Phylogeographic separation and formation of sexually discrete lineages in a global population of *Yersinia pseudotuberculosis*

Tristan Seecharran,1 Laura Kalin-Manttari,2 Katja Koskela,3 Simo Nikkari,3 Benjamin Dickins,1 Jukka Corander,4 Mikael Skurnik2 and Alan McNally5,*

**Abstract**

*Yersinia pseudotuberculosis* is a Gram-negative intestinal pathogen of humans and has been responsible for several nationwide gastrointestinal outbreaks. Large-scale population genomic studies have been performed on the other human pathogenic species of the genus *Yersinia*, *Yersinia pestis* and *Yersinia enterocolitica* allowing a high-resolution understanding of the ecology, evolution and dissemination of these pathogens. However, to date no purpose-designed large-scale global population genomic analysis of *Y. pseudotuberculosis* has been performed. Here we present analyses of the genomes of 134 strains of *Y. pseudotuberculosis* isolated from around the world, from multiple ecosystems since the 1960s. Our data display a phylogeographic split within the population, with an Asian ancestry and subsequent dispersal of successful clonal lineages into Europe and the rest of the world. These lineages can be differentiated by CRISPR cluster arrays, and we show that the lineages are limited with respect to inter-lineage genetic exchange. This restriction of genetic exchange maintains the discrete lineage structure in the population despite co-existence of lineages for thousands of years in multiple countries. Our data highlights how CRISPR can be informative of the evolutionary trajectory of bacterial lineages, and merits further study across bacteria.

**DATA SUMMARY**

All of the raw sequence data for this project has been deposited into the European Nucleotide Archive. Individual accession numbers for each pair of fastq files are indicated in Table S1 (available in the online Supplementary Material). Additionally de novo assemblies of all genomes used are available in Enterobase (https://enterobase.warwick.ac.uk/species/index/yersinia), searchable by the strain name allocated in Table S1.

**INTRODUCTION**

The genus *Yersinia* belongs to the Gram-negative bacterial family *Enterobacteriaceae*, and is a model genus for studying the evolution of bacterial pathogens [1]. Three species of *Yersinia* are well-recognised human pathogens: the plague bacillus *Yersinia pestis*, and the enteropathogenic *Yersinia pseudotuberculosis* and *Yersinia enterocolitica* [1]. *Y. pseudotuberculosis*, which causes infection in a broad range of hosts, including domesticated and wild animals, has also been associated with foodborne infection in humans – known as yersiniosis. Transmission of the bacterium is usually through the faecal–oral route, and human infection can result from the ingestion of contaminated food products or water, or otherwise by direct contact with an infected animal or human [2–5]. *Y. pseudotuberculosis* is also found widely in the environment, including soil [6], and in animals it causes a tuberculosis-like disease [6]. Human cases of *Y. pseudotuberculosis* infections are usually sporadic, however several large outbreaks have been reported in Finland and recently in New Zealand [7, 8]. Classical identification and typing of *Y. pseudotuberculosis* is based on the lipopolysaccharide O-antigen, resulting in a total of 21 known serotypes [9]. However the efficacy of serotyping is very limited due to a large proportion of strains belonging to serotypes O:1a, O:1b and O:3 [8, 10].

The population structure of *Y. pseudotuberculosis* has been elucidated by multi-locus sequence typing (MLST) [10]. This added further granularity to the serotype
differences, grouping serotype O:3 strains into a distinct
clonal group designated ST19 which are characterized by a
truncation in the Yersiniabactin locus resulting in the loss of the
genes encoding iron transport across the bacterial mem-
brane. Serotype O:1 strains formed a distinct clade of
strains encompassing a large number of sequence type com-
plexes, suggesting a highly diverse population of bacteria
within the serotype O:1 group of Y. pseudotuberculosis. In
addition to MLST genotyping, recent work has also analysed
clustered regularly interspaced short palindromic repeat
(CRISPR) loci of 335 isolates of Y. pseudotuberculosis [11].
The CRISPR-Cas system is an RNA-based immune system
that regulates invasion of plasmids and viruses in bacteria
and archaea [12]. CRISPRs are constructed from a chain of
21 to 47 bp repeated sequences called direct repeats (DR),
in which DRs are unique spacer sequences. These
spacers represent foreign DNA originating predominantly
from bacteriophages and plasmids. In Y. pseudotuberculosis,
the CRISPR spacers are stored in three genomic loci named
YP1, YP2 and YP3, and we identified 1902 distinct spacer
sequences, and that

Y. pestis

carry a relatively low number of
spacers compared with Y. pseudotuberculosis.

