Metagenomic sequencing of clinical samples reveals a single widespread clone of *Lawsonia intracellularis* responsible for porcine proliferative enteropathy

Rebecca J. Bengtsson[1,2], Bryan A. Wee[1,3], Gonzalo Yebra[1], Rodrigo Bacigalupe[1,4], Eleanor Watson[5], Roberto M. C. Guedes[6], Magdalena Jacobson[7], Tomasz Stadejek[8,9], Alan L. Archibald[1], J. Ross Fitzgerald[1,*] and Tahar Ait-Ali[1,*]

Abstract

Lawsonia intracellularis is a Gram-negative obligate intracellular bacterium that is the aetiological agent of proliferative enteropathy (PE), a common intestinal disease of major economic importance in pigs and other animal species. To date, progress in understanding the biology of *L. intracellularis* for improved disease control has been hampered by the inability to culture the organism *in vitro*. In particular, our understanding of the genomic diversity and population structure of clinical *L. intercellularis* is very limited. Here, we utilized a metagenomic shotgun approach to directly sequence and assemble 21 *L. intracellularis* genomes from faecal and ileum samples of infected pigs and horses across three continents. Phylogenetic analysis revealed a genetically monomorphic clonal lineage responsible for infections in pigs, with distinct subtypes associated with infections in horses. The genome was highly conserved, with 94 % of genes shared by all isolates and a very small accessory genome made up of only 84 genes across all sequenced strains. In part, the accessory genome was represented by regions with a high density of SNPs, indicative of recombination events importing novel gene alleles. In summary, our analysis provides the first view of the population structure for *L. intracellularis*, revealing a single major lineage associated with disease of pigs. The limited diversity and broad geographical distribution suggest the recent emergence and clonal expansion of an important livestock pathogen.

DATA SUMMARY

The sequencing reads generated in the current study have been deposited in the Sequence Read Archive (SRA) under the BioProject numbers PRJNA554776 and PRJNA432360. The draft genome for sample LR189 is deposited in the National Center for Biotechnology Information (NCBI) under the BioProject number PRJNA432360. SRA and draft genome accession numbers are listed.

INTRODUCTION

Modern microbiology techniques predominately rely on culture-based analyses and the ability to grow the organism of interest *in vitro* prior to downstream analyses. This has impeded the detection, surveillance and investigation of bacterial pathogens with substantial medical and economic importance that are either non-culturable or difficult to culture, limiting our understanding of their biology and our ability to design treatments and control methods.

The microaerophilic, obligate intracellular bacterium *Lawsonia intracellularis* belongs to the family *Desulfovibrionaceae* of the class *Deltaproteobacteria*, and is the sole species described to date for the genus *Lawsonia* [1]. *L. intracellularis* is the aetiological agent of a non-zoonotic enteric disease known as proliferative enteropathy (PE) in pigs, and has been

detected in a wide range of wild and domestic animal species [2–5]. Two clinical manifestations have been described in pigs: (i) a mild, self-limiting form commonly affecting weaners or young growing animals between 6 to 20 weeks of age and (ii) an acute, severe form with a high mortality rate more commonly observed in animals between 4 to 12 months of age [6]. In addition, the number of equine PE outbreaks in foals has increased worldwide [7–9]. Although hyperplastic lesions resulting in weight loss are observed in pigs and horses, clinical signs and pathology differ between the two hosts [10–12]. Horses tend to develop hypoproteinaemia with acute but non-haemorrhagic diarrhoea, and infection is often self-limiting or subclinical [11]. Previously, cross-species experimental infections in pigs and horses have failed to develop clinical signs, indicating host specificity of *L. intracellularis* subtypes [13, 14].

Biological characterization of *L. intracellularis* has been severely hampered by its fastidious *in vitro* growth requirements [15, 16], and the molecular basis of *L. intracellularis* pathogenesis remains undetermined. To date, only six *L. intracellularis* genome sequences have been deposited in the National Center for Biotechnology Information (NCBI) database, three of which were obtained from cell-cultured samples, an approach that requires extensive passaging in immortalized cell lines. However, such a method is labour-intensive, slow and may risk the introduction of mutations during passaging. The first complete genome for a *L. intracellularis* strain was sequenced using the Sanger-based method in 2006, revealing a 1.4 Mb chromosome and three plasmids of 27, 39 and 194 kbp. Comparative analysis of two additional pathogenic porcine isolates has been carried out, revealing limited genetic differences [17], but the population structure and phylogeny of this pathogen remains unknown.

In the current study, we utilized culture-independent metagenomic sequencing of 21 clinical samples from field outbreaks of porcine and equine PE from 7 countries to assemble metagenome-derived *L. intracellularis* whole genomes. In addition, 3 cell-passaged samples were also sequenced, generating a total of 24 genome assemblies. Comparative genomic and phylogenetic analysis uncovered the population structure of *L. intracellularis* for the first time, revealing a genetically monomorphic clone responsible for infections in pigs and distinct subtypes associated with equine infections. These data have provided novel insights into the genetic diversity of the bacterium, enhancing our understanding of its biology.

METHODS

Bacterial strains, DNA extraction and microbial DNA enrichment

All samples sequenced in this study are listed in Table S1 (available in the online version of this article). Data from six additional strains (GenBank GCA_000055945.1, GCA_000331715.1, GCA_001975945.1, GCA_003312265.1, GCA_003312285.1 and GCA_003312305.1) were retrieved

Impact Statement

Proliferative enteropathy is a common enteric disease in pigs caused by *Lawsonia intracellularis*, highly prevalent across major pig-farming countries worldwide. Infection caused by the pathogen has a significant impact on animal health, welfare and production, imposing a huge economic burden in the pork production industry. Due to the difficulty in culturing the pathogen *in vitro*, little is known regarding its pathogenesis and the genomic diversity of the population remains unknown. The current work explores metagenomic sequencing to obtain *L. intracellularis* genome sequences, providing novel insights into the genetic diversity and evolution of the bacterium, enhancing our understanding of its biology.

from the NCBI. All samples in this study were stored at −80 °C prior to extraction. DNA was extracted from faecal samples using the DNeasy PowerSoil kit (Qiagen), using PowerBead tubes and FastPrep homogenizer (MP Biomedicals) to homogenize and mechanically lyse bacterial cells. DNA from tissue samples were extracted using DNeasy Blood and Tissue kits (Qiagen), with the exception of sample 4242, which was extracted using phenol/chloroform and precipitated by ethanol, as previously described [18]. The NEBNext Microbiome DNA Enrichment kit (New England Biolabs) was used to deplete host DNA from tissue DNA samples. All commercial kits were used according to the manufacturer's recommendations. The quality of DNA was measured using an Agilent 4200 TapeStation System (Agilent Genomics) and its quantity was measured using a Qubit 3.0 fluorometer with the Qubit dsDNA BR Assay kit (Invitrogen). *L. intracellularis* genomic DNA was quantified from all samples by qPCR using primers targeting the *aspA* gene.

