	111	ST186
B06	4	107	112	69	3	102	103	ST187
B12	5	37	105	27	18	49	35	ST134
B31	5	37	105	27	18	49	35	ST134
B37	5	37	105	27	18	49	35	ST134
B67	3	11	12	9	8	132	10	ST182
SAP_774	3	108	113	22	88	133	112	ST183
SAP_822	55	109	114	68	89	82	113	ST184
SAP_865	3	110	12	25	90	129	114	ST188
SAP_859	3	110	12	25	90	129	114	ST188
SAP_894	3	3	115	59	91	86	115	ST185
SAP_898	55	109	114	68	89	82	113	ST184

MLST allele data for the 12 *B. pilosicoli* isolates. Novel alleles and sequence types (STs) are indicated in bold. Allele abbreviations are as follows: *adh*, alcohol dehydrogenase; *alp*, alkaline phosphatase; *est*, esterase; *gdh*, glutamate dehydrogenase; *glpK*, glucose kinase; *pgm*, phosphoglucomutase; and *thi*, acetyl-CoA acetyltransferase.

Clusters of orthologous groups (COGs)

COGs were calculated by separating the core and accessory genome proteins and using the `cdd2cog_script` from `bacsgenomics-scripts` [36] to assign functional groups.

Recombination

B. pilosicoli assemblies were mapped to the closed genome of *B. pilosicoli* 95/1000 using Parsnp v1.2 [52] under default parameters. This generated a core genome alignment in XMFA format, which was converted to FASTA format using the `xmfa2fasta.pl` script (https://github.com/kjolley/seq_scripts/blob/master/xmfa2fasta.pl; date of accession: 14 August 2020). The recombination analysis program Clonal-FrameML (CFML) v1.11 requires both a FASTA alignment and phylogenetic tree as inputs. A maximum-likelihood (ML) tree was created by converting the FASTA alignment to PHYLIP format using EMBOSS `seqret` v6.6 [53] and running PhyML v3.20180621 [54] on the PHYLIP alignment using the HKY85 model. This process also calculated a transition/transversion ratio to be used as the `-kappa` input parameter in CFML. ML trees were converted to binary format using the R package `ape` v3.0 [55]. Bootstrapping (`-emsim`) was used until recombination parameter variance values were $<10^{-6}$. Recombination was visualized on the core genome alignment using the `cfml_results.R` script available in the ClonalFrameML github (<https://github.com/xavierdidelot/ClonalFrameML>; date of accession: 1 September 2020). Recombinant genes of interest to the study were found by extracting ‘importations’ detected by CFML and compared via BLASTX to the reference genome used to create the input core genome alignment. This identified genes that were functionally annotated by submitting the sequences to the Uniprot (discussed further below)

and EggNOG V5.0 (<http://eggno5.embl.de/#/app/home>) databases [56].

Average nucleotide identity (ANI)

ANI analyses were performed using FastANI [57], an orthology mapper utilized to calculate p-dist values (%), and the Python module PyANI [58] was used to infer and visualize ANI via multiple tests and heatmaps. These included (i) 1020 bp fragment BLASTN+ analysis, (ii) ANI-BLAST_all via legacy BLASTN on 1020 bp fragments (ANIb), (iii) MuMer alignment (ANIb/m) and (iv) tetranucleotide frequency analysis [59]. ANI was performed to distinguish isolate SAP_772 from the best representative genome assembly of the other eight *Brachyspira* spp. ($n=8$) as well as the newly assembled isolates of *B. pilosicoli* ($n=12$) and four reference genomes of *B. pilosicoli* ($n=4$).

Putative virulence factor dataset generation and clustering

A list of 231 putative virulence factors was created by combining and filtering replicate UniProt accession numbers from (i) a list of proteins with pathogenic and virulence potential [60] and (ii) a list of 26 mostly surface-exposed proteins identified as cross-reacting bands using immune pig sera [61]. The UniProt accessions were converted into coding sequence (CDS) nucleotide accessions from UniProt (<https://www.uniprot.org/>) and then downloaded using the batch entrez service (<https://www.ncbi.nlm.nih.gov/sites/batchentrez>). Both sets of identifiers are listed with full annotations and cluster identities in Data S2. The virulence factors in coding sequence format were converted into a format suitable for use with SRST2 by following the steps for generating a custom

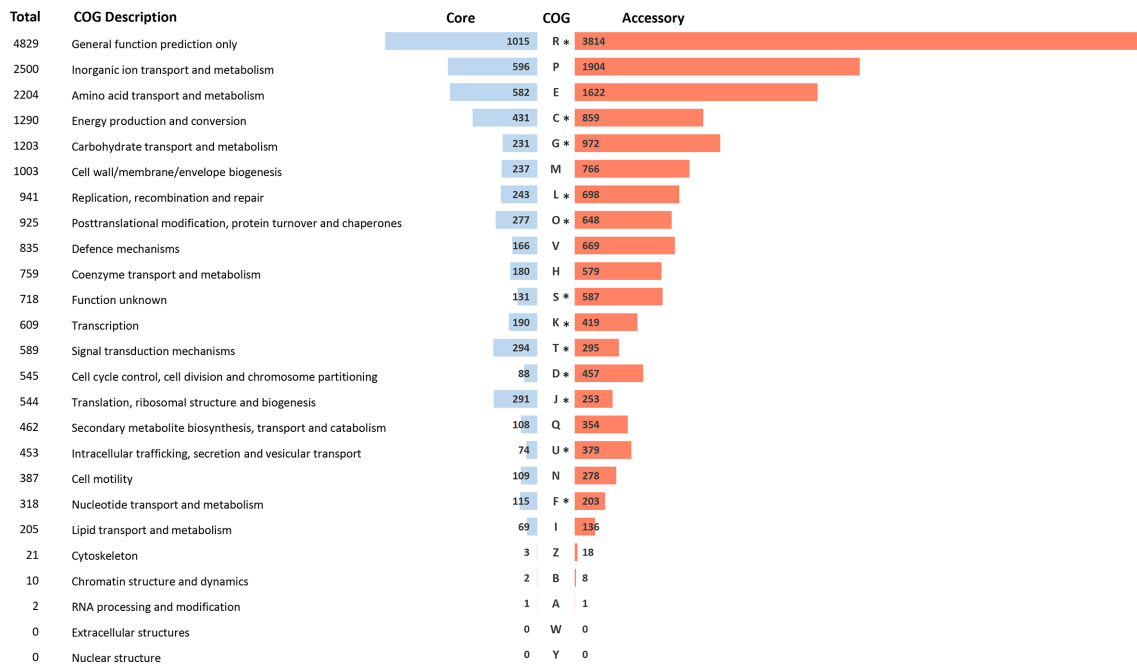


Fig. 1. COG category metrics and proportional distribution in the pangenome of *B. pilosicoli*. The diagram shows the percentage proportion of COGs in the core (blue) vs accessory genomes (red). *, there is a significant difference in the distribution of the numbers of a category present in the core and accessory genome compared to COGs in the core/accessory not assigned to that given COG category. (Contingency table, χ^2 -corrected, 1 degree of freedom, twin-tailed, Bonferroni corrected P value<0.002.)

database (<https://github.com/katholt/srst2#generating-srst2-compatible-clustered-database-from-raw-sequences>). As part of the database clustering step, the database of 231 putative virulence factors was clustered into 207 putative virulence factors. This was due to sequences that shared a gene name and annotation despite having a sequence identity of less than 90%. Gene names were edited manually with alphabetic characters in order to be recognized as alternate alleles by SRST2. For example, *arp_A*, *arp_B* were both annotated ‘*arp*’ but occupied separate sequence clusters). SRST2 v.1.8 was run with default parameters and used to map isolate read data (.fastq) against the virulence factor list to generate a list of (i) exact matches; (ii) >90% coverage matches, where a minimum of 90% isolate reads aligned to a virulence factor sequence, and the difference is given in SNPs and insertions/deletions (SRST2 interprets these matches as allelic variants of virulence factors); (iii) uncertain matches identified due to poor read depth, truncation or base ambiguity; and (iv) a lack of matches.

