HHS Public Access

Author manuscript

Diagn Microbiol Infect Dis. Author manuscript; available in PMC 2017 April 01.

Published in final edited form as:

Diagn Microbiol Infect Dis. 2016 April ; 84(4): 275–280. doi:10.1016/j.diagmicrobio.2015.12.003.

Whole genome multilocus sequence typing as an epidemiologic tool for *Yersinia pestis*

Luke C. Kingry[a], **Lori A. Rowe**[b], **Laurel B. Respicio-Kingry**[a], **Charles B. Beard**[a], **Martin E. Schriefer**[a], and **Jeannine M. Petersen**[a,*]

[a]Division of Vector-Borne Diseases, Bacterial Diseases Branch, Centers for Disease Control and Prevention, Fort Collins, CO 80523

[b]Division of Scientific Resources, Biotechnology Core Facility Branch, Centers for Disease Prevention and Control, Atlanta, GA 30329

Abstract

Human plague is a severe and often fatal zoonotic disease caused by *Yersinia pestis*. For public health investigations of human cases, nonintensive whole genome molecular typing tools, capable of defining epidemiologic relationships, are advantageous. Whole genome multilocus sequence typing (wgMLST) is a recently developed methodology that simplifies genomic analyses by transforming millions of base pairs of sequence into character data for each gene. We sequenced 13 US *Y. pestis* isolates with known epidemiologic relationships. Sequences were assembled de novo, and multilocus sequence typing alleles were assigned by comparison against 3979 open reading frames from the reference strain CO92. Allele-based cluster analysis accurately grouped the 13 isolates, as well as 9 publicly available *Y. pestis* isolates, by their epidemiologic relationships. Our findings indicate wgMLST is a simplified, sensitive, and scalable tool for epidemiologic analysis of *Y. pestis* strains.

Keywords

wgMLST; Whole genome sequencing; *Yersinia pestis*; Molecular epidemiology; Plague

1. Introduction

Yersinia pestis is the etiological agent of plague, a severe and often fatal illness most commonly transmitted to humans by the bite of an infected flea (Kugeler et al., 2015; Stenseth et al., 2008). The bacterium can also be transmitted to humans through the handling of tissues or fluids from a plague-infected animal or less commonly, by inhaling respiratory droplets from infected humans, cats, or dogs (Kugeler et al., 2015; Wang et al., 2011). The pneumonic form of plague is the most severe, and mortality rates in untreated patients approach 100% (Kugeler et al., 2015). Because of this extreme virulence, history of weaponization, and the possibility of person-to-person transmission, *Y. pestis* has been

*Corresponding author. Tel.: +1-970-266-3524. nzp0@cdc.gov (J.M. Petersen).

classified as a tier 1 priority pathogen, of highest concern with respect to an intentional event (Inglesby et al., 2000; Register, 2012).

Y. pestis is characterized as having a monomorphic genome (Achtman, 2008). Diversity observed among *Y. pestis* strains is limited and due primarily to single nucleotide polymorphisms (SNPs), variable number of tandem repeats (VNTRs), small insertions and deletions (INDELs), and insertion sequence (IS)–mediated rearrangements (Achtman et al., 2004; Auerbach et al., 2007; Colman et al., 2009; Drancourt et al., 2004; Gibbons et al., 2012; Huang et al., 2002; Klevytska et al., 2001; Lowell et al., 2005; Morelli et al., 2010; Motin et al., 2002; Touchman et al., 2007). Four classically defined biovars of *Y. pestis* exist worldwide; however, North American strains comprise only a single biovar, *orientalis*, due to a limited introduction into the continent in the early 20th century (Link, 1955; Morelli et al., 2010; Zhou et al., 2004). SNP typing of hundreds of US *Y. pestis* strains confirmed that all belong to only a single lineage (1.ORI) (Achtman et al., 2004). Among 1.ORI strains from the United States, Canada, and Madagascar, fewer than 40 SNPs have been identified by whole genome sequencing (WGS), and the mutation rate for *Y. pestis* has been estimated at 10^{-9} to 10^{-8} per site per year (Antonation et al., 2014; Auerbach et al., 2007; Gibbons et al., 2012; Morelli et al., 2010; Parkhill et al., 2001; Touchman et al., 2007; Vogler et al., 2013).