Genome sequencing

Genomic DNA libraries were prepared using a TruSeq Nano 550 bp Gel Free kit (Illumina) and library preparation and sequencing service was provided by Edinburgh Genomics. Sequencing was performed on Hi-Seq 2500, Hi-Seq 4000 or Mi-Seq instruments (Illumina).

Sequence processing

The quality of sequence read FASTQ files was assessed using FastQC (Babraham Bioscience Technologies, Cambridge, UK). All raw reads were adapter- and quality-trimmed with Trimmomatic (v 0.36) on paired-end mode [19]. Kraken taxonomic sequence classifier (v 1.0) was used to remove host reads from the filtered reads. A custom Kraken database was constructed containing *Sus scrofa* (GenBank GCA_ 000003025.6), *Rattus norvegicus* (GenBank GCA_000001895.4), *Mus musculus* (GenBank GCA_000001635.8) and *Equus caballus* (GenBank GCA_002863925.1) genomes [20]. Host reads that mapped

to the custom database were removed from the FASTQ files and unclassified reads were used for downstream analyses.

Genome assembly and annotation

Draft genomes of *L. intracellularis* were assembled *de novo* for each isolate using MEGAHIT with the meta-large pre-set parameter [21]. BWA MEM [22] was used to map reads back to the draft assemblies, and the contig coverage within each metagenome assembly was calculated using the script jgi_summarize_bam_contig_depths from the MetaBAT2 package [23]. The assembled contigs within each sample were binned using MetaBAT2. Assessment of genome bins on the level of completeness and contamination was performed using CheckM [24]. In addition, draft genomes were assembled using a reference-guided *de novo* approach using Kraken to extract *L. intracellularis* FASTQ reads prior to assembly. A custom Kraken database containing E40504, N343 and PHE/MN1-00 reference genome sequences was constructed and used to classify *L. intracellularis* reads. Three reference genomes were included to reduce bias towards a single reference. The classified *L. intracellularis* reads were used as input for assembly with the SPAdes genome assembler (v 3.11.1) [25]. The quality of each genome assembly using both methods was assessed using QUAST (v 4.6.0) [26] based on the PHE/MN1-00 reference genome. Coding sequences were inferred and annotated with Prokka [27].

Pangenome analysis

Pangenome analysis was performed using Roary with the 95% BLASTP threshold without splitting paralogues, and the roary2svg.pl script was used to plot and visualize Roary output [28].

Variant calling and phylogenetic inference

Processed reads were mapped onto the *L. intracellularis* PHE/MN1-00 reference genome (GCA_000055945.1) using snippy (v 4.0) [29]. Mapping quality and sequencing coverage of *L. intracellularis* from each sample were assessed using QualiMap (v 2.2.1) [30]. For single nucleotide polymorphisms (SNPs) to be called, a minimum depth of 5× reads was required with a minimum fraction of 0.9 for a variant present in the reads. The prophage-associated genomic island and ribosomal RNA (rRNA) genes were excluded, as these are highly conserved in bacteria leading to unspecific mapping from other species. SNPs were also called using Harvest v1.2 [31] with the same reference genome. The bam file output from snippy was visualized using Artemis [32] and differences in SNPs identified from Snippy and Harvest were checked manually and curated with false positives removed. Detection of recombination within alignments was performed using Gubbins [33] and visualized using Phandango [34].

Maximum-likelihood (ML) phylogenetic inference was performed on non-recombinant core genome SNPs using IQ-TREE (v 1.6.3) [35] with 1000 bootstrap replicates.

ModelFinder implemented in the IQ-TREE software package was used to find the nucleotide substitution model that best fit the dataset [36].

RESULTS

Optimization of an approach for the recovery of *L. intracellularis* genomes directly from clinical samples

In order to facilitate whole-genome sequencing and comparative genomics of *L. intracellularis*, we developed and optimized a shotgun metagenomic approach for sequencing DNA extracted from faecal and ileal samples. For ileum and cell-cultured samples, DNA was extracted and enriched for microbial DNA before quantification of genome copy number by quantitative PCR of the *aspA* gene. At a concentration of 3.04×10^3 genome copies per ng of DNA, we obtained a mean sequencing depth of 19×, with 38% of reads mapping to *L. intracellularis* (Table S1, Fig. S1). For faecal samples, DNA was extracted and genome copy number was quantified as before, with the *L. intracellularis* genome copy number being directly proportional to the mean sequencing depth and coverage across the reference genome PHE/MN1-00 (GeneBank GCA_000055945.1) (Fig. S2). Multiple libraries were pooled and multiplexed on a single flow cell lane, generating between 55 and 227 million reads per sample (Table S1). Samples with *L. intracellularis* copy numbers of $\geq 9 \times 10^4$ ng⁻¹ genomic DNA achieved genome coverage of >98% across the reference genome. A reference-guided *de novo* assembly approach was adapted for recovery of *L. intracellularis* genomes from each metagenome dataset. Reads were mapped against three *L. intracellularis* reference genomes – PHE/MN1-00 (GCA_000055945.1), N343 (GCA_000331715.1) and E40504 (GCA_001975945.1) – before assembly with metaSPAde to construct a consensus sequence [37]. In this manner, 24 *L. intracellularis* draft whole-genome sequences with >98% genome coverage were obtained, of which 21 were obtained directly from clinical samples and 3 were obtained from isolates propagated in McCoy cells (Table S3).