To date the most comprehensive genome-scale analysis of
Y. pseudotuberculosis centred around a country-wide out-
break in New Zealand [8]. Incorporation of publicly avail-
able genomes into this dataset also suggested a highly
diverse species and that the New Zealand strains repres-
ented a geographically isolated clade of Y. pseudotuberculo-
sis. The paucity of a specifically designed, geographically
and temporally distributed dataset of Y. pseudotuberculosis
gene sequences of the YP1, YP2 and YP3, and we identified 1902 distinct spacer
sequences [11]. One central finding was that strains of Y.
pseudotuberculosis and strains of Y. pestis shared very few
spacers, and that Y. pestis carries a relatively low number of
spacers compared with Y. pseudotuberculosis.

To date the most comprehensive genome-scale analysis of
Y. pseudotuberculosis centred around a country-wide out-
break in New Zealand [8]. Incorporation of publicly avail-
able genomes into this dataset also suggested a highly
diverse species and that the New Zealand strains repres-
ented a geographically isolated clade of Y. pseudotuberculo-
sis. The paucity of a specifically designed, geographically
and temporally distributed dataset of Y. pseudotuberculosis
gene sequences of the YP1, YP2 and YP3, and we identified 1902 distinct spacer
sequences [11]. One central finding was that strains of Y.
pseudotuberculosis and strains of Y. pestis shared very few
spacers, and that Y. pestis carries a relatively low number of
spacers compared with Y. pseudotuberculosis.

To date the most comprehensive genome-scale analysis of
Y. pseudotuberculosis centred around a country-wide out-
break in New Zealand [8]. Incorporation of publicly avail-
able genomes into this dataset also suggested a highly
diverse species and that the New Zealand strains repres-
ented a geographically isolated clade of Y. pseudotuberculo-
sis. The paucity of a specifically designed, geographically
and temporally distributed dataset of Y. pseudotuberculosis
gene sequences of the YP1, YP2 and YP3, and we identified 1902 distinct spacer
sequences [11]. One central finding was that strains of Y.
pseudotuberculosis and strains of Y. pestis shared very few
spacers, and that Y. pestis carries a relatively low number of
spacers compared with Y. pseudotuberculosis.

To date the most comprehensive genome-scale analysis of
Y. pseudotuberculosis centred around a country-wide out-
break in New Zealand [8]. Incorporation of publicly avail-
able genomes into this dataset also suggested a highly
diverse species and that the New Zealand strains repres-
ented a geographically isolated clade of Y. pseudotuberculo-
sis. The paucity of a specifically designed, geographically
and temporally distributed dataset of Y. pseudotuberculosis
gene sequences of the YP1, YP2 and YP3, and we identified 1902 distinct spacer
sequences [11]. One central finding was that strains of Y.
pseudotuberculosis and strains of Y. pestis shared very few
spacers, and that Y. pestis carries a relatively low number of
spacers compared with Y. pseudotuberculosis.

To date the most comprehensive genome-scale analysis of
Y. pseudotuberculosis centred around a country-wide out-
break in New Zealand [8]. Incorporation of publicly avail-
able genomes into this dataset also suggested a highly
diverse species and that the New Zealand strains repres-
ented a geographically isolated clade of Y. pseudotuberculo-
sis. The paucity of a specifically designed, geographically
and temporally distributed dataset of Y. pseudotuberculosis
gene sequences of the YP1, YP2 and YP3, and we identified 1902 distinct spacer
sequences [11]. One central finding was that strains of Y.
pseudotuberculosis and strains of Y. pestis shared very few
spacers, and that Y. pestis carries a relatively low number of
spacers compared with Y. pseudotuberculosis.