Statistics

General statistics (mean, standard deviation) were calculated using LibreOffice Calc v6.3.0 for Debian-based Linux systems. To determine which COG groups were over-represented in the core, accessory and pangenome, the composition of each COG category was examined in a 2x2 contingency table using the Chi-square test with twin-tailed P -value and Yate’s correction [62]. This was done using the GraphPad Prism calculator (<https://www.graphpad.com/quickcalcs/contingency1.cfm>,

date of accession: 1 September 2020). A Bonferroni-adjusted P -value of 0.002 (0.05 P -value/25/COG tests) was used as the significance threshold.

RESULTS

Genome assembly

A total of 14 newly sequenced genomes that were all initially thought to be *B. pilosicoli* isolates were assembled. Subsequently, 12 of these were confirmed to be *B. pilosicoli*, 1 was identified as *B. murdochii* and 1 was an unknown *Brachyspira* species (SAP_772). The identity and source of the isolates, and the genome assembly quality control metrics such as GC content, length, contig number and N50 are shown in Table 1. Average *B. pilosicoli* genome length was 2645822 bp (+/-106304) with a GC content of 27.87% (+/-0.06). The 12 *B. pilosicoli* isolates were clustered into 8 sequence types (STs), based on a combination of 7 housekeeping loci (alleles of *adh*, *alp*, *est*, *gdh*, *glpK*, *pgm* and *thi* genes). We identified seven of these STs as novel (Table 3). This brings the total count of *B. pilosicoli* sequences type up to 94 STs.

Isolate SAP_772 produced a good-quality draft assembly. The data from the quality control from the A5 pipeline revealed that an extremely high proportion (99.9%) of reads passed the error correction and the contigs were kept to a low number compared to other *B. pilosicoli* assemblies. In addition to post-assembly read error correction, post-assembly quality control via Kraken revealed no unclassified contigs.

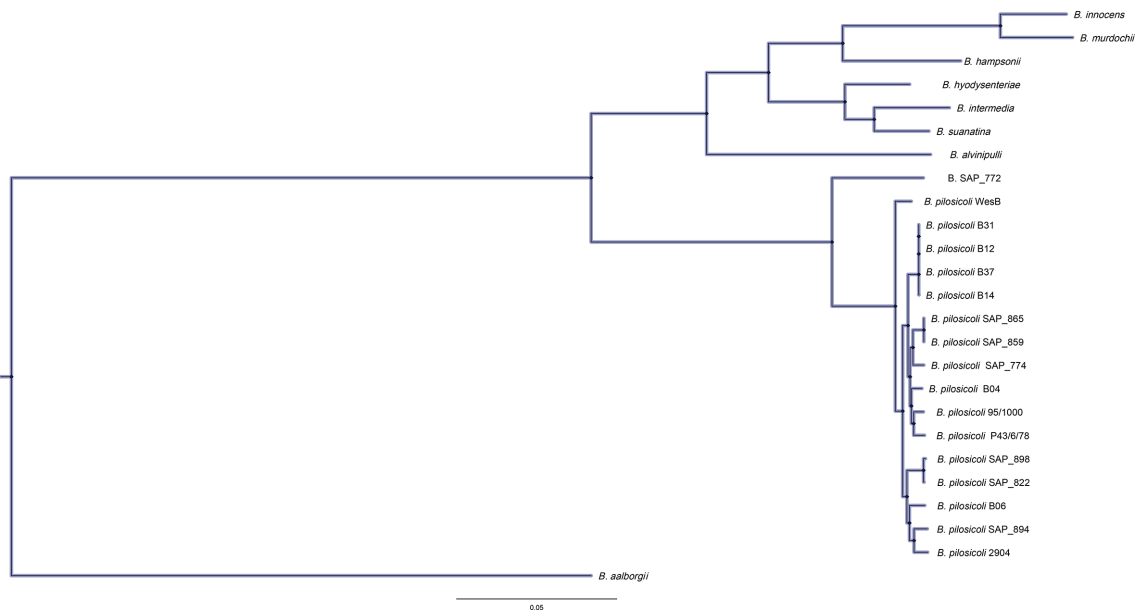


Fig. 2. Phylogenetic tree. Phylogenetic inference reveals the extent of genomic diversity both within the genus *Brachyspira* and interspecifically among *B. pilosicoli* isolates. The tree was generated using RaxML version 8.2 and a GTR model. An alignment of 27 CDS was found to be core (95% Seq-ID) to this dataset generated using the MAFFT aligner, rooted with *B. aalborgii* genome, which was found to be the most distant.

***B. pilosicoli* pangenome analysis**

A pangenome approach using the Roary pipeline provided a quantitative measure and insight into the shared genomic content of 12 study *B. pilosicoli* isolates and 4 reference *B. pilosicoli* isolates. The *B. pilosicoli* pangenome comprised 4590 genes. The core genome (present in 99–100% of isolates tested) was 1751 genes. There were no genes present in the soft-core category (present in 95–99% isolates). The accessory genome (Data S3) consisted of the shell (1760 genes present in the genomes of 15–95% of isolates tested) and the cloud (1079 genes present in the genomes of 1–15% of isolates tested).

COG analysis

Distribution of COG categories was determined for the core and accessory genome (Fig. 1). The COG categories that were found most commonly among the *B. pilosicoli* isolates were inorganic ion transport and metabolism (COG P, $n=2500$), amino acid transport and metabolism (COG E, $n=2204$) and energy production and conversion (COG C, $n=1290$). Relative to the accessory genome, the *B. pilosicoli* core genome contained a higher proportion of genes involved with signal transduction and translation (COG T) and ribosomal structure and biogenesis (COG J). Of the 25 COG categories, 13 were present at significantly different distributions between the core and accessory genome compared to total COGs. These data are presented alongside statistics in Data S4.

Phylogenetics

A core genome alignment (99% gene coverage among strains defined as core) was generated via Roary for the

genus *Brachyspira*. Based on a core alignment of 9393 sites (27/21395 total genes), *Brachyspira* isolate SAP_772 occupied one branch (red), but still clustered closer to *B. pilosicoli* than any other species (Fig. 2). Among the genus *Brachyspira*, *B. innocens* and *B. murdochii* were found to group together, as did *B. intermedia* and *B. suanatina* (blue). Rooting by the lowest common ancestor (Spirochaete: *Brevinema andersonii*, based on 16S rRNA phylogeny of bacteria and archaea strains) was not possible owing to the loss of tree resolution following its inclusion.

Recombination

Recombination analysis was performed on the *B. pilosicoli* core genome ($n=16$ isolates, $n=1751$ genes). These metrics included the relative rate of recombination to mutation: $R/\theta=0.38\pm 5.9E^{-6}$, the mean length of detected recombinant regions: $1/\delta=1.9E^{-2}\pm 1.5E^{-8}$ and mean divergence level between recipient and donor: $Nu=5.1E^{-2}\pm 4.0E^{-8}$. The parameters were multiplied together to calculate the relative effect of recombination compared to the relative effect of mutation ($r/m=1.02$). The metrics revealed that while homologous recombination occurred approximately one-third (0.38: R/θ) as often as mutation, the overall effect of recombination (in introducing diversity) on the core genome of *B. pilosicoli* was approximately the same as the effect of mutation ($r/m=1.02$). To make a broad comparison, 40 public genome assemblies of *B. hyodysenteriae* were analysed similarly. Here, the R/θ was 0.81 and the resultant r/m value was 3.99. The visualization showed that recombination events were found scattered throughout the phylogeny of *B. hyodysenteriae* (Data S5).