Molecular epidemiologic tools for *Y. pestis* are valuable for public health preparedness to track the geographic origin of isolates, to define exposure sources for human cases, to investigate outbreaks, and to distinguish between naturally occurring and intentional events. Currently, a hierarchal approach is used for molecular typing of *Y. pestis* strains. These assays are PCR based and rely on canonical SNPs (discovered from sequencing a limited number of full genomes) to first determine phylogenetic placement followed by a higher discriminatory secondary typing method, such as 43 locus VNTR analysis, for strain differentiation (Riehm et al., 2015; Riehm et al., 2012; Vogler et al., 2013).

With rapidly decreasing costs for WGS, a single approach that captures multiple types of sequence features could be advantageous for molecular epidemiologic investigations. Whole genome multilocus sequence typing (wgMLST) is an appealing approach as it captures various types of nucleotide differences (SNPs, VNTRs, and INDELs) for every open reading frame (ORF) of an organism, thereby allowing genome-wide comparisons by expanding upon the traditional 7–10 gene multilocus sequence typing (MLST) (Jolley and Maiden, 2014; Maiden, 2006; Perez-Losada et al., 2013). A benefit of wgMLST is the designation of alleles for each ORF in the genome, which transforms millions of base pairs of nucleotide sequence into character data for each gene. This in turn reduces the computing power necessary for whole genome comparisons. The usefulness and sensitivity of wgMLST for strain tracking and outbreak investigations has recently been demonstrated for several bacterial pathogens (Cody et al., 2013; Jolley et al., 2012; Jolley and Maiden, 2013; Sheppard et al., 2012).

The genome of the North American *Y. pestis* reference strain, CO92, encodes 3979 protein-coding genes which are located on the 4.6 Mb main chromosome and the 3 extrachromosomal plasmids, pMT1, pCD1, and pPCP1 (96 kb, 70 kb, and 9 kb,

respectively) (Parkhill et al., 2001). To test the usefulness of wgMLST for identification of epidemiologic relationships among *Y. pestis* strains, sequence diversity across all 3979 ORFs (4,046,060 bp) was assessed for 13 North American *Y. pestis* isolates with connections to one another representative of those encountered in public health investigations.

2. Materials and methods

2.1. Bacterial isolates and growth

Thirteen banked *Y. pestis* isolates, comprising 4 unrelated groups of isolates, with known epidemiologic relationships (Table 1), were chosen for analysis (Centers for Disease Control and Prevention, Division of Vector-Borne Diseases, Fort Collins, CO). Links between isolates were determined by epidemiological investigations of human cases and from field collections in which isolates were obtained from associated environmental samples (animals, fleas, and soil) (Table 1). All isolates were originally recovered from diagnostic specimens by either direct culture on sheep blood agar (SBA) or by passage of the original specimen through laboratory mice, followed by culture of infected tissues on SBA (Table 1). Isolates were streaked from frozen glycerol stocks to SBA and incubated at 35 °C for 48 hours, followed by subculture and incubation for an additional 24 hours at 35 °C. Multiple colonies were picked, pooled, and DNA extracted using the QIAamp DNA Mini kit and associated tissue protocol (Qiagen, Valencia, CA, USA). DNA concentrations were measured using the Nanodrop 2000 (Fisher Scientific, Pittsburgh, PA, USA).

2.2. WGS, contig assembly, and quality assessment

WGS was performed using the Roche 454 platform (Roche, Indianapolis, IN, USA) by the Biotechnology Core Facilities Branch, Genomic Sequencing Laboratory (Atlanta, GA, USA). Sequencing libraries were prepared per manufacturers protocols. Contigs were assembled de novo using either Newbler 2.5.3 (Roche) or CLC Genomics Workbench 7.0.3 (Qiagen).

Raw sequence reads were imported into CLC Genomics Workbench 7.0.3 and sequencing QC, raw read mapping, and alignments performed. Sequencing QC reports were to ensure raw read files for each strain had equivalent read length distribution, GC content, overrepresentation of sequence reads, and individual 5-mer distribution. Read mapping parameters included a mismatch cost of 2, insertion cost of 3, deletion cost of 3, length fraction of 0.5, and similarity fraction of 0.8. Reads that mapped to more than 1 region of the reference were allowed to map randomly.