To date the most comprehensive genome-scale analysis of
Y. pseudotuberculosis centred around a country-wide out-
break in New Zealand [8]. Incorporation of publicly avail-
able genomes into this dataset also suggested a highly
diverse species and that the New Zealand strains repres-
ented a geographically isolated clade of Y. pseudotuberculo-
sis. The paucity of a specifically designed, geographically
and temporally distributed dataset of Y. pseudotuberculosis
gene sequences of the YP1, YP2 and YP3, and we identified 1902 distinct spacer
sequences [11]. One central finding was that strains of Y.
pseudotuberculosis and strains of Y. pestis shared very few
spacers, and that Y. pestis carries a relatively low number of
spacers compared with Y. pseudotuberculosis.

To date the most comprehensive genome-scale analysis of
Y. pseudotuberculosis centred around a country-wide out-
break in New Zealand [8]. Incorporation of publicly avail-
able genomes into this dataset also suggested a highly
diverse species and that the New Zealand strains repres-
ented a geographically isolated clade of Y. pseudotuberculo-
sis. The paucity of a specifically designed, geographically
and temporally distributed dataset of Y. pseudotuberculosis
gene sequences of the YP1, YP2 and YP3, and we identified 1902 distinct spacer
sequences [11]. One central finding was that strains of Y.
pseudotuberculosis and strains of Y. pestis shared very few
spacers, and that Y. pestis carries a relatively low number of
spacers compared with Y. pseudotuberculosis.

To date the most comprehensive genome-scale analysis of
Y. pseudotuberculosis centred around a country-wide out-
break in New Zealand [8]. Incorporation of publicly avail-
able genomes into this dataset also suggested a highly
diverse species and that the New Zealand strains repres-
ented a geographically isolated clade of Y. pseudotuberculo-
sis. The paucity of a specifically designed, geographically
and temporally distributed dataset of Y. pseudotuberculosis
gene sequences of the YP1, YP2 and YP3, and we identified 1902 distinct spacer
sequences [11]. One central finding was that strains of Y.
pseudotuberculosis and strains of Y. pestis shared very few
spacers, and that Y. pestis carries a relatively low number of
spacers compared with Y. pseudotuberculosis.

To date the most comprehensive genome-scale analysis of
Y. pseudotuberculosis centred around a country-wide out-
break in New Zealand [8]. Incorporation of publicly avail-
able genomes into this dataset also suggested a highly
diverse species and that the New Zealand strains repres-
ented a geographically isolated clade of Y. pseudotuberculo-
sis. The paucity of a specifically designed, geographically
and temporally distributed dataset of Y. pseudotuberculosis
gene sequences of the YP1, YP2 and YP3, and we identified 1902 distinct spacer
sequences [11]. One central finding was that strains of Y.
pseudotuberculosis and strains of Y. pestis shared very few
spacers, and that Y. pestis carries a relatively low number of
spacers compared with Y. pseudotuberculosis.

To date the most comprehensive genome-scale analysis of
Y. pseudotuberculosis centred around a country-wide out-
break in New Zealand [8]. Incorporation of publicly avail-
able genomes into this dataset also suggested a highly
diverse species and that the New Zealand strains repres-
ented a geographically isolated clade of Y. pseudotuberculo-
sis. The paucity of a specifically designed, geographically
and temporally distributed dataset of Y. pseudotuberculosis
gene sequences of the YP1, YP2 and YP3, and we identified 1902 distinct spacer
sequences [11]. One central finding was that strains of Y.
pseudotuberculosis and strains of Y. pestis shared very few
spacers, and that Y. pestis carries a relatively low number of
spacers compared with Y. pseudotuberculosis.

To date the most comprehensive genome-scale analysis of
Y. pseudotuberculosis centred around a country-wide out-
break in New Zealand [8]. Incorporation of publicly avail-
able genomes into this dataset also suggested a highly
diverse species and that the New Zealand strains repres-
ented a geographically isolated clade of Y. pseudotuberculo-
sis. The paucity of a specifically designed, geographically
and temporally distributed dataset of Y. pseudotuberculosis
gene sequences of the YP1, YP2 and YP3, and we identified 1902 distinct spacer
sequences [11]. One central finding was that strains of Y.
pseudotuberculosis and strains of Y. pestis shared very few
spacers, and that Y. pestis carries a relatively low number of
spacers compared with Y. pseudotuberculosis.