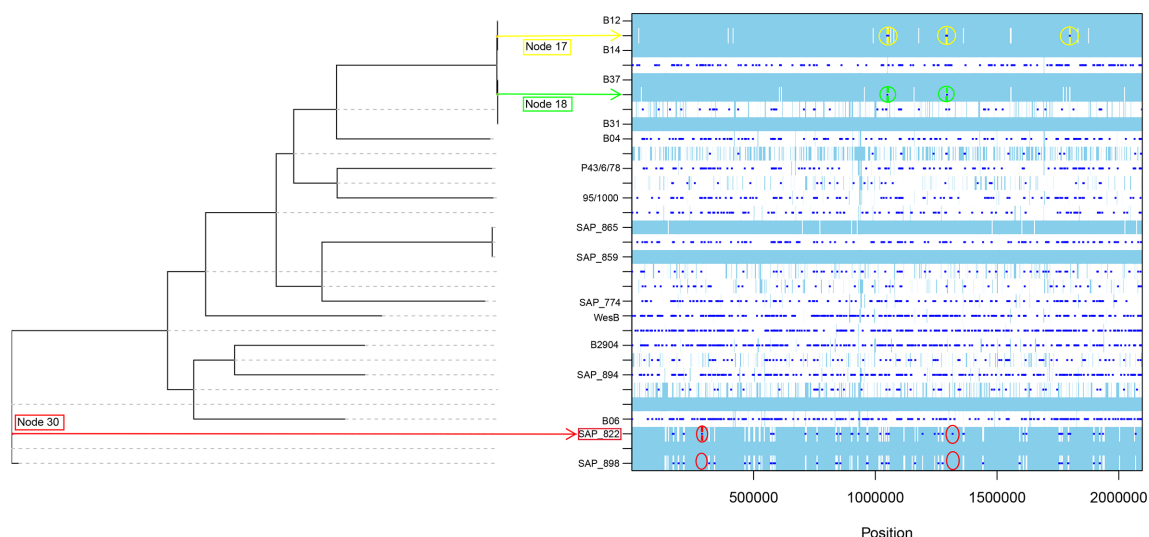


Fig. 3. Recombination detection on the core genome of *B. pilosicoli*. Recombination events and sites are indicated along the branches of the phylogenetic tree. A given branch may display results for a specific isolate (see isolate label) or for an ancestral phylogenetic node (no isolate label). Recombination events are indicated by dark blue bars, while light blue sites are used to indicate no substitution and white sites indicate that a convergent mutation has occurred (one base or more) at that point in the phylogeny. When multiple convergent events occur within short nucleotide distance of each other a recombination event or ‘importation’ is identified by the software at a specific position in the core alignment affecting x number of isolates or x number of ancestral phylogenetic nodes which affect multiple isolates. Recombination events discussed in the Results section are indicated with yellow, green and red circles. Their corresponding genes are identified with full annotation data in Data S6.

Recombination events were detected frequently throughout the core genome phylogeny of *B. pilosicoli* but were not universally distributed among every isolate examined (Fig. 3). Among closely related isolates, recombination analysis can identify recombination events (specifically, ‘importations’ of DNA bases that do not match the clonal genealogy). These data can be used to identify the genes affected by this allelic variation or novel gene gain that separates one isolate from another isolate(s) or group of isolates on a node from another.

Data S6 shows that nine putative recombination events were detected in five genes, i.e. amidotransferase; *gatA* (alanine cation symporter); *alsT* (amino acid carrier); *cpdB* (nucleotidase); hypothetical proteins hp1 (putative growth rate regulator) and hp2 (putative cell division activator). These events were detected in the ancestral node separating ST 134 isolates B12 and B14 from B37. Furthermore, two importations were detected among another *alsT* gene and a hypothetical protein (no obvious function) in the node separating B37 from B31. In addition, recombination events were found in ST 184 isolates indicating gene gain in isolate SAP_822. These importations of DNA were (i) absent in recombination events and (ii) absent of convergent mutation in SAP_898 at the same or similar position in the core genome alignment. While both genes gained were hypothetical with unknown function in the COG database, Hp 4 contained DUF3298 domains that are also found in peptidoglycan deacetylase proteins such as PdaC of *Bacillus subtilis*. Interestingly, of all the genes, Hp 5 was found to be a spirochete-specific oxidoreductase containing VMA domain (a potential ‘intein’ or self-splicing,

parasitic, mobile domain). Homologues of this protein are found in a select group of enteric pathogens that includes *B. intermedia*, *B. hampsonii* and *Treponema* spp.

Comparing the *B. pilosicoli* phylogenetic tree, before and after recombination correction, revealed an alternated phylogeny not just with regards to the branch lengths but also to the topology of the tree (Data S7). Accounting for recombination resulted in the generation of a new ancestral node in the phylogenetic tree, revealing that isolates SAP_822 and SAP_898 did not share the same branch as indicated previously.

Discovery of a potential new species similar to *B. pilosicoli*

In the primary analysis of the pangenome data, an anomalous result was seen by the inclusion of isolate SAP_772, which was originally thought to be a *B. pilosicoli* based on Kraken species designation. Thus, SAP_772 was excluded from any analysis related to the *B. pilosicoli* core genome. In comparison to the data mentioned above, the *B. pilosicoli* pangenome including SAP_772 expanded the total gene count from 4590 genes to 5772 genes spread across the now 17 isolates tested. Core genome size decreased from 1751 to 983 genes and the accessory genome (shell+cloud) increased in size. Again, no genes were identified as belonging to the soft-core genome category (the genes present in 95–99% of isolates tested). The pangenome is displayed qualitatively in Fig. 4 and confirms that gene content distinguishes *B. pilosicoli* SAP_772 from the rest of the isolates tested, with large portions of the genome

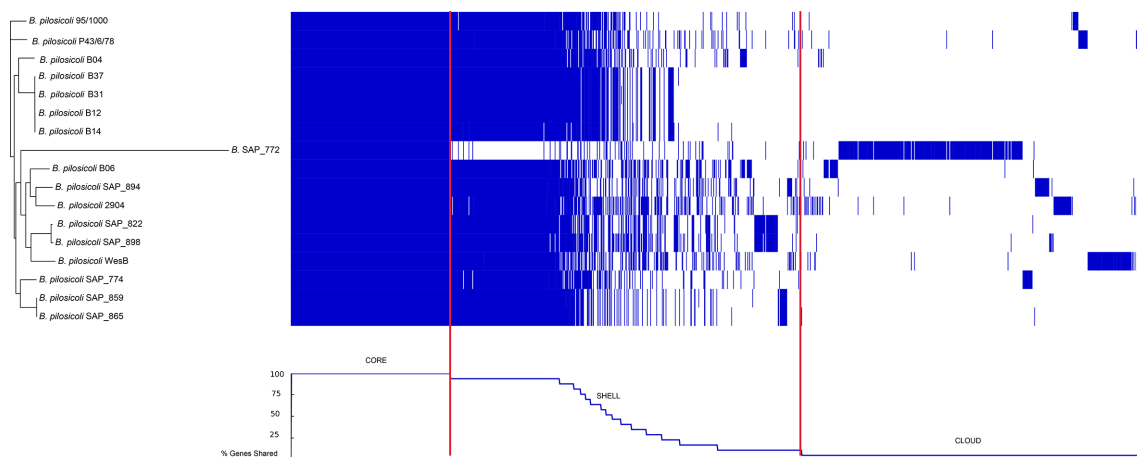


Fig. 4. Gene distribution map of the *B. pilosicoli* pangenome plus *Brachyspira* SAP_772. On the left of the figure is an unrooted, ML phylogeny based on an accessory genome alignment (for more accurate inference of phylogenetic relationships see Fig. 2). In the centre, blue segments represent gene presence and white segments represent gene absence. The pangenome is displayed, starting from the core genome on the left and transitioning into the accessory genome (shell and cloud genomes) with increasing gene sequence disparity. The bottom graph displays a trace showing the percentage of isolates containing blue segments of gene presence. Similar to the centre, the graph starts with the core genome ($n=17$ genomes) and falls steadily as it transitions into the accessory [shell: $n=2-16$ isolates (approximately 95% - >1% of the *B. pilosicoli* plus *B. SAP_772* cohort) and cloud: $n=1$ isolate, approx. <1% of the cohort) genome displaying blocks of genes shared by fewer and fewer genomes. Red lines are used to mark the transitions from the core genome to the shell and cloud genomes.