Phylogenetic analysis of *L. intracellularis* reveals a single genetically monomorphic porcine clone

To characterize the population structure of *L. intracellularis*, we analysed a total of 30 genomes comprising porcine (*n*=27) and equine (*n*=3) isolates from 7 countries across 3 continents, including the 6 publicly available genomes (Table 1). Of these, 24 had been obtained directly from clinical samples in Brazil (*n*=1), Japan (*n*=3) [38], Poland (*n*=9), the UK (*n*=9) and Sweden (*n*=2). The remaining six genomes were obtained from cell-cultured isolates originating from the USA (*n*=3), Denmark (*n*=2) and the UK (*n*=1). A multiple genome sequence alignment of 1673690 bp was produced with 6257 core genome SNPs identified outside putative recombinant regions as detected by Gubbins [33]. ML phylogenetic trees were constructed using either SNPs identified from a core-genome sequence alignment (Fig. 1) or SNPs identified from mapping short-read sequences to the core reference

**Table 1.** *L. intracellularis* isolates used in the current study

Isolate name	Country of origin	Host	Source	Year of isolation	No. CDS	Accession
5189	UK	Porcine	Cell cultured	1993	1422	SRR9841585
DKp23	Denmark	Porcine	Cell cultured	2003	1430	SRR9841584
15540	Denmark	Porcine	Cell cultured	NA	1418	SRR9866663
LR189	UK	Porcine	Ileum	1993	1416	PRDD00000000
ED	UK	Porcine	Ileum	2015	1419	SRR9866662
Thirsk2	UK	Porcine	Ileum	2017	1422	SRR9866665
630	UK	Porcine	Faecal	2016	1420	SRR9866661
682	UK	Porcine	Faecal	2016	1420	SRR9866660
SRUC1	UK	Porcine	Faecal	2016	1422	SRR9866667
SRUC3	UK	Porcine	Faecal	2016	1421	SRR9866666
1886	Poland	Porcine	Faecal	2014	1421	SRR9866664
9761	Poland	Porcine	Faecal	2014	1419	SRR9866659
661	Poland	Porcine	Faecal	2014	1417	SRR9866658
5939	Poland	Porcine	Faecal	2014	1422	SRR9866671
2746	Poland	Porcine	Faecal	2014	1421	SRR9866670
8163	Poland	Porcine	Faecal	2014	1418	SRR9866672
6073	Poland	Porcine	Faecal	2014	1419	SRR9866675
5626	Poland	Porcine	Faecal	2014	1422	SRR9866674
3387	Poland	Porcine	Faecal	2014	1424	SRR9866677
2069	Sweden	Porcine	Faecal	2003	1418	SRR9866669
4242	Sweden	Porcine	Ileum	2003	1417	SRR9866668
F22	Brazil	Porcine	Faecal	2016	1432	SRR9866654
PHE/MN1-00*	US	Porcine	Cell cultured	NA	1439	GCA_000055945
N343*	US	Porcine	Cell cultured	NA	1434	GCA_000331715
Fu*	Japan	Porcine	Ileum	NA	1411	GCA_003312285
Ni*	Japan	Porcine	Ileum	NA	1412	GCA_003312265
Ib2*	Japan	Porcine	Ileum	NA	1412	GCA_003312305
E40504*	US	Equine	Cell cultured	NA	1408	GCA_008363085
H9	UK	Equine	Faecal	2017	1416	SRR9866657
H14	UK	Equine	Faecal	2017	1414	SRR9866650

*Genome from NCBI.

NA, data not available; CDS, coding sequence.

genome isolate PHE/MN1-00. Each approach yielded an indistinguishable phylogeny comprising a single major clade that contained all 27 porcine clinical isolates, and 2 distinct equine-associated branches represented by the UK and US equine isolates, respectively.

The isolates represented in the porcine clade contained a maximum pairwise distance of 343 SNPs, indicating very limited genetic diversity. In contrast, isolates from the 2 equine clades had longer phylogenetic branches, with a distance of 3222 SNPs separating the 2 clades, suggesting greater genetic diversity. Among the porcine isolates, a total
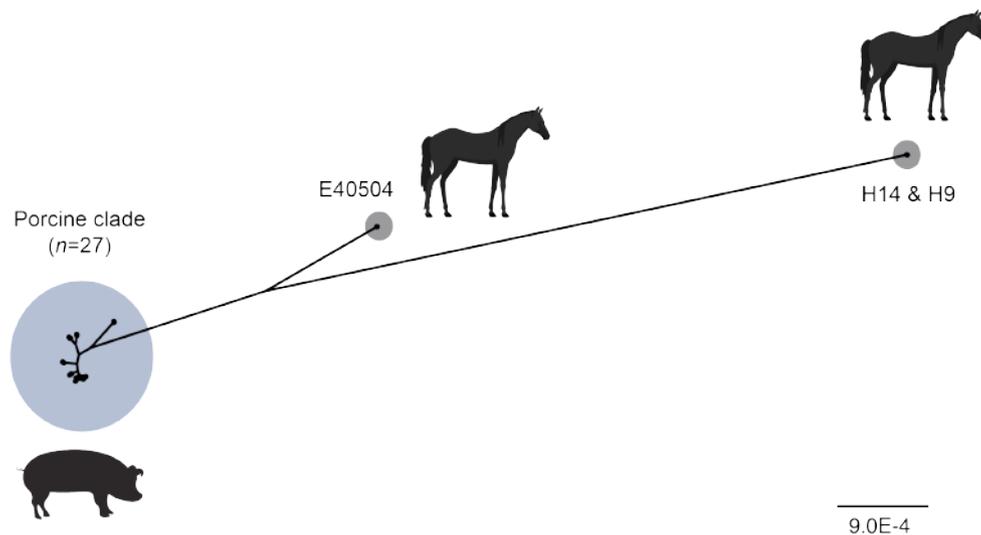
Fig. 1. Unrooted ML phylogenetic tree of *L. intracellularis*. Twenty-four genomes were generated in this study and six were obtained from the NCBI. The phylogeny was reconstructed using IQTREE based on 6257 core genome SNPs after filtering for putative recombinant sites (5260 SNPs), with the best-fitting substitution model selected by ModelFinder (TVM+F+I). The core genome was defined as the chromosomal and plasmids sequence with the prophage-associated genome island region excluded. The phylogenetic tree revealed the host-associated genetic structure of *L. intracellularis*, of which three phylogroups are formed. The 27 porcine isolates are clustered into a clonal group highlighted in blue, and the equine isolates are clustered into 2 distinct groups highlighted in grey. The scale bar represents the number of nucleotide substitutions per variable site.

of 721 polymorphic sites were identified, of which 557 SNPs were found within 414 genes, comprising 29% of the total genes. Most SNPs resulted in predicted amino acid replacements (384 non-synonymous versus 156 synonymous) and 8 nonsense mutations were identified, leading to predicted truncation or loss of function.