To date the most comprehensive genome-scale analysis of
Y. pseudotuberculosis centred around a country-wide out-
break in New Zealand [8]. Incorporation of publicly avail-
able genomes into this dataset also suggested a highly
diverse species and that the New Zealand strains repres-
ented a geographically isolated clade of Y. pseudotuberculo-
sis. The paucity of a specifically designed, geographically
and temporally distributed dataset of Y. pseudotuberculosis
gene sequences of the YP1, YP2 and YP3, and we identified 1902 distinct spacer
sequences [11]. One central finding was that strains of Y.
pseudotuberculosis and strains of Y. pestis shared very few
spacers, and that Y. pestis carries a relatively low number of
spacers compared with Y. pseudotuberculosis.

To date the most comprehensive genome-scale analysis of
Y. pseudotuberculosis centred around a country-wide out-
break in New Zealand [8]. Incorporation of publicly avail-
able genomes into this dataset also suggested a highly
diverse species and that the New Zealand strains repres-
ented a geographically isolated clade of Y. pseudotuberculo-
sis. The paucity of a specifically designed, geographically
and temporally distributed dataset of Y. pseudotuberculosis
gene sequences of the YP1, YP2 and YP3, and we identified 1902 distinct spacer
sequences [11]. One central finding was that strains of Y.
pseudotuberculosis and strains of Y. pestis shared very few
spacers, and that Y. pestis carries a relatively low number of
spacers compared with Y. pseudotuberculosis.
Accessory genome analysis

The large-scale blast score ratio (LS-BSR) pipeline [21] was used to create pan-genomes from genome assemblies of all strains. The included post-matrix script (filter_BSR_variome.py) was run to isolate the accessory genomes from the pangenomes. The resulting accessory genome matrix was then transposed according to the order of the strains on the phylogenetic tree. The output was used to visualize the presence or absence of all accessory genes in each individual genome by generating a heat map using the ggplot2 package of the R statistical software. Genes with >90% prevalence and also those found in fewer than five strains were excluded from this analysis. The included Python script comparee_BSR.py from LS-BSR was used to look for unique coding sequences (CDSs) between two defined populations in the pangenome matrix. Comparisons were made between the 'European' clade of strains and the 'Asian' clade, as well as between each CRISPR cluster and the rest of the population. Any unique CDSs detected were compared to the non-redundant nucleotide database using nucleotide BLAST (http://blast.ncbi.nlm.nih.gov/) to determine the genes they encode.

KPAX2 software was used to cluster the strains on the basis of their CRISPR spacer profiles [22]. Input to the software was a binary matrix with columns representing an absence/presence variable for each of the 2969 spacers in each detected CRISPR cassette. KPAX2 was used with default prior hyperparameters and an upper bound for the number of clusters equal to 50. Five independent runs of the inference algorithm were performed and the clustering solution with the highest posterior probability was chosen. All estimation runs converged to a number of clusters well below the chosen upper bound, indicating that it was sufficiently large to accommodate the region of high posterior density. To analyse the association between CRISPR spacer patterns and the accessory genome content, we calculated an average accessory genome dissimilarity (Hamming distance normalized by the number of CRISPR spacers) matrix for all detected CRISPR clusters with >1 strain (18 clusters).

To assess the significance of the observed dissimilarity pattern, we used a standard permutation test. Under the null hypothesis of no association between CRISPR cluster and the accessory genome content, the cluster label of a strain can be permuted randomly. For each of 10 000 random permutations of the labels we then re-calculated the average dissimilarity for each cluster and recorded how often the observed value is smaller than the observed dissimilarity in the original data matrix. Under a global significance level of 5%, 12/18 CRISPR clusters had a significantly smaller average distance than expected under the null hypothesis.

Detection of core genome recombination events

Core genome alignments were constructed using Parsnp [18]. Core genome recombination events were detected by performing BratNextGen analysis on the core genome alignment [23]. BratNextGen was run using the default prior settings, 20 iterations of the HMM estimation algorithm and 100 runs executed in parallel for the permutation test of significance at the 5% level.