2.3. Genome-wide identification of MLST alleles

Contig files from the assembly of each isolate were converted into individual BLAST databases and queried using nucleotide sequences for 3979 ORFs representing the protein coding genes of the *Y. pestis* CO92 reference genome in fasta format available at http://www.ncbi.nlm.nih.gov/genome/153?project_id=57621. All database creation and BLAST queries were performed using the standalone BLAST 2.2.29+ for UNIX available at http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download

(Camacho et al., 2009). Each BLAST output file was parsed into a list containing the top hit for each CO92 ORF and the percent identity score. SNPs were defined as any single variable nucleotide at a given position, VNTRs were defined as any adjacent repetitive sequence greater than or equal to 2 nucleotides, and INDELs were defined as any nucleotide gain or loss greater than or equal to 1 nucleotide with no adjacent repetitive sequence. Mutations adjacent to or within homopolymeric tracks were ignored due to the likelihood of sequencing errors at these loci. For all 13 isolates, SNPs, VNTRs, and INDELs were verified by inspecting raw reads mapped to the CO92 reference to ensure atleast 8× coverage and 90% of the base calls confirming the mutation. One instance of low SNP coverage (3×) was observed for the NM02-4452m isolate. The SNP was retained in the dataset because all other strains in group 2 demonstrated 100% agreement at this position. Mutations in ORFs annotated as phage-associated or a transposase were ignored due to the potential for ambiguous mapping. ORFs with verified mutation/s were assigned unique allele designations (sequential whole integers) based on comparison to the CO92 reference. Alleles were assigned and analyzed based on all mutation types (SNPs, VNTRs, and INDELs) as well as separately by individual mutation type.

2.4. wgMLST using publicly available Y. pestis genomes

Contig sequences corresponding to genomes of 9 *Y. pestis* strains from New Mexico were downloaded from GenBank (accession: PRJNA72923) in fasta format (Gibbons et al., 2012). Contig sequences for each strain were converted into individual BLAST databases and then queried with the 3979 ORFs from CO92 as described above. ORFs with percent identity less than 100% were determined for each strain and converted to allele designations. Mutations identified in the contig sequences could not be verified at the raw read level, as these data were not publicly available.

2.5. Cluster analysis

Clustering of isolates was performed using BioNumerics 7.1 (Applied Maths, Austin, TX, USA). Character data (whole integers) were used as an input for each allele and isolates clustered by categorical coefficients and unpaired group weighted means of averages. Isolates were clustered using allele assignments encompassing all mutation types (SNPs, VNTRs, and INDELs) as well as by individual mutation types.

3. Results

3.1. WGS of US Y. pestis isolates

Thirteen human, animal, flea, and environmental *Y. pestis* isolates originating from 3 US states, New Mexico, Colorado, and Arizona, were subjected to WGS. This sample set included 4 groups consisting of 2–4 isolates with known epidemiologic relationships (same transmission chain, same outbreak/epizootic, or an environmental investigation associated with a human plague case); none of the 4 groups of isolates had known connections to one another (Table 1). Group 1 included 3 linked isolates from Arizona recovered in 2007 from a mountain lion (AZ07-7301), a human who necropsied the mountain lion, and blood contaminated soil collected at the site where the mountain lion was found dead (AZ07-7298 and AZ07-7462, respectively) (Eisen et al., 2008; Wong et al., 2009). The soil was collected

~3 weeks after the mountain lion was found dead at the site. Group 2 included 4 linked isolates from New Mexico in 2002: 1 isolate from a human plague case (NM02-4452h), the same isolate passaged in the laboratory through a mouse (NM02-4452m), and isolates from 2 fleas recovered from the infected individual's yard 4 days after symptom onset in the patient (NM02-4476-306 and NM02-4477-309) (Colman et al., 2009). Group 3 was composed of 4 linked isolates recovered from 4 squirrels (*Sciurus niger*) over the course of a 3-month plague epizootic that occurred in a localized area of Denver, Colorado (Denver county) in 2007 (April: CO07-2003, CO07-2014, CO07-2015; June CO07-3070). Group 4 included 2 linked isolates from Moffatt county, Colorado, in 2007: an isolate from a rabbit (CO07-6570) and another from a flea collected from the rabbit (CO07-6570-120). Denver and Moffat counties are separated by approximately 403 km and are located on the eastern and western slopes of the Rocky Mountains, respectively.