Within the porcine clade, three sub-lineages of porcine-derived *L. intracellularis* (Fig. 2) were identified, which we refer to as sub-lineage I, consisting of 2 isolates from Japan, and sub-lineage II, comprising a cluster of 4 UK isolates and a cluster of 4 isolates from Poland. Sub-lineage III is more geographically diverse, populated by isolates from Japan, Europe and the Americas, consistent with wider dispersal of isolates across multiple continents in comparison to the other sub-lineages, which exhibit greater geographical restriction. However, our conclusions regarding the geographical spread of these lineages are limited by the small number of isolates included in the study.

L. intracellularis has a very limited accessory genome

Pan-genomic analysis of *L. intracellularis* using Roary revealed the gene content of this pathogen to be highly conserved, with the total number of predicted genes in each genome varying from 1393 to 1411 genes, and an average of only 1.3% gene content variation among the population (Fig. S4). The number of unique gene clusters predicted across all 30 isolates was 1458, with 1374 (94.3%) identified core genes conserved in all 30 genomes (Fig. 3a). The accessory genome is made up of a combined total of 84 gene clusters (5.7% of

the pangenome), comprising 24 genes identified in at least 2 genomes and 60 strain-specific genes (Fig. 3b). Of note, 15 accessory genes were located within a previously described 18 kb prophage-associated genomic island [39], identified in 4 porcine strains DKp23, F22, N343 and PHE/MN1-00, representing a single monophyletic clade (Fig. 2). The highly conserved prophage-associated genomic island contains a *tetM* gene encoding resistance to tetracycline, a commonly used antibiotic for the treatment of PE. The island exhibits a G+C content of 60%, much higher than the average 33% G+C content of the rest of the chromosome, which is a strong indication of horizontal acquisition (Fig. S5). Examination of the remaining 66 genes identified as accessory revealed divergent gene orthologues that were distinguished from core genes due to protein truncations, in-frame deletions or amino acid sequence identity below the default BLASTP threshold of 95%. Repeating the analysis with Roary using a lower 80% BLASTP threshold reduced the number of inferred accessory genes to 37 (Fig. S4), excluding genes located within the prophage genomic-associated island.

A 111 bp in-frame deletion in the *LI_RS03480* gene encoding a putative ZIP family divalent metal cation transporter was observed in two cell-passaged isolates, including *L. intracellularis* strains 15540 (80 passages) and DKp23 (23 passages), whereas the intact form of the gene was present among all clinical isolates (Fig. S5). We observed direct repeats flanking the deleted region of *LI_RS03480* (Fig. S5), suggesting that intra-molecular recombination may have resulted in the excision of this region (Fig. S5).

Tree scale: 0.00001 ⊢⊣
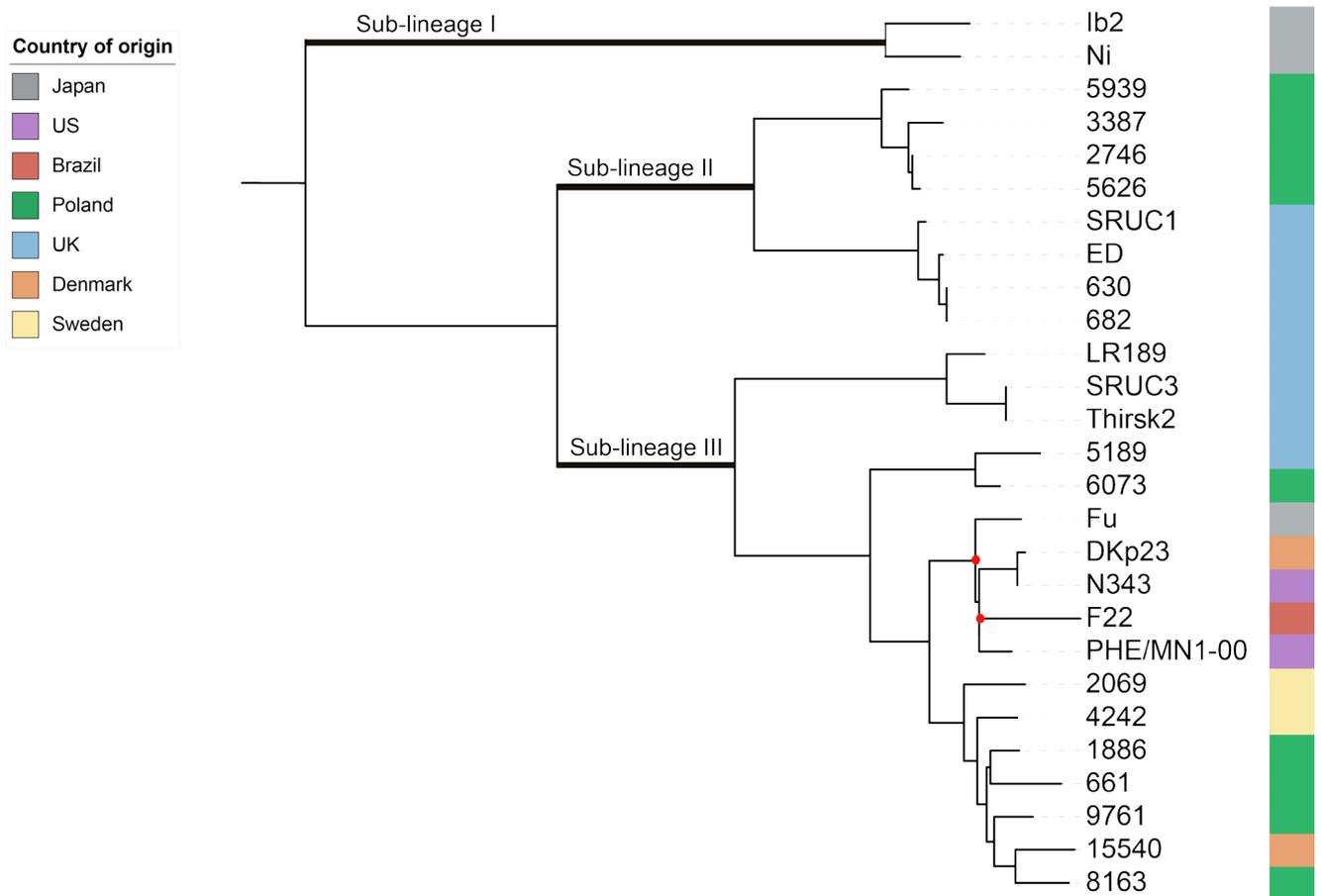
Fig. 2. High-resolution ML phylogeny of 27 *L. intracellularis* isolates from the porcine clade. The dataset was composed of isolates originated from Japan (*n*=3), US (*n*=2), Brazil (*n*=1), Poland (*n*=9), UK (*n*=8), Denmark (*n*=2) and Sweden (*n*=2). The phylogeny was estimated based on 721 SNPs, called against the core genome of reference strain PHE/MN1-00 (NCBI GenBank accession no. GCA_000055945.1). The tree was midpoint rooted and constructed using IQTREE with the best-fitting substitution model selected by ModelFinder (HKY+F+I). Phylogenetic tree branches in thick lines highlight the three major sub-lineages (sub-lineages I, II and III) formed among the isolates. All nodes displayed contained bootstrap support of >80%, with the exception of two nodes, as indicated by the red dots. The scale bar represents the number of nucleotide substitutions per variable site.