that would otherwise be core, missing from the isolate. The accessory genome was quantified (Data S3), and this revealed that *B. pilosicoli* SAP_772 possessed the greatest allelic and genomic diversity, with 1197 novel genes and/or gene variants. This diversity was ~3.5-fold greater than the next most diverse isolate (reference strain *B. pilosicoli* WesB, with 343 novel gene and/or allelic variants) and 5-fold greater than the third most diverse isolate (reference strain *B. pilosicoli* 2904, with 220 novel gene and/or allelic variants).

To discriminate the intraspecific genetic diversity between isolate SAP_772 and other *Brachyspira* spp. isolates, a cohort was created where the genomes of confirmed *B. pilosicoli* isolates ($n=16$; 12 newly assembled and 4 publicly sourced) was combined with 8 representative genomes of the other *Brachyspira* species. This cohort was subjected to species delineation using whole-genome average nucleotide identity (ANI). This metric demonstrates a percentage of the average nucleotide matches among all orthologous genes between two or more comparator genomes. An ANI of >95% is indicative of the interspecific border. FastANI analysis (Data S8) showed that isolate SAP_772 had an average ANI of 80.31% \pm 1.43 when compared to the genus *Brachyspira*, and an average ANI of 91.63% \pm 0.09 when compared to *B. pilosicoli*. PyANI was used to calculate ANI and visualize the data via heatmaps using four experimental methods (Fig. 5). In three out of four analyses (ANI-muscle, ANI-BLAST and ANI-BLAST_all vs all, Fig. 5a-c) PyANI confirmed the FastANI results, i.e. that SAP_772 was divergent from the rest of the *Brachyspira* spp., and although more related to *B. pilosicoli*, was beyond the boundary for inclusion in this species. In the final method,

tetranucleotide analysis (ANI-Tetra, Fig. 5d), SAP_772 nested amongst the *B. pilosicoli* isolates.

Detection of genes encoding putative virulence factors

All study isolates ($n=12$ *B. pilosicoli* isolates plus *B. SAP_772*) were screened for the presence of putative virulence factors using SRST2 and a curated list of 207 previously identified genes (Data S2). The raw data are presented in Data S9 and are quantified in Table 4, which shows the number of matches (exact matches plus matches within 90% sequence identity) alongside an absence of matches. In this screening, 11 out of 12 of the *B. pilosicoli* isolates displayed exact matches or putative allele assignment of 196.6 \pm 3.9 out of a total of 207 virulence factors. One strain of *B. pilosicoli* (B14) had fewer matches than the other *B. pilosicoli* isolates, but still had more than isolate SAP_772. Isolate SAP_772 both (i) showed matches between approximately 160 virulence factors and (ii) had the greatest paucity of matches ($n=16$ virulence factors). Undetected matches indicate that <90% of the gene in question was covered by the sequence reads, exceeding the threshold for either an exact match or putative allele sharing >90% sequence identity. Instances where an undetected match was unique to a given isolate are detailed in Table 5. This indicates that isolate SAP_772 has 10 undetected putative virulence factors that are otherwise represented among the rest of the *B. pilosicoli* isolates. Lastly, there were nine putative virulence factors that were undetected in multiple isolates, which are displayed in Data S10. These include CDS B204_orf2390 (basic membrane lipoprotein); BP951000_1477

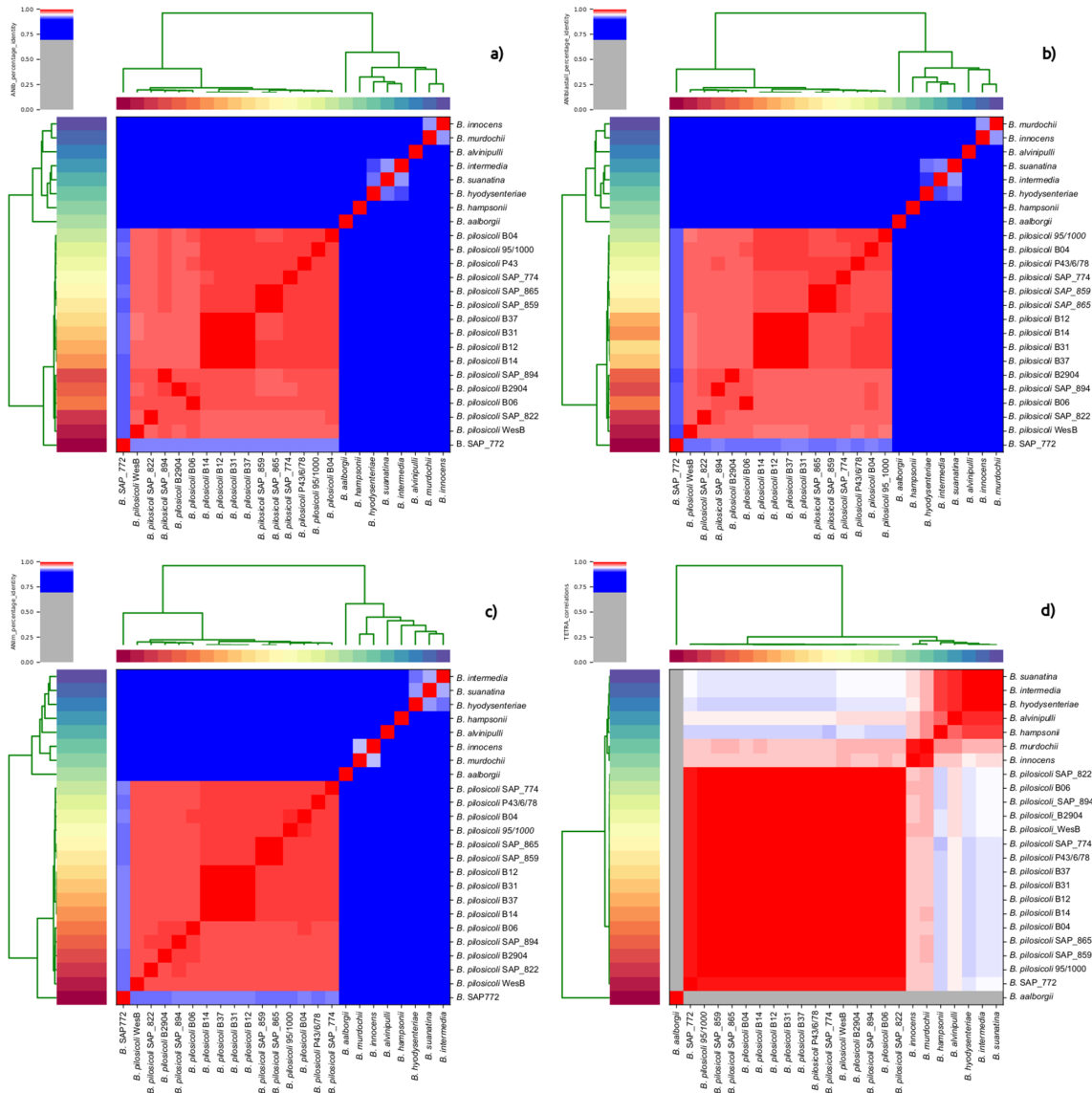


Fig. 5. Average nucleotide identity (ANI). These four heatmaps visualize four tests performed by PyANI. These are (a) 1020 bp fragment BLASTN+ analysis, ANIb (b) ANI-BLAST_all via legacy BLASTN on 1020 bp fragments, ANIb-allvall, (c) MUSCLE alignment ANI-M and, lastly, (d) tetranucleotide frequency analysis ANI-TETRA. As shown by the scale, red indicates increasing homology, while blue denotes decreases in homology between isolates examined in the heatmaps.