Dating analysis

To date the geographic split within the species, and the formation of the distinct CRISPR clusters, we used BEAST [24]. The core genome alignment for the 73 strains for which isolation dates were known was obtained using Parsnp, was stripped of recombination detected using BratNextGen, and the resulting alignment was used as input with all known dates of isolation to date individual taxa. By assessing ESS scores for priors, the following parameters were chosen for the best fitting model: HKY model of substitution with estimated base frequencies and a relaxed molecular clock. The analysis was run for a total of 50 million iterations with the initial 5 million used as burn-in. From this a maximum clade credibility tree was inferred and visualized in Figtree. For the skyline analysis a stepwise constant variant was selected with the age of youngest tip set to zero.

RESULTS

Phylogeographic structure of *Y. pseudotuberculosis* signals an Asian ancestry

A maximum-likelihood phylogeny was reconstructed from a core genome alignment of 134 *Y. pseudotuberculosis* genome sequences (Fig. 1). The phylogeny has a clear two-clade structure with a seemingly ancestral clade containing high diversity and long branch lengths, and a second clade containing much lower levels of diversity. Annotation of the tree with geographical source of isolation identifies a very clear geographic split in the phylogeny, with the ancestral highly diverse clade containing primarily Asian isolates and the second low diversity clade containing primarily European isolates. Between these two clades is a small transitional cluster of isolates originating from South Africa, North and South America. Such a phylogenetic structure is consistent with an Asian ancestry for *Y. pseudotuberculosis*, with two separate migrations into Africa and the Americas, and into Europe. A European migration is consistent with a bottleneck event leading to successful establishment of a small number of clones. Annotation of the phylogeny with serotype and host species (Fig. S1) identified that the European clade is further split into serotype 1a and serotype 1b clusters, and that there is no pattern of phylogenetic grouping associated with host.

Phylogenetic dating suggests recent geographical divergence

Of the 134 genomes sequenced, isolation dates are available for 73 isolates which represent the full diversity of the phylogeny. To date the evolutionary split of the 'European' clade of strains from the 'Asian' clade we used BEAST, analyzing only the 73 strains for which isolation dates were available (Fig. S2). The analysis suggests a time to most recent common ancestor (tMRCA) for the dataset of 33 591 years before present [95% confidence interval (CI) 49 460–20 476], with the divergence of the European and Asian
clades occurring approximately 12 500 years ago (95% CI 18 500–6625). This period marks the transition between
the Neolithic and Mesolithic eras at the end of the last ice age,
and the beginning of livestock domestication and wheat and barley farming. A Bayesian skyline plot analysis of the dated
phylogeny also supports the possibility of a strong bottle-
neck occurring in the population within the European clade,
occurring in the very recent past (Fig. S2).

**Phylogenetic clusters within *Y. pseudotuberculosis* associate with discrete CRISPR cassette patterns**

We sought to determine any obvious genotypic traits associated with the phylogeographic split in our population. Bayesian clustering of the presence/absence of all 2969 known *Y. pseudotuberculosis* CRISPR spacers present in the dataset identified a total of 33 distinct clusters of CRISPR spacer cassettes (Table S1). Annotation of these CRISPR clusters on the phylogenetic tree shows that the clusters form phylogenetically distinct groups of *Y. pseudotuberculosis* (Fig. 2). The most recent of these clusters has a tMRCA to the rest of the population of 5222 years (95% CI 7676–2768) before present, suggesting that this clustering is not a recent phenomenon nor is it due to any temporal artefacts of sampling. Indeed comparison of the CRISPR cluster pattern to year and geographical source of isolation (Table S1) suggests that this clustering is not a result of strains isolated in the same short time span or localized source. To confirm this, we mapped the geographical source of isolation of all of the CRISPR clusters (Fig. 3). This shows that CRISPR clusters are widely distributed across the world with some correlation to the higher phylogeographic split observed earlier. It also shows the highest diversity in CRISPR clusters occurs in Asia consistent with an Asian ancestry of *Y. pseudotuberculosis*.