To determine genome coverage for all 13 strains, raw reads were individually mapped to the CO92 reference genome. Average coverage ranged from 16 to 59× (mean 31×) for the 13 genomes, and average read length was 387 bp. De novo assembly of contigs resulted in contig N50 values ranging from 18 to 32 kb with maximum contig lengths from 60 to 100 kb long. Contig N50 values were equivalent for de novo contig assembly using Newbler or CLC Genomics Workbench.

3.2. Identification of MLST alleles

Of the 3979 predicted protein-coding ORFs (4,046,060 bp) encoded in the *Y. pestis* CO92 reference strain, allelic diversity among the 13 *Y. pestis* genomes was confined to 39 (0.98%) of the ORFS. A total of 81 different alleles were assigned. The majority of sequence diversity within ORFs was present in the form of SNPs (n = 30), followed by VNTRs (n = 5) and INDELs (n = 5); 25 of which are newly described here (Table 2). Sequence ambiguity at mutation sites was not evidenced; base calls for all identified mutations showed >99% agreement for raw read data. Of the 30 SNPs identified, 7 caused synonymous mutations, and 23 resulted in nonsynonymous mutations. Two of the nonsynonymous SNPs resulted in premature stop codons in the YPO0442 and YPO2948 genes. Only 1 of the 30 SNPs was identified in an extrachromosomal plasmid; this nonsynonymous SNP was located within the plasminogen activating protease (*pla*) ORF encoded in the pPCP1 plasmid. Of the 5 VNTRs identified, none disrupted the reading frame, and repeat lengths varied from 6 to 18 bp. The 5 identified INDELs ranged in size from a single nucleotide deletion to an 18-bp insertion. Two of the INDELs were 9 bp (deletion) and 18 bp (insertion) in size and did not disrupt the coding frame of the respective corresponding genes, YPO1989 and YPO1236. The remaining 3 INDELs, 1 insertion and 2 deletions, involved a single nucleotide resulting in a frameshift in the corresponding gene sequence.

Of the 39 ORFs, YPO0776 displayed the highest amount of diversity with 4 different alleles observed among the 4 groups of linked isolates. YPO1397 had 3 allele calls among the 4 groups of linked isolates; all other ORFs had only 2 allele calls. Among the mutation types, VNTRs showed the most sequence diversity between linked groups and thus the most allele calls for a single ORF; no VNTR variability was observed between isolates within the same transmission chain.

3.3. wgMLST-based cluster analysis of Y. pestis isolates

Cluster analysis based on allele designations encompassing all mutation types (SNPs, VNTRs, and INDELs) for the 3979 ORFs accurately grouped the 13 *Y. pestis* isolates into 4 main groups (Fig. 1A; Table 1). Groups 1–4 were separated by 18, 6, 2, and 9 unique alleles, respectively. Four alleles were shared among all sequenced isolates as compared to CO92 and included an 18-bp insertion in YPO1236, 2 nonsynonymous SNPs in YPO2029 and YPO4060, and a synonymous SNP in YPO3352. These 4 allele differences were also present in all 9 of the New Mexico *Y. pestis* sequences retrieved from NCBI and may represent strain specific alleles unique to CO92 or sequencing errors in the CO92 genome. No sequence differences in ORFs were observed among linked isolates within groups 1, 3, and 4 (isolates in the same transmission chain and isolates from the same epizootic). In contrast, among group 2 isolates, differences in 2 alleles separated the *Y. pestis* isolate recovered from the human (NM02-4452h) and the same isolate passaged through a mouse (NM02-4452m) from the 2 flea isolates recovered from the individual's yard 4 days after onset of the patient's illness (NM02-4476-306 and NM02-4477-309) (Fig. 1A and B). The differences were due to 2 SNPs, 1 synonymous SNP in YPO1638 and a nonsynonymous SNP (C to T transition) in the *pla* gene encoded on the pPCP1 plasmid.