Chromosomal regions of high SNP density suggest recombination has impacted on the evolution of *L. intracallularis*

For accurate phylogenetic reconstruction, Gubbins was used to identify and remove genomic regions with a high density of polymorphisms. The resulting tree displayed the same topology, but the phylogenetic distances separating the three lineages were reduced (Fig. S3), particularly the terminal branch length of equine strain E40504 (Fig. S3). Gubbins identified multiple chromosomal regions containing high SNP density along two phylogenetic branches: (i) the branch leading to the UK equine clade from the node of diversification between the porcine clade and the E40504 isolate and (ii) the terminal branch of E40504 (Figs S6 and Fig. 4). Putative recombination events were not detected along branches within the porcine clade. In total, 28 high-SNP-density regions were identified on the branch leading to the UK equine clade,

representing 88536 bp, and 5% of the core genome. The estimated *r/m* value, indicating the relative frequency of recombination to point mutation, was 0.65, consistent with mutation being more important than recombination in genome diversification. In addition, 18 high-SNP-density regions were predicted along the E40504 terminal branch, with a total length of 100182 bp affecting 6% of its core genome, with an estimated *r/m* value of 2.0. A total of 119 genes were identified within the regions of elevated SNP density, including 19 encoding proteins with less than 95% aa sequence identity to their respective orthologues (Table S4), consistent with the pan-genome analysis. Functional annotation of these genes revealed that most (11 of 19) were hypothetical proteins with no BLASTP hits, with the remaining genes encoding putative proteins with various predicted clusters of orthologous groups (COGs) functional categories, including cell metabolism, signal transduction, membrane biogenesis, cell cycle/mitosis

A

Porcine clade (*n=27*)