**Phylogenetic clusters are associated with patterns of genetic recombination in *Y. pseudotuberculosis***

Given the role of CRISPR in generating lasting immunity to foreign DNA, we sought to investigate if the phylogenetic clusters within *Y. pseudotuberculosis* were associated with any signature of gene sharing. We created a pangenome matrix for all 134 genomes using LS-BSR, and then extracted the accessory genome. This accessory genome matrix was then used to annotate the core phylogenetic tree alongside CRISPR clusters (Fig. 4). There are clear patterns of accessory genome profile concordant with the pattern of CRISPR clusters within the tree. Attempts to identify genes unique to any given phylogenetic cluster were largely
Fig. 2. Phylogeny of 134 *Y. pseudotuberculosis* isolates annotated with CRISPR clusters as determined by Bayesian clustering of concatenated sequence of CRISPR spacer arrays. The 'Asian' and 'European' clades are defined by tree branch colouring.
unsuccessful. Only one unique coding sequence was detected in the ‘European’ clade of strains and no unique CDSs in the ‘Asian’ clade of strains. Of CRISPR clusters that contained more than one representative strain, CRISPR 22 contained nine unique CDSs. Other clusters include CRISPR 28 with two unique CDSs, and CRISPRs 1, 9 and 11 each with one unique CDS, when compared to the rest of the population. This analysis suggests that each cluster contains a unique combination of accessory genes rather than unique genes per se. To confirm this, we calculated the average accessory genome dissimilarity for all CRISPR clusters containing more than one strain. This showed that in 12 of 18 clusters, strains have significantly more similar gene profiles to strains in the same cluster than to strains in other clusters ($P<0.05$ based on a 10 000 random permutations test). This suggests that gene sharing between strains in the Y. pseudotuberculosis population is largely restricted to within individual phylogenetic clades. Analysis of core genome recombination identified a distribution of core genome recombination events which is highly concordant with the CRISPR clusters (Fig. 5). Despite very high levels of recombination across the dataset, the recombination occurring is not eroding the phylogenetic or CRISPR cluster signal, suggesting that inter-cluster horizontal transfer of genetic material is largely inhibited or occurs at very low frequency compared to intra-cluster recombination events.

**DISCUSSION**

The genus *Yersinia* has acted as a model for developing our understanding of microbial pathogenesis, molecular microbiology and microbial evolution [1]. *Yersinia* was the first bacterial genus to have all representative species sequenced allowing fine-scale analysis of how pathogenesis evolved in the three human pathogenic members of the genus [15]. This analysis showed a striking degree of parallelism in how human-pathogenesis evolved in the pathogenic *Yersinia* [15]. However, finer scale evolutionary genomics studies of *Y. pestis* and *Y. enterocolitica* have shown very divergent mechanisms of intra-species evolution. *Y. pestis* is a recently evolved clone of *Y. pseudotuberculosis* which is globally disseminated and host-restricted with very low levels of diversity allowing fine-scale transmission events to be successfully reconstructed [13]. In complete contrast to this, pathogenic *Y. enterocolitica* have evolved from a non-pathogenic ancestor and have split into ecologically distinct clades which move rapidly across host species [14].

By sequencing the genomes of a globally and temporally distributed set of 134 isolates of *Y. pseudotuberculosis* from a wide range of hosts and environments, we show that evolution in this species is driven by completely different mechanisms again. Our data show that *Y. pseudotuberculosis* is unique between the enteropathogenic species of the genus *Yersinia* in that it shows a clear phylogeographic split in its population. This was once postulated to be the case for *Y. enterocolitica* [25] with Old World and New World strains, however comprehensive population genomics have shown this is not the case [14, 15]. The discovery of Asian ancestry in *Y. pseudotuberculosis* is in line with the postulated ancestry of *Y. pestis* [13, 26], though our data appear to show the most outstanding genetic variation occurring in Japan, not China. Whilst this does not appear to be an artefact of sampling.