To compare the relative sensitivity of allele designations incorporating all mutation types with those based on a single mutation type, data for only SNPs, VNTRs, or INDELs were extracted in silico, and cluster analyses were conducted (Fig. 1A–D). Clustering of the isolates based on SNPs (Fig. 1B) was essentially identical to that observed using SNPs, VNTRs, and INDELs, but with lower resolution between the 4 groups (Fig. 1A). Utilizing only INDELs or VNTRs (Fig. 1C and D), cluster analysis accurately separated all the groups; however, the isolates within group 2 were not differentiated as compared to cluster analysis using SNPs.

3.4. Application of Y. pestis wgMLST to previously sequenced US isolates

To demonstrate the scalability of the method, wgMLST was performed using the 3979 predicted protein-coding ORFs and genome sequence data publicly available from NCBI for 9 *Y. pestis* isolates from New Mexico (Fig. 2) (Gibbons et al., 2012). The isolates were recovered from June through August 2009 in and around Santa Fe (AGJS01.1, AGJU01.1, AGJV01.1, AGJY01.1, AGJZ01.1, AGKA01.1), Las Vegas (AGJT01.1), and Edgewood (AGJW01.1, AGJX01.1), New Mexico, and sources included 4 humans, 1 feline, 1 rabbit, and 3 prairie dogs (Gibbons et al., 2012). For 1 set (AGJW01.1, AGJX01.1), there was a known epidemiologic relationship between the source isolates (2 patients with the same exposure source), while the remaining 7 source isolates had only geographic and temporal information available.

Comparison of all 3979 ORFs across the 9 *Y. pestis* genomes identified 29 ORFs with allelic diversity. Sixteen of the allelic ORFs were unique to the 9 genomes, while 13 allelic ORFs were shared with the 13 US strains genome sequenced in this study. Sequence diversity within the 16 unique ORFs was present solely in the form of SNPs; all of which were described previously (Gibbons et al., 2012). When cluster analysis was performed based on allele designations for the 3979 ORFs, the 2 epidemiologically linked isolates (AGJW01.1,

AGJX01.1) were accurately clustered, displaying no sequence difference across 4,046,060 bp of analyzed genome sequence. Additionally, 3 isolates from Santa Fe (AGJY01.1, AGJZ01.1, AGKA01.1) recovered from animals over a 1.5-month period grouped together and differed from one another by only a single VNTR.

4. Discussion

wgMLST represents an appealing epidemiological tool for public health investigations as the vast majority of the genome can be interrogated in a straightforward manner. By analyzing 3979 predicted protein-encoding ORFs in *Y. pestis* and assigning alleles based on mutational differences, wgMLST was used to compare 83.7% (4,046,060 bp) of the entire *Y. pestis* genome sequence. As expected, sequence diversity was limited with allelic diversity among the 13 *Y. pestis* genomes sequenced in this study confined to <1% of the analyzed ORFs. Three different types of mutations, SNPs, INDELs, and VNTRs, were identified across the 13 US *Y. pestis* genomes. Because wgMLST alleles were assigned irrespective of mutation type, all genetic differences were readily combined and simultaneously analyzed by cluster analysis. For the 13 *Y. pestis* isolates sequenced in this study, the highest resolution between isolates with no relationship to one another was observed when alleles were assigned using all mutation types.

In this study, we chose to interrogate *Y. pestis* isolates where epidemiological linkages were known in order to gain information from expected results and to inform interpretation of these data. Isolates were analyzed from human, animal, flea, and soil sources 1) with a known chain of transmission, 2) from a localized epizootic, and 3) from a routine environmental investigation of a human plague case with an uncertain exposure source. Notably, exact sequence matches across 83.7% (4,046,060 bp) of the *Y. pestis* genome were demonstrated for 1) isolates recovered from blood contaminated soil and the animal that contaminated it as well as isolates from the infected animal and the individual who necropsied it (group 1); 2) an isolate recovered from a human case and the same isolate recovered post amplification in laboratory mice (group 2); 3) isolates from 4 animal hosts recovered during a geographically localized epizootic (group 3); 4) isolates recovered from an animal host and a flea which fed on this host (group 4); and 5) isolates from 2 patients with exposure to the same point source (AGJW01.1, AGJX01.1). These findings suggest that *Y. pestis* wgMLST alleles are not rapidly changing and perfect allele matches may be the norm for *Y. pestis* isolates with direct links to one another.