H9 & H14   E40504

B

30 taxa, 100 clusters (accessory only)
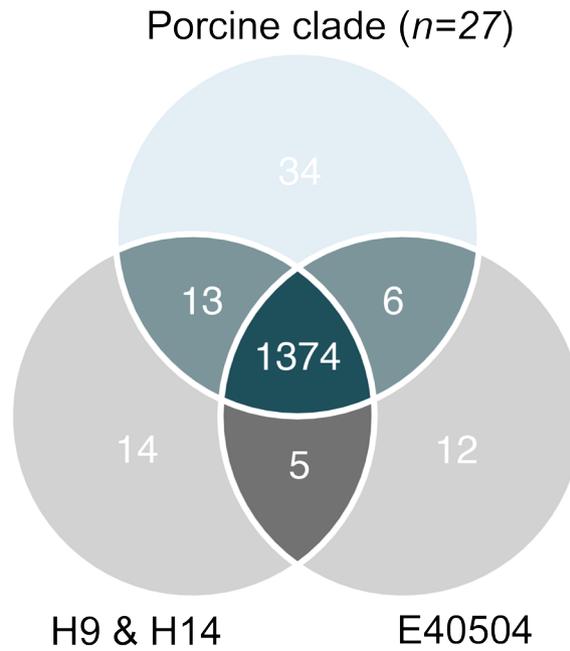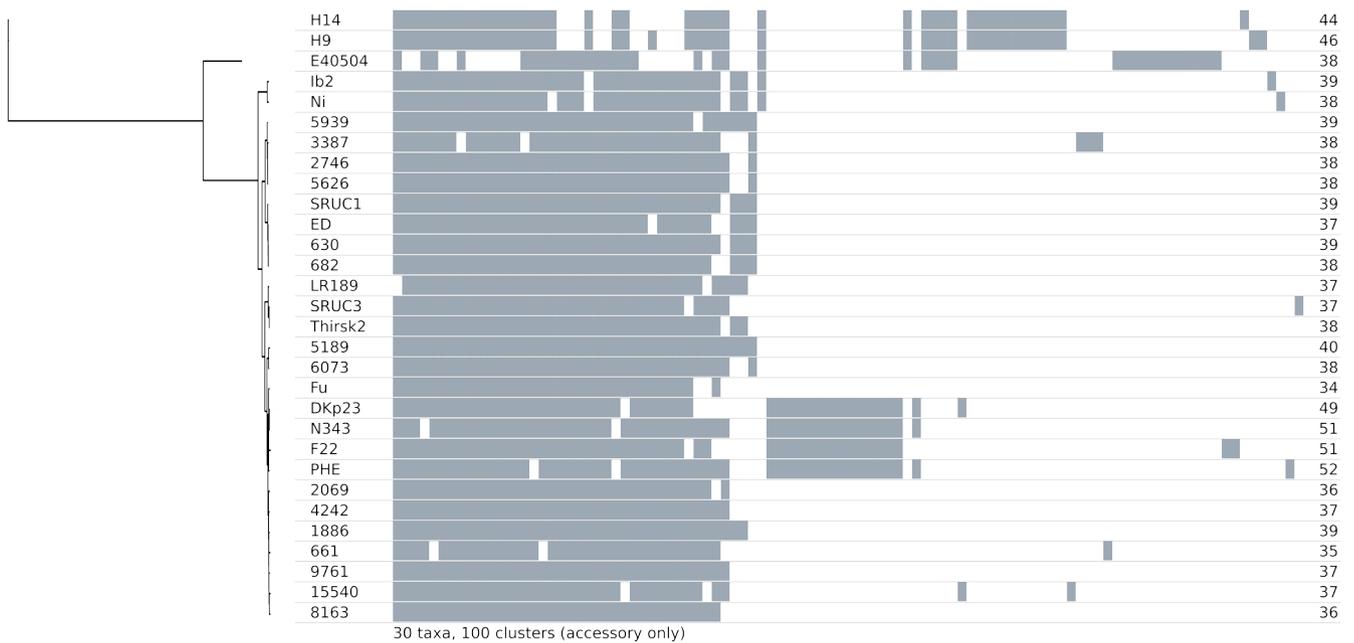
Fig. 3. Pan-genome analysis of 30 *L. intracellularis* isolates. (a) Venn diagram displaying output of Roary performed with default parameters, which identified a complement of 1374 core genes shared among all the isolates. Thirty-four, 14 and 12 clusters of genes unique to isolates within the porcine clade, the UK equine clade (isolates H14 and H9), and the US equine isolate E40504 were identified, respectively. (b) ML phylogeny based on core genome SNPs outside regions of inferred recombination (left) and distribution of accessory genes for each of the isolates (right), with the number of accessory genes stated in the column on the right.
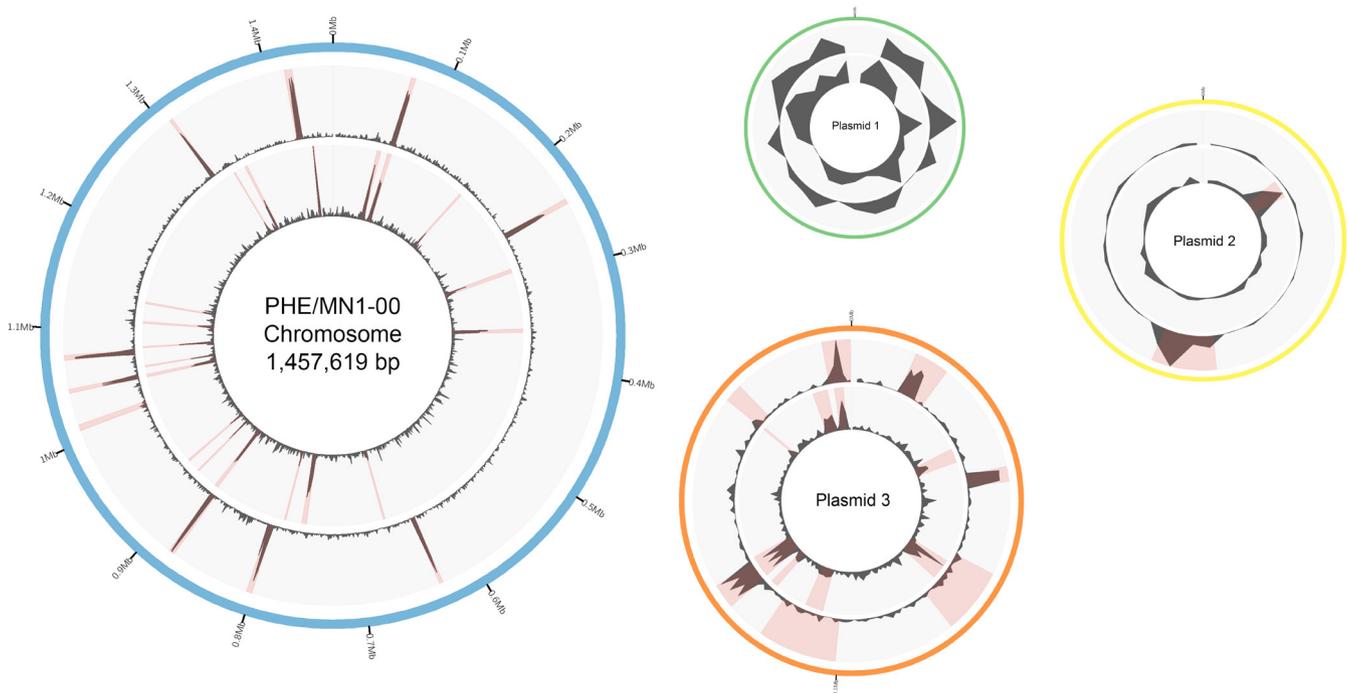
Fig. 4. Frequency of SNPs per 1000 bp across the chromosome and three plasmids of *L. intracellularis* genomes of equine-derived H14 isolate (inner grey ring) and E40504 isolate (outer grey ring). Polymorphisms were called against the core genome of porcine-derived isolate PHE/MN1-00 (NCBI GenBank accession no. GCA_000055945.1). Regions across the genomes highlighted in red represent regions of elevated diversity detected by Gubbins.

control and hypothetical proteins with unknown function (Table S4). Of note, two genes, *LI_RS063*20 and *LI_RS06315* (old locus tag LI1159 and LI1158, respectively), belong to a predicted type III secretion operon, which were previously found to be highly expressed during infection [40].