**Fig. 3.** Global map showing the geographical sources of isolation of strains belonging to each of the 33 identified CRISPR clusters. The map shows clear co-existence of multiple CRISPR cluster-type strains in a number of countries.
in this study it cannot be discounted that a more thorough and dense genomic sampling would provide a different result. However, of interest is the fact that a sub-clade of *Y. pseudotuberculosis* exists which causes Far East scarlet-like fever and is associated with Japan and tropical South-East Asia [27], suggesting larger variation in this region and a potential focus of ancestry for the species. Although accurate dating from a relatively small sample set is difficult, our tMRCA for the entire *Y. pseudotuberculosis* dataset is in the same range (10 000–40 000 years ago) as that postulated for the emergence of *Y. pestis* [26], and it is tempting to speculate that this emergence coincided with a larger population and dispersal event across *Y. pseudotuberculosis*.

Previous work by our group analysed patterns of accessory gene sharing to conclude that *Y. enterocolitica* was formed of ecologically distinct phylogroups [14]. This hypothesis was formed on the basis that the limited inter-clade sharing of genes could not be due to stearic hindrance by O-antigen nor be genetic exclusion as no such mechanisms existed. Our data on *Y. pseudotuberculosis* also identify clearly distinct phylogroups within the phylogeographic clades. These phylogroups have unique combinations of accessory genes with little variation in the accessory genome, and a very similar pattern of core genome homologous recombination. Similar to *Y. enterocolitica*, it is highly improbable that this might be driven by some factor which precludes physical contact given the limited serotypes present in *Y. pseudotuberculosis* [4]. Rather our analysis is strongly indicative of genetic restriction between phylogroups, and that this can be correlated with patterns of acquired CRISPR cassettes. The primary evidence for this genetic restriction is in the fact that different phylogroups co-exist in the same geographical locations. In the absence of any active barrier to recombination, one would expect the signal that identifies each CRISPR cluster to be eroded relatively quickly in time [28], resulting in a lack of clear phylogroup signatures [29]. This would be especially so given the large levels of recombination detected in the core genome of *Y. pseudotuberculosis*. As the phylogenetic clusters have co-existed in locations for around 5000 years or more and continue to display a clear signal of within cluster similarity, our data suggest very limited genetic exchange.
between clusters and maintenance of the distinct phylogroups.

CRISPR has been shown to play a role in dictating the accessory genomes of *Pseudomonas aeruginosa* [30], and CRISPR analysis correlated with phylogenetic structure in a study of *Shigella* genomes [31]. Previous data has shown that CRISPR evolution in bacteria, and particularly in the family *Enterobacteriaceae* is driven by vertical and not horizontal evolution [32]. Together our data create a hypothesis for *Y. pseudotuberculosis* evolution whereby large population perturbations give rise to geographically isolated clones. During the early formation of these clones, exposure to geographically localized exogenous DNA creates a CRISPR array of immunity, which correlates with the repertoire of genetic material that can be transferred and acquired from the gene pool. As each of these clones then globally disseminates, they find themselves in co-existence with other clones of *Y. pseudotuberculosis*, but gene transfer between clones is restricted. This restriction is such that transfer of genetic material cannot occur at levels required to erode the clonal phylogenetic signature in the population, and consequently distinct phylogroups of *Y. pseudotuberculosis* persist in the population.

**Funding information**
T. S. is funded by a Nottingham Trent University, Vice Chancellors studentship awarded to T. S. and A. M. J. C. is partially funded by the COIN Centre of Excellence (Academy of Finland). This work was part of the European Defence Agency (EDA) project B-1325-ESM4-GP.

**Acknowledgements**
Katarzyna Leskinen, Joanna Zur and Monika Rajtor, from the University of Helsinki are thanked for the help in genomic DNA isolation.

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

**Ethical statement**
There are no ethical considerations applicable to the work presented.

**Data bibliography**
Accession numbers for all raw sequence data are in Table S1.

**Fig. 5.** BratNextGen analysis of core genome recombination events for 134 isolates of *Y. pseudotuberculosis*. Horizontal coloured bars show the indicated recombination events for each strain (y-axis); at the relative base pair positions (x-axis). CRISPR clusters and continent of origin are shown as coloured bars. Where segments are the same colour this indicates recombination events that are shared between those strains.
References

1. McNally A, Thomson NR, Reuter S, Wren BW. 'Add, stir and reduce': Yersinia spp. as model bacteria for pathogen evolution. Nat Rev Microbiol 2016;14:177–190.