(phage capsid protein); BP951000_2112 (peptidase); *bspA* (membrane surface antigen); *gmhB* (lipooligosaccharide (LOS) biosynthesis phosphatase); *lepB* (signal peptidase I); *rfbF* (LOS biosynthesis cytidyltransferase); *vsh* (phage capsid protein); and *vspD* (variable surface protein).

DISCUSSION

To our knowledge, our study is the most comprehensive genomic analysis of *B. pilosicoli* to date and gives insight into the intraspecific genomic diversity of this organism. *Brachyspira* species have a single circular chromosome varying in size from ~2.5 to 3.3 Mb, with a similar G+C content of ~27%, each encoding for >2300 proteins [4, 5, 34, 60, 63–69]. Close

similarities were observed in the 16S rRNA gene sequences of most species, which implies relatively recent speciation events [70]. *B. pilosicoli* has the second smallest genome (~2.6 Mb) of all *Brachyspira* species reported and has no described extra-chromosomal elements, with only *B. aalborgi* being marginally smaller. Nevertheless, *B. pilosicoli* strains, taken together, encode a greater diversity of proteins than other *Brachyspira* species, which is possibly attributable to a greater number of gene duplications than other species of the genus [2, 60, 66, 70].

A previous study [60] compared *B. pilosicoli* strain 95/1000 with *B. hyodysenteriae* strain WA1 and *B. murdochii* strain 56–150 and found that the smaller genome of *B. pilosicoli*

Table 4. Matches and no matches to *B. pilosicoli* and *B. SAP_772* isolates subject to virulence factor analysis

Sample	Matches/ putative alleles	No match
<i>B. pilosicoli</i> B12	200	3
<i>B. pilosicoli</i> B31	200	4
<i>B. pilosicoli</i> B37	199	4
<i>B. pilosicoli</i> SAP_774	199	4
<i>B. pilosicoli</i> B04	198	6
<i>B. pilosicoli</i> SAP_859	198	4
<i>B. pilosicoli</i> SAP_865	197	4
<i>B. pilosicoli</i> B06	196	7
<i>B. pilosicoli</i> SAP_898	194	3
<i>B. pilosicoli</i> SAP_894	190	9
<i>B. pilosicoli</i> SAP_822	188	10
<i>B. pilosicoli</i> B14	171	6
<i>B. SAP_772</i>	159	16

This table lists the lack of matches versus exact matches and instances of putative allele assignment using SRST2 default parameters. Exact matches are instances where sequencing reads have mapped in perfect alignment with a virulence factor CDS. SRST2 assigns putative allele status to any sequence sharing a minimum of 90% coverage with reads mapping to it. Uncertain results flagged by the pipeline were inspected manually and removed from counts.

95/1000 contained more genes than the other species in the COGs for cell motility (COG-N), intracellular trafficking, secretion and vesicular transport (COG-U), energy production and conversion (COG-C), co-enzyme transport and metabolism (COG-H) and lipid transport and metabolism (COG-I). From our study, the greater diversity of genes involved particularly in energy production and conversion (COG C) in the *B. pilosicoli* isolates agrees with this prior analysis.

Phylogenetic inference coupled with ANI suggests a close relatedness between *B. murdochii*, *B. innocens* and *B. suanatina*, as well as between *B. intermedia* and *B. hyodysenteriae*. *B. pilosicoli* and *B. aalborgi* stand apart, with the latter remaining the most phylogenetically distant from all other *Brachyspira* species. This observation is supported by previous studies using DNA–DNA hybridization or whole-genome comparison, which also found a high level of homology between *B. murdochii*, *B. intermedia*, *B. innocens* and *B. hyodysenteriae* [60, 71, 72].

In contrast to the clonal population structure of other *Brachyspira* species (the sole exception so far being *B. aalborgi* [66]), the population structure of *B. pilosicoli* is thought to be driven by high levels of recombination [16, 66, 70]. This observation has been based on multilocus enzyme electrophoresis [73],

Table 5. List of putative virulence factors undetected among one sample of isolate reads from *B. pilosicoli* or *B. SAP_772*

Isolate	No.	Undetected putative virulence factors
SAP_772	10	B2904_orf1521, iron/sulphur flavoprotein
		BP951000_1807, membrane lipoprotein
		<i>murD</i> , member of the <i>mur</i> operon of lipooligosaccharide (LOS) biosynthetic modification ligases and synthases
		<i>pldB</i> , phospholipase
		<i>rfaD</i> , LOS biosynthetic epimerase
		<i>tolR</i> , outer-membrane biopolymer transport protein
		<i>arp_G*</i> , ankyrin repeat-containing protein
		<i>batA_B*</i> , aerotolerance-related membrane protein
		<i>mcpB_H*</i> , methyl-accepting chemotaxis protein B
		<i>mcpB_K*</i> , methyl-accepting chemotaxis protein B
SAP_894	3	BP951000_0437, peptidase C14, caspase catalytic subunit p20
		BP951000_1159/ <i>pppI</i> , family intracellular peptidase
		<i>ftsH</i> , ATP-dependant zinc metalloprotease
SAP_822	3	ADK31727/ <i>pep</i> , peptidase
		BP951000_0437, peptidase C14, caspase catalytic subunit p20
		BP951000_1779, probable metal-dependent glycoprotease
B06	2	B2904_orf2005, lipoprotein
		B2904_orf651, lipoprotein
B14	1	<i>pep_A*</i> , peptidase
SAP_774	1	BP951000_2039, putative periplasmic binding protein
B04	0	
B12	0	
SAP_859	0	
SAP_865	0	
B31	0	
SAP_898	0	
B37	0	

Accession numbers, cluster identities and functional annotation available for putative virulence factors are detailed in Data S1. *These proteins had the same functional annotation as another protein named the same but possessed <90% sequence identity. They are labelled with an underscore and capital letter to distinguish them for virulence factor screening.

pulsed-field gel electrophoresis [74], MLST [16] and variable-number tandem repeat analysis [75]. Through the higher resolution whole-genome analysis, we detected a plethora of recombination events spread throughout the core genome of *B. pilosicoli*. Principally, we found that isolates phylogenetically clustered according to location data: an exception to this was *B. pilosicoli* B06, which was found to be genetically more similar to isolates sourced from the UK poultry farms, rather than from the Australian farms. This was confirmed in both genus-specific and species-specific phylogenetic comparisons. This may be either an example of a common strain-specific evolution in an avian enteric pathogen, or maybe an adaptation of isolates found in egg-laying chickens. The UK chickens in this study only included egg-laying birds and not broiler birds. This makes sense from an evolutionary point of view, as the species has the potential to inhabit and infect multiple hosts and may therefore import a variety of exogenous DNA from other species found within these varied environments. While recombination and substitution events were detected in most isolates examined and in every ancestral node of the phylogenetic tree, a paucity of these events was seen when comparing isolates SAP_859, SAP_865, B37, B31, B14 and B12. These data, combined with low accessory genome size and extremely high ANI (when comparing the isolates within both groups), suggested evidence of gene loss and gain events and/or allelic variation specific to isolates all within the same sequence type. A recent study acknowledged that besides introducing diversity, homologous recombination acted as a cohesive force on population structure among prokaryotic species with >95% ANI and a minimum r/m value of 0.25 [76].