DISCUSSION

The fastidious growth requirement of *L. intracellularis* has severely restricted our capacity to examine its evolution and molecular pathogenesis. In the current work, a shotgun metagenomic sequencing approach was developed and applied to obtain *L. intracellularis* genome sequences through direct sequencing of clinical samples, including faecal and tissue samples. Our study represents a proof-of-principle study for whole-genome sequencing of *L. intracellularis* directly from clinical samples and a first view of the population structure of this major livestock pathogen. Although a limited number of isolates were included in our study, comparative genomic analysis revealed a host-associated genetic structure with isolates infecting pigs segregating to a single major clonal lineage distinct from isolates infecting horses. Cross-species experimental infection studies of pigs with the equine *L. intracellularis* E40504 could not establish disease in pigs, suggesting an equine-specific host tropism [13, 14].

Our analysis revealed remarkably low levels of intraspecies diversity in gene content. Much of the variation observed was due to the strain-dependent 18 kb prophage-associated

genomic island, identified in four closely related porcine isolates. This island was previously described by Vannucci *et al.* as a DLP12-associated island, and the authors considered the element to be defective and of limited pathogenic value, since its presence did not correlate with a more virulent phenotype [39]. Consistent with this observation, the majority of isolates sequenced in the current study, along with the three Japanese strains, were derived from clinical PE cases, but did not contain this prophage-associated island.

The highly clonal population of porcine *L. intracellularis* isolates displaying very limited genetic variation suggests that the population may have evolved from clonal expansion of a single or small number of strains. The lack of correlation between sampling times and divergence precluded robust estimation of the mutation rate and a reliable prediction of the time to the most recent common ancestor of the porcine clone. Although highly homogenous in gene content, comparative analysis of equine and porcine *L. intracellularis* isolates identified multiple regions of elevated genetic diversity, which likely resulted from the import of homologous DNA by recombination, though an elevated substitution rate due to selection acting on specific chromosomal loci cannot be ruled out. However, the population structure is clonal, as these regions only accounted for approximately 5% of the *L. intracellularis* genome, and genome diversification mainly occurs through mutation. Given the obligate intracellular lifestyle of *L. intracellularis*, the effect of genetic drift due to

a population bottleneck introduced during host transmission may have a profound limiting effect on genetic diversity [41, 42]. Functional classification based on COGs of genes located within regions of elevated SNP density predicted proteins involved in a broad range of functions. Notably, one of the regions presenting a significant excess of SNPs displaying high sequence diversification between the E40504 strain and porcine isolates corresponds to a previously reported type III secretion operon [40]. Genes within this operon are among the most highly expressed *L. intracellularis* transcripts during peak of infection and were hypothesized to be expressed in the porcine enterocyte endosome [40]. Thus, sequence variations of these genes between isolates targeting the two host species suggest a possible important role in pathogenesis, and their exact role during infection warrants further investigation.

The wide distribution of porcine-derived isolates indicates a global transmission of isolates from sub-lineage III across Europe, East Asia and the Americas. We speculate that the spread is likely linked to the expanding international livestock trade, as millions of live animals are being transported between countries [43, 44]. Previously, such events have facilitated the transmission and spread of foot-and-mouth disease in the UK and classical swine fever in the USA [45, 46]. Since porcine PE cases can be subclinical in nature, high-level pig movements can easily facilitate spread of the disease without being detected. Thus, our finding highlights the need for improved surveillance and novel control strategies for *L. intracellularis*.

The current work presents several limitations that need to be considered. First, the reliance on reference genomes during assembly could theoretically interfere with the discovery of unique strain-dependent genomic regions. However, we employed multiple different approaches, including reference-guided and reference-free assembly methods, to establish the optimal approach for genome sequence extraction and assembly, with only minor differences in output observed. Secondly, the presence of repetitive DNA sequences remains a technical challenge for short-read assembly and mapping and thus chromosomal rearrangements and structural variations may not be captured. Finally, the samples included were biased towards isolates derived from clinical disease cases, which may only represent a proportion of the population. Thus, examination of isolates from subclinical and clinically apparent cases will be required to fully assess the general population of *L. intracellularis*, and elucidate predisposing factors contributing to disease outbreak to inform better control strategies.

In the current study, we have demonstrated the potential of metagenomic sequencing to investigate the population genomics of an obligate intracellular pathogen. This has in turn provided novel insights into the genome biology of *L. intracellularis* and the first glimpse into the evolutionary history of a major bacterial pathogen of pigs and other animals. This study provides a framework for future investigations into the population biology of unculturable bacterial pathogens.

Author contributions
T. A. A., R. F. and A. L. A. conceived and designed the project. R. J .B. performed the experiments. E. W, R. M. C. G, M. J and T. S provided key clinical samples. R. J. B., B. A. W., G. Y., R. B., R. F and T.A.A. analysed and interpreted the data. R. J. B. wrote the manuscript, with major contributions by J. R. F. and T. A. A. All authors reviewed and contributed to the editing of the manuscript. R. J. B. was supervised by J. R. F. and T.A.A.

Conflicts of interest
The authors declare that there are no conflicts of interest.

Data Bibliography
1. Bengtsson RJ, SRA, PRJNA, 554776 (2019).
2. Bengtsson RJ, SRA, PRJNA, 432360 (2019).

References
1. Gebhart CJ, Barns SM, Mcorist S, Lin G-F, Lawson GH. Ileal symbiont intracellularis, an obligate intracellular bacterium of porcine intestines showing a relationship to Desulfovibrio species. *International Journal of Systematic and Evolutionary Microbiology* 1993;43:533–538.

2. Klein EC, Gebhart CJ, Duhamel GE. Fatal outbreaks of proliferative enteritis caused by Lawsonia intracellularis in young colony-raised rhesus macaques. *J Med Primatol* 1999;28:11–.

3. Lawson GHK, Gebhart CJ. Proliferative enteropathy. *J Comp Pathol* 2000;122:77–100.

4. Pusterla N, Mapes S, Rejmanek D, Gebhart C. Detection of *Lawsonia intracellularis* by real-time PCR in the feces of free-living animals from equine farms with documented occurrence of equine proliferative enteropathy. *J Wildl Dis* 2008;44:992–.

5. Murakata K, Sato A, Yoshiya M, Kim S, Watarai M *et al*. Infection of different strains of mice with *Lawsonia intracellularis* derived from rabbit or porcine proliferative enteropathy. *J Comp Pathol* 2008;139:8–15.