2. Han TH, Paik IK, Kim SJ. Molecular relatedness between isolates Yersinia pseudotuberculosis from a patient and an isolate from mountain spring water. J Korean Med Sci 2003;18:425–428.

3. Behra GD, Garg DN, Batra HV, Chandiramani NK. Isolation of Yersinia pseudotuberculosis from bovine calves with enteric disorders. Microbiol Immunol 1984;28:237–241.

4. Savin C, Martin L, Bouchier C, Filali S, Chenau J et al. The Yersinia pseudotuberculosis complex: characterization and delineation of a new species, Yersinia wautersii. Int J Med Microbiol 2014;304:452–463.

5. Chiles MC et al. Pathogenic Yersinia pseudotuberculosis and Yersinia enterocolitica DNA is detected in bowel and mesenteric nodes from Crohn's disease patients. Mod. Pathol 2002;15:518.

6. Naktin J, Beavis KG. Yersinia enterocolitica and Yersinia pseudotuberculosis. Clin Lab Med 1999;19:523–536.

7. Nuorti JP, Niskanen T, Hallanvuo S, Mikkola J, Kela E et al. A widespread outbreak of Yersinia pseudotuberculosis O:3 infection from iceberg lettuce. J Infect Dis 2004;189:766–774.

8. Williamson DA, Baines SL, Carter GP, da Silva AG, Ren X et al. Genomic insights into a sustained national outbreak of Yersinia pseudotuberculosis. Genome Biol Evol 2016;8:ewv285–14.

9. Bogdanovich T, Carniel E, Fukushima H, Skurnik M. The complete genome sequence of Yersinia pseudotuberculosis from bovine calves with enteric disorders. Microbiol Immunol 2002;46:829–832.

10. Koskela KA, Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR et al. Detection of recombination events in bacterial genomes from large population samples. Nucleic Acids Res 2012;40:6e6.

11. Letunic I, Bork P. Interactive Tree of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res 2011;39:W475–W478.

12. Sahli JW, Caporaso JG, Rasko DA, Keim P. The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. PeerJ 2014;2:e332.

13. Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA et al. Phylogenetic distribution of CRISPR-Cas systems in antibi-resistant Pseudomonas aeruginosa. MBio 2015;6:e01796-15.

14. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014;30:2068–2069.

15. 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.

16. Yersinia enterocolitica subsp. paelearctica serogroup O:3 infection strain for tracing human infections. J Bacteriol 2011;193:2067–2068.

17. Cui Y, Yu C, Yan Y, Li D, Li Y et al. Historical variations in mutation rate in an epidemic pathogen, Yersinia pestis. Proc Natl Acad Sci USA 2013;110:577–582.

18. Bockaert R, Heled J, Kühnert D, Vaughan T, Wu CH et al. BEAST 2: a software platform for bayesian evolutionary analysis. PLoS Comput Biol 2014;10:e1003537.

19. Batzilla J, Höper D, Antonenka U, Heesemann J, Rakin A. Complete genome sequence of Yersinia enterocolitica subsp. paelearctica serogroup O:3. J Bacteriol 2011;193:2067–2068.

20. Eppinger M, Rosovitz MJ, Fricke WF, Rasko DA, Kokorina G et al. The complete genome sequence of Yersinia pseudotuberculosis IP31758, the causative agent of Far East scarlet-like fever. PLoS Genet 2007;3:e142.

21. Sheppard SK, Mccarthy ND, Falush D, Maiden MC. Convergence of Campylobacter species: implications for bacterial evolution. Science 2008;320:237–239.

22. van Belkum A, Soriaga LB, Lafave MC, Akella S, Veyrieras JB et al. Polymorphism of CRISPR-Cas systems in antibiotics-resistant Campylobacter jejuni strains erodes the signal for tracing human infections. Isme J 2016;10:721–729.

23. Achtman M, Zurth K, Morelli G, Torrea G, Guioulea A et al. Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis. Proc Natl Acad Sci USA 1999;96:14043–14048.

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.