Correcting for recombination in the phylogenetic inference of *B. pilosicoli* revealed not only corrections to the branch lengths, but also to the topology of the tree, as two isolates (SAP_898 and SAP_822) were seen to diverge from a common ancestor, rather than occupying the same branch. Our study inferred an r/m value of 1.02, based upon a core genome alignment of >2000000 sites, which was based upon mapping 15 isolates to the *B. pilosicoli* 95/1000 reference genome. The r/m metric infers that recombination is just as likely as mutation to have an impact upon the evolution of the core genome of this organism. This value is 5 times greater than that previously calculated for the genus *Brachyspira* [77], although this estimate was based on 7 housekeeping loci across 36 *Brachyspira* spp. strains. A comparison of r/m values of multiple bacterial genera found *Brachyspira* to generate one of the lowest signals [78]. To our knowledge, no other study has studied genome-wide recombination for this species. *B. pilosicoli* does not have the highly recombinogenic activity of organisms such as *Neisseria* spp. [78] and it is not as clonally structured as *B. hyodysenteriae* and *B. intermedia* [16].

Through the introduction of novel and varied fragments, homologous recombination was thought to be responsible for shaping the evolution of the zoonotic pathogen *B. pilosicoli*, in contrast to the more clonal population structure of other *Brachyspira* spp. such as *B. hyodysenteriae* and *B. intermedia* [16, 72]. Our study also revealed a dynamic impact on

horizontal gene transfer events, affecting the core genome of *B. hyodysenteriae* to a far greater degree than *B. pilosicoli*. This is likely because r/m is always calculated relative to a sample. A *B. hyodysenteriae* cohort with double the isolates and greater inter-sample phylogenetic diversity gave the potential of many more sites in the core alignment analysed. This likely resulted in the greater detection of importations of novel DNA relative to *B. pilosicoli*. *B. hyodysenteriae* has a history of horizontal gene transfer events with (a) phage-like virulence plasmids and (b) bacteriophage members of *B. hyodysenteriae* and phylogenetically related members of the *Brachyspira* such as *B. intermedia* and *B. murdochii* [63].

Our study revealed an additional isolate that was unable to be species typed against any other *Brachyspira* and suggests a possible new species that is related to, yet distinguishable from, *B. pilosicoli*. This was indicated by the high accessory genome content of isolate SAP_772 compared to *B. pilosicoli* isolates and by genus-wide phylogenetic inference. SAP_772 was further examined using ANI, which is a quantitative way of defining the species boundary and a well-characterized *in silico* substitute for DNA–DNA hybridization [59]. ANI was used to quantify the distance within the genus *Brachyspira* and between/within *B. pilosicoli*. Organisms that belong to the same species have an ANI value of >95% and this was estimated using multiple methods and two pipelines. *Brachyspira* isolate SAP_772 was found to be similar to *B. pilosicoli*, compared to the rest of the *Brachyspira* spp., by FastANI, phylogenetic inference and three of the four PyANI methods (ANI-muscle, ANI-BLAST and ANI-BLAST_all). However, despite this similarity, the ANI scores and visualizations suggest SAP_772 falls outside the ANI boundary for a species (ANI ≤95%), whereas all other *B. pilosicoli* isolates were within this boundary in every ANI testing method used. One PyANI method, tetranucleotide correlation, showed isolate SAP_772 nesting within *B. pilosicoli*. However, this analysis also resulted in *B. suanatina*, *B. intermedia* and *B. hyodysenteriae*, and *B. alvinipulli* and *B. hampsonii* clustering together. This may indicate a paucity of resolution when using the tetranucleotide correlation with these five *Brachyspira* species.

Furthermore, our study generated an indexed, clustered database of 207 putative virulence factors based on proteins identified in previous studies [60, 61]. On average, the *B. pilosicoli* isolates harboured 196 matches or putative allele assignments. This indicated that the dataset was relatively representative of *B. pilosicoli* as a species and warranted investigation into putative virulence factors that were undetected. Isolate SAP_772 displayed both a reduced virulence factor match score and an increased number of no matches compared to the confirmed *B. pilosicoli* isolates. In particular, our data indicate that SAP_772 lacks a large number of outer-envelope, LOS biosynthesis genes, the presence of which has displayed serological heterogeneity among *B. pilosicoli* strains [79]. The reduced virulence factor score is most likely due to the presence of allelic variants of the 'absent' *B. pilosicoli*-specific VF panel. Taken together, this evidence suggests isolate SAP_772 as a candidate for

full phenotypic characterization to support a potentially novel species designation.

In summary, the contribution of new *B. pilosicoli* isolate genomes generated during the current study has allowed us to quantify and visualize the impact of homologous recombination as a force importing DNA on the core genome of *B. pilosicoli* for the first time. The *in silico* discovery of a potentially novel species represented by an isolate similar to *B. pilosicoli*, but lacking putative virulence factors, suggests potentially greater diversity among the genus *Brachyspira* than previously reported. The identification of a more robust core genome, alongside screening for common putative virulence factors, will aid researchers in developing a vaccine against an organism that is responsible for morbidity among livestock and has the potential for zoonotic transmission amongst a variety of species including pigs, poultry, wildlife species, dogs and humans [66].

Funding information

M. C. and R. M. L. R. are joint senior authors on this manuscript and they were in receipt of a Houghton Trust UK (*Avian Pathology*) research award (unnumbered) for research into diseases of avian species and domestic poultry.

Acknowledgements

The authors acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work. We thank Dr Tom La, Murdoch University, Australia, for assistance with assigning sequence types in MLST.

Author contributions

Conceptualization: M. C., R. M. L. R. Methodology: A. P., M. V. H., A. J., J. P. Software: A. P. Validation: A. P. Formal analysis: A. P., M. V. H., D. W. C., M. C. Investigation: A. P., M. V. H., A. J., J. P. Data curation: A. P. Writing – original draft preparation: A. P., M. V. H., D. W. C., M. C. Writing – review and editing: A. P., M. V. H., D. W. C., D. J. H., R. M. L. R., M. C. Visualization: A. P., M. V. H., M. C. Supervision: M. C. Project administration: M. C., R. M. L. R. Funding: M. C., R. M. L. R.

Conflicts of interest

The authors declare that there are no conflicts of interest.