6. Rowland A, Rowntree P. A haemorrhagic bowel syndrome associated with intestinal adenomatosis in the pig. *Veterinary Record* 1972;91:235–241.

7. Shimizu C, Shibahara T, Takai S, Kasuya K, Chikuba T *et al*. *Lawsonia intracellularis* and virulent *Rhodococcus equi* infection in a thoroughbred Colt. *J Comp Pathol* 2010;143:303–308.

8. Van Den Wollenberg L, Butler C, Houwers D, de Grootv M, Panhuijzen H *et al*. *Lawsonia intracellularis*-associated proliferative enteritis in weanling foals in the Netherlands. *Tijdschr Diergeneeskd* 2011;136:565–570.

9. Pusterla N, Gebhart CJ. Equine proliferative enteropathy - a review of recent developments. *Equine Vet J* 2013;45:403–409.

10. Pusterla N, Gebhart C. Equine proliferative enteropathy caused by *Lawsonia intracellularis*. *Equine Vet Educ* 2009;21:415–419.

11. Lavoie JP, Drolet R, Parsons D, Leguillette R, Sauvageau R *et al*. Equine proliferative enteropathy: a cause of weight loss, colic, diarrhoea and hypoproteinaemia in foals on three breeding farms in Canada. *Equine Vet J* 2000;32:418–425.

12. Pusterla N, Wattanaphansak S, Mapes S, Collier J, Hill J *et al*. Oral infection of weanling foals with an equine isolate of *Lawsonia intracellularis*, agent of equine proliferative enteropathy. *J Vet Intern Med* 2010;24:622–627.

13. Vannucci FA, Pusterla N, Mapes SM, Gebhart C. Evidence of host adaptation in *Lawsonia intracellularis* infections. *Vet Res* 2012;43:53.

14. Sampieri F, Vannucci FA, Allen AL, Pusterla N, Antonopoulos AJ *et al*. Species-Specificity of equine and porcine Lawsonia

intracellularis isolates in laboratory animals. *Canadian Journal of Veterinary Research* 2013;77:261–272.

15. Lawson GH, McOrist S, Jasni S, Mackie RA. Intracellular bacteria of porcine proliferative enteropathy: cultivation and maintenance *in vitro*. *J Clin Microbiol* 1993;31:1136–1142.

16. Vannucci FA, Wattanaphansak S, Gebhart CJ. An alternative method for cultivation of *Lawsonia intracellularis*. *J Clin Microbiol* 2012;50:1070–1072.

17. Sait M, Aitchison K, Wheelhouse N, Wilson K, Lainson FA *et al*. Genome sequence of *Lawsonia intracellularis* strain N343, isolated from a sow with hemorrhagic proliferative enteropathy. *Genome Announc* 2013;1:e00027–13.

18. Jacobson M, Aspan A, Königsson MH, Segerstad CHaf, Wallgren P *et al*. Routine diagnostics of *Lawsonia intracellularis* performed by PCR, serological and post mortem examination, with special emphasis on sample preparation methods for PCR. *Vet Microbiol* 2004;102:189–201.

19. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–2120.

20. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014;15:R46.

21. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagen-omics assembly via succinct de Bruijn graph. *Bioinformatics* 2015;31:1674–1676.

22. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754–1760.

23. Kang DD, Froula J, Egan R, Wang Z, MetaBAT WZ. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 2015;3:e1165.

24. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015;25:1043–1055.

25. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M *et al*. SPAdes: a new genome assembly algorithm and its applica-tions to single-cell sequencing. *Journal of Computational Biology* 2012;19:455–477.

26. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013;29:1072–1075.

27. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioin-formatics* 2014;30:2068–2069.

28. 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.

29. Seemann T. snippy: fast bacterial variant calling from NGS reads 2015.

30. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* 2015;32:btv566–4.

31. 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.

32. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* 2012;28:464–469.

33. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA *et al*. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* 2015;43:e15–e..

34. Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM *et al*. Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics* 2018;34:292–293.

35. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015;32:268–274.

36. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylo-genetic estimates. *Nat Methods* 2017;14:587–.

37. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaS-PAdes: a new versatile metagenomic assembler. *Genome Res* 2017;27:824–834.

38. Nishikawa S, Ogawa Y, Eguchi M, Rambukkana A, Shimoji Y. Draft genome sequences of *Lawsonia intracellularis* swine strains causing proliferative enteropathy in Japan. *Microbiol Resour Announc* 2018;7:e01021–18.

39. Vannucci FA, Kelley MR, Gebhart CJ. Comparative genome sequencing identifies a prophage-associated genomic island linked to host adaptation of *Lawsonia intracellularis* infections. *Vet Res* 2013;44:49.

40. Vannucci FA, Foster DN, Gebhart CJ. Laser microdissection coupled with RNA-seq analysis of porcine enterocytes infected with an obligate intracellular pathogen (*Lawsonia intracellularis*). *BMC Genomics* 2013;14:421.

41. Achtman M, Wagner M. Microbial diversity and the genetic nature of microbial species. *Nat Rev Microbiol* 2008;6:431–.

42. Smith NH, Gordon SV, de la Rua-Domenech R, Clifton-Hadley RS, Hewinson RG. Bottlenecks and broomsticks: the molecular evolu-tion of *Mycobacterium bovis*. *Nat Rev Microbiol* 2006;4:670–.

43. Fèvre EM, Bronsvoort BMdeC, Hamilton KA, Cleaveland S. Animal movements and the spread of infectious diseases. *Trends Microbiol* 2006;14:125–131.

44. Balka G, Podgórska K, Brar MS, Bálint Ádám, Cadar D *et al*. Genetic diversity of PRRSV 1 in central eastern Europe in 1994–2014: origin and evolution of the virus in the region. *Sci Rep* 2018;8:7811.

45. Gibbens JC, Wilesmith JW, Sharpe CE, Mansley LM, Michalo-poulou E *et al*. Descriptive epidemiology of the 2001 foot-and-mouth disease epidemic in Great Britain: the first five months. *Veterinary Record* 2001;149:729–743.

46. Brown VR, Bevins SN. A review of classical swine fever virus and routes of introduction into the United States and the potential for virus establishment. *Frontiers in Veterinary Science* 2018;5:31.