Notably, isolates collected during an environmental investigation of a human plague case did not demonstrate an exact match by wgMLST to the patient isolate (group 2). The 2 *Y. pestis* isolates were derived from fleas collected in the patient's yard 4 days after onset of the patient's symptoms and included in this analysis as the patients' yard was previously implicated as the most likely source of infection based on a 43 loci VNTR analysis (Colman et al., 2009). By wgMLST, the patient isolate differed by 2 SNPs as compared to the 2 flea isolates. Neither of the 2 SNPs has been described previously. For both isolates, coverage for these 2 SNPs was 10× (1,864,049) and 198× (7431, pCP1) with 100% agreement. Additionally, an independent full genome sequence of the patient isolate (NM02-4452; BioProject PRJNA224116), recently deposited in NCBI, verified the presence of both SNPs

(Johnson et al., 2015). A difference of 2 SNPs between *Y. pestis* strains appears to be significant. Only 2 SNPs separated unrelated *Y. pestis* strains in this study recovered 2 years and 628 km apart in Santa Fe, New Mexico (AGJU01.1, AGJV01.1), and Denver, Colorado (group 3). Based on the perfect matches observed for other linked isolates and the differences among group 2 strains resulting from SNPs, opposed to more rapidly changing VNTRs, we hypothesize that multiple strains were circulating in the area at the time of the patient's exposure. Indeed, Santa Fe County, New Mexico, has been shown to produce multiple *Y. pestis* genotypes in a single year (Gibbons et al., 2012), a phenomenon that has also been noted in both China and Madagascar (Riehm et al., 2015; Zhang et al., 2009). wgMLST of additional *Y. pestis* isolates derived from fleas collected in the patient's yard will be important for determining if an exact match to the patient isolate can be identified.

The ORF showing the highest level of allelic diversity between the *Y. pestis* isolates was YPO0776, which encodes a siderophore biosynthesis protein. Four different alleles, due to both VNTR and SNP differences, were assigned for the 4 groups of nonlinked isolates. The VNTR in YPO0776 has been described previously (M33) (Klevytska et al., 2001; Vogler et al., 2007). Compared against CO92, the differences in YPO0776 included 1 repeat gain in group 1, a 2 repeat loss in group 4, and a 1 repeat loss and a nonsynonymous SNP (T to G, 845,495) in group 3. Iron acquisition is a requirement for growth of *Y. pestis*, and knockouts in genes required for iron acquisition and utilization cause growth defects and affect the ability of the bacterium to cause disease in mice (Bearden and Perry, 1999). It is unknown whether YPO0776 is required for growth of *Y. pestis* as there are no functional characterizations of this gene in the literature. Potentially, the repeat changes within the YPO0776 VNTR may allow the bacteria to grow under conditions with differing iron availability.

Use of wgMLST as a molecular epidemiologic tool for *Y. pestis* lends itself to de novo discovery of novel mutations and alleles with each new strain that is sequenced. Thirty-nine allelic ORFs containing 18 new SNPs and 4 new INDELS were discovered among the 13 *Y. pestis* genomes sequenced in this study and analysis of 9 publicly available *Y. pestis* genomes identified 16 additional ORFs with allelic diversity (Gibbons et al., 2012). While this study focused only on ORFs, intergenic regions can easily be incorporated into wgMLST analyses. In the case of the 13 *Y. pestis* genomes sequenced in this study, preliminary analyses identified 14 extragenic SNPs that differed between each of the 4 groups of isolates, whereas no extragenic SNP differences were identified between linked isolates in each of the 4 groups. As more *Y. pestis* isolates are analyzed by wgMLST, comparison of alleles based on SNPs (extragenic, intragenic, or both) or all mutation types will be important for determining which are the most useful for identification of epidemiologic relationships between isolates.

The findings presented here suggest that wgMLST shows promise as a single-platform, high-resolution method for investigation of plague cases in the United States. Establishing interpretative criteria for defining epidemiologic relationships and identifying newer next generation platforms and a standardized bioinformatic approaches are important next steps. Addition and analysis of wgMLST alleles from isolates outside the 1.ORI lineage will