References

- Ludwig W, Euzéby J, Whitman WB. Draft taxonomic outline of the Bacteroidetes, Planctomycetes, chlamydiae, spirochaetes, Fibrobacteres, Fusobacteria, Acidobacteria, Verrucomicrobia, Dictyoglomi and Gemmatimonadetes. www.bergeys.org/outlines/Bergeys_Vol_4_Outlinepdf01/05/12 2008.
- Hampson DJ. The spirochete *Brachyspira pilosicoli*, enteric pathogen of animals and humans. *Clin Microbiol Rev* 2018;31.
- La T, Phillips ND, Hampson DJ. An investigation into the etiological agents of swine dysentery in Australian pig herds. *PLoS One* 2016;11:e0167424.
- Mirajkar NS, Phillips ND, La T, Hampson DJ, Gebhart CJ. Characterization and recognition of *Brachyspira hampsonii* sp. nov., a novel intestinal spirochete that is pathogenic to pigs. *J Clin Microbiol* 2016;54:2942–2949.
- Mushtaq M, Zubair S, Råsbäck T, Bongcam-Rudloff E, Jansson DS. *Brachyspira suanatina* sp. nov., an enteropathogenic intestinal spirochaete isolated from pigs and mallards: genomic and phenotypic characteristics. *BMC Microbiol* 2015;15:208.
- Osorio J, Carvajal A, Naharro G, Rubio P, La T et al. Identification of weakly haemolytic *Brachyspira* isolates recovered from pigs with diarrhoea in Spain and Portugal and comparison with results from other countries. *Res Vet Sci* 2013;95:861–869.
- Mappley LJ, La Ragione RM, Woodward MJ. *Brachyspira* and its role in avian intestinal spirochaetosis. *Vet Microbiol* 2014;168:245–260.
- Mirajkar NS, Bekele AZ, Chander YY, Gebhart CJ. Molecular epidemiology of novel pathogen “*Brachyspira hampsonii*” reveals relationships between diverse genetic groups, regions, host species, and other pathogenic and commensal *Brachyspira* species. *J Clin Microbiol* 2015;53:2908–2918.
- Burch DGS, Strugnell BW, Steventon A, Watson EN, Harding C. Survey of 222 flocks in Great Britain for the presence of *Brachyspira* species and their effect on production. Proceedings of the 5th International Conference on Colonic Spirochaetal Infections in Animals and Humans Leon, Spain 2009.
- Hovind-Hougen K, Birch-Andersen A, Henrik-Nielsen R, Orholm M, Pedersen JO et al. Intestinal spirochetosis: morphological characterization and cultivation of the spirochete *Brachyspira aalborgi* gen. nov., sp. nov. *J Clin Microbiol* 1982;16:1127–1136.
- Mikosza AS, Hampson DJ. Human intestinal spirochetosis: *Brachyspira aalborgi* and/or *Brachyspira pilosicoli*? *Anim Health Res Rev* 2001;2:101–110.
- Trivett-Moore NL, Gilbert GL, Law CL, Trott DJ, Hampson DJ. Isolation of *Serpulina pilosicoli* from rectal biopsy specimens showing evidence of intestinal spirochetosis. *J Clin Microbiol* 1998;36:261–265.
- Smith JL. Colonic spirochetosis in animals and humans. *J Food Prot* 2005;68:1525–1534.
- Hampson DJ, Oxberry SL, La T. Potential for zoonotic transmission of *Brachyspira pilosicoli*. *Emerg Infect Dis* 2006;12:869–870.
- Koopman MB, Käsbohrer A, Beckmann G, van der Zeijst BA, Kusters JG. Genetic similarity of intestinal spirochetes from humans and various animal species. *J Clin Microbiol* 1993;31:711–716.
- Neo E, La T, Phillips ND, Alikani MY, Hampson DJ. The pathogenic intestinal spirochaete *Brachyspira pilosicoli* forms a diverse recombinant species demonstrating some local clustering of related strains and potential for zoonotic spread. *Gut Pathog* 2013;5:24.
- Trott DJ, Stanton TB, Jensen NS, Duhamel GE, Johnson JL et al. *Serpulina pilosicoli* sp. nov., the agent of porcine intestinal spirochetosis. *Int J Syst Bacteriol* 1996;46:206–215.
- Le Roy CI, Mappley LJ, La Ragione RM, Woodward MJ, Claus SP. *Brachyspira pilosicoli*-induced avian intestinal spirochaetosis. *Microb Ecol Health Dis* 2015;26:28853.
- Stephens CP, Hampson DJ. Intestinal spirochete infections of chickens: a review of disease associations, epidemiology and control. *Anim Health Res Rev* 2001;2:83–91.
- Jansson DS, Persson M, Zimmerman U, Johansson KE. Phenotypic and genetic diversity among intestinal spirochaetes (genus *Brachyspira*) in free-living wild mallards (*Anas platyrhynchos*) sampled in southern Sweden. *Syst Appl Microbiol* 2011;34:566–575.
- Oxberry SL, Trott DJ, Hampson DJ. *Serpulina pilosicoli*, waterbirds and water: potential sources of infection for humans and other animals. *Epidemiol Infect* 1998;121:219–225.
- Shivaprasad HL, Duhamel GE. Cecal spirochetosis caused by *Brachyspira pilosicoli* in commercial turkeys. *Avian Dis* 2005;49:609–613.
- Jansson DS, Bröjer C, Gavier-Widén D, Gunnarsson A, Fellström C. *Brachyspira* spp. (*Serpulina* spp.) in birds: a review and results from a study of Swedish game birds. *Anim Health Res Rev* 2001;2:93–100.
- Webb DM, Duhamel GE, Mathiesen MR, Muniappa N, White AK. Cecal spirochetosis associated with *Serpulina pilosicoli* in captive juvenile ring-necked pheasants. *Avian Dis* 1997;41:997–1002.
- Backhans A, Jansson DS, Aspán A, Fellström C. Typing of *Brachyspira* spp. from rodents, pigs and chickens on Swedish farms. *Vet Microbiol* 2011;153:156–162.
- Backhans A, Johansson K-E, Fellström C. Phenotypic and molecular characterization of *Brachyspira* spp. isolated from wild rodents. *Environ Microbiol Rep* 2010;2:720–727.

27. Duhamel GE, Trott DJ, Muniappa N, Mathiesen MR, Tarasiuk K et al. Canine intestinal spirochetes consist of *Serpulina pilosicoli* and a newly identified group provisionally designated "*Serpulina canis*" sp. nov. *J Clin Microbiol* 1998;36:2264–2270.
28. Hidalgo A, Rubio P, Osorio J, Carvajal A. Prevalence of *Brachyspira pilosicoli* and "*Brachyspira canis*" in dogs and their association with diarrhoea. *Vet Microbiol* 2010;146:356–360.
29. Oxberry SL, Hampson DJ. Colonisation of pet shop puppies with *Brachyspira pilosicoli*. *Vet Microbiol* 2003;93:167–174.
30. Hampson DJ, Lester GD, Phillips ND, La T. Isolation of *Brachyspira pilosicoli* from weanling horses with chronic diarrhoea. *Vet Rec* 2006;158:661–662.
31. Duhamel GE. Comparative pathology and pathogenesis of naturally acquired and experimentally induced colonic spirochetosis. *Anim Health Res Rev* 2001;2:3–18.
32. Margawani KR, Robertson ID, Hampson DJ. Isolation of the anaerobic intestinal spirochaete *Brachyspira pilosicoli* from long-term residents and Indonesian visitors to Perth, Western Australia. *J Med Microbiol* 2009;58:248–252.
33. Caro-Quintero A, Ritalahti KM, Cusick KD, Löffler FE, Konstantinidis KT. The chimeric genome of Sphaerochaeta: nonspiral spirochetes that break with the prevalent dogma in spirochete biology. *mBio* 2012;3.
34. Bellgard MI, Wanchanthuek P, La T, Ryan K, Moolhuijzen P et al. Genome sequence of the pathogenic intestinal spirochete *Brachyspira hyodysenteriae* reveals adaptations to its lifestyle in the porcine large intestine. *PLoS One* 2009;4:e4641.
35. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014;15:R46.
36. Leimbach A. Bac-Genomics-Scripts: bovine *E. coli* mastitis comparative genomics edition. *Genome Announc* 2016.
37. Phillips ND, La T, Hampson DJ. A cross-sectional study to investigate the occurrence and distribution of intestinal spirochaetes (*Brachyspira* spp.) in three flocks of laying hens. *Vet Microbiol* 2005;105:189–198.
38. Mikosza AS, La T, de Boer WB, Hampson DJ. Comparative prevalences of *Brachyspira aalborgi* and *Brachyspira (Serpulina) pilosicoli* as etiologic agents of histologically identified intestinal spirochetosis in Australia. *J Clin Microbiol* 2001;39:347–350.
39. Coil D, Jospin G, Darling AE. A5-miseq: an updated pipeline to assemble microbial genomes from illumina MiSeq data. *Bioinformatics* 2015;31:587–589.
40. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 2014;30:2114–2120.
41. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res* 2012;22:549–556.
42. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 2012;28:1420–1428.
43. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013;29:1072–1075.
44. Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ et al. SRST2: rapid genomic surveillance for public health and hospital microbiology Labs. *Genome Med* 2014;6:90.
45. La T. *B. pilosicoli* MLST scheme. Available at https://pubmlst.org/bigdb?db=pubmlst_brachyspira_seqdef&page=schemeInfo&scheme_id=4 2018.
46. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.
47. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 2013;30:772–780.
48. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015;31:3691–3693.
49. Price MN, Dehal PS, Arkin AP. FastTree 2-approximately maximum-likelihood trees for large alignments. *PLoS One* 2010;5:e9490.
50. Rambaut A. FigTree, a graphical viewer of phylogenetic trees. <http://tree.bio.ed.ac.uk/software/figtree>. *Inst Evol Biol Univ Edinburgh* 2016.
51. Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM et al. Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics* 2017.
52. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol* 2014;15:524.
53. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. *Trends Genet* 2000;16:276–277.
54. Guindon S, Delsuc F, Dufayard J-F, Gascuel O. Estimating maximum likelihood phylogenies with PhyML. *Methods Mol Biol* 2009;537:113–137.
55. Popescu A-A, Huber KT, Paradis E. Ape 3.0: new tools for distance-based phylogenetics and evolutionary analysis in R. *Bioinformatics* 2012;28:1536–1537.
56. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 2019;47:D309–D314.
57. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 2018;9:5114.
58. Pritchard L, Glover RH, Humphris S, Elphinstone JG, Toth IK. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Analytical Methods* 2016;8:12–24.
59. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A* 2009;106:19126–19131.
60. Wanchanthuek P, Bellgard MI, La T, Ryan K, Moolhuijzen P et al. The complete genome sequence of the pathogenic intestinal spirochete *Brachyspira pilosicoli* and comparison with other *Brachyspira* genomes. *PLoS One* 2010;5:e11455.
61. Casas V, Rodríguez-Asiain A, Pinto-Llorente R, Vadillo S, Carrascal M et al. *Brachyspira hyodysenteriae* and *B. pilosicoli* proteins recognized by sera of challenged pigs. *Front Microbiol* 2017;8:723.
62. Bakshi U, Sarkar M, Paul S, Dutta C. Assessment of virulence potential of uncharacterized *Enterococcus faecalis* strains using pan genomic approach - Identification of pathogen-specific and habitat-specific genes. *Sci Rep* 2016;6:38648.
63. Black M, Moolhuijzen P, Barrero R, La T, Phillips N et al. Analysis of multiple *Brachyspira hyodysenteriae* genomes confirms that the species is relatively conserved but has potentially important strain variation. *PLoS One* 2015;10:e0131050.
64. Häfström T, Jansson DS, Segerman B. Complete genome sequence of *Brachyspira intermedia* reveals unique genomic features in *Brachyspira* species and phage-mediated horizontal gene transfer. *BMC Genomics* 2011;12:395.
65. Lin C, den Bakker HC, Suzuki H, Lefébure T, Ponnala L et al. Complete genome sequence of the porcine strain *Brachyspira pilosicoli* P43/6/78(T). *Genome Announc* 2013;1.
66. Mappley LJ, Black ML, AbuOun M, Darby AC, Woodward MJ et al. Comparative genomics of *Brachyspira pilosicoli* strains: genome rearrangements, reductions and correlation of genetic complement with phenotypic diversity. *BMC Genomics* 2012;13:454.
67. Mirajkar NS, Johnson TJ, Gebhart CJ. Complete Genome Sequence of *Brachyspira hyodysenteriae* Type Strain B-78 (ATCC 27164). *Genome Announc* 2016;4.
68. Pati A, Sikorski J, Gronow S, Munk C, Lapidus A et al. Complete genome sequence of *Brachyspira murdochii* type strain (56-150). *Stand Genomic Sci* 2010;2:260–269.

69. Stanton TB, Postic D, Jensen NS. *Serpulina alvinipulli* sp. nov., a new *Serpulina* species that is enteropathogenic for chickens. *Int J Syst Bacteriol* 1998;48 Pt 3:669–676.
70. Hampson DJ, Wang P. Colonic spirochetes: what has genomics taught us? *Curr Top Microbiol Immunol* 2018;415:273–294.
71. Stanton TB, Fournié-Amazouz E, Postic D, Trott DJ, Grimont PA et al. Recognition of two new species of intestinal spirochetes: *Serpulina intermedia* sp. nov. and *Serpulina murdochii* sp. nov. *Int J Syst Bacteriol* 1997;47:1007–1012.
72. Zuerner RL, Stanton TB, Minion FC, Li C, Charon NW et al. Genetic variation in *Brachyspira*: chromosomal rearrangements and sequence drift distinguish *B. pilosicoli* from *B. hyodysenteriae*. *Anaerobe* 2004;10:229–237.
73. Stanton TB, Trott DJ, Lee JI, McLaren AJ, Hampson DJ et al. Differentiation of intestinal spirochaetes by multilocus enzyme electrophoresis analysis and 16S rRNA sequence comparisons. *FEMS Microbiol Lett* 1996;136:181–186.
74. Atyeo RF, Oxberry SL, Hampson DJ. Pulsed-field gel electrophoresis for sub-specific differentiation of *Serpulina pilosicoli* (formerly '*Anguillina coli*'). *FEMS Microbiol Lett* 1996;141:77–81.
75. Neo E, La T, Phillips ND, Hampson DJ. Multiple locus variable number tandem repeat analysis (MLVA) of the pathogenic intestinal spirochaete *Brachyspira pilosicoli*. *Vet Microbiol* 2013;163:299–304.
76. González-Torres P, Rodríguez-Mateos F, Antón J, Gabaldón T. Impact of homologous recombination on the evolution of prokaryotic core genomes. *MBio* 2019;10:e02494-18.
77. Råsbäck T, Johansson K-E, Jansson DS, Fellström C, Alikhani MY et al. Development of a multilocus sequence typing scheme for intestinal spirochaetes within the genus *Brachyspira*. *Microbiology* 2007;153:4074–4087.
78. Vos M, Didelot X. A comparison of homologous recombination rates in bacteria and archaea. *Isme J* 2009;3:199–208.
79. Lee BJ, Hampson DJ. Lipo-oligosaccharide profiles of *Serpulina pilosicoli* strains and their serological cross-reactivities. *J Med Microbiol* 1999;48:411–415.

Five reasons to publish your next article with a Microbiology Society journal

1. The Microbiology Society is a not-for-profit organization.
2. We offer fast and rigorous peer review – average time to first decision is 4–6 weeks.
3. Our journals have a global readership with subscriptions held in research institutions around the world.
4. 80% of our authors rate our submission process as 'excellent' or 'very good'.
5. Your article will be published on an interactive journal platform with advanced metrics.

Find out more and submit your article at microbiologyresearch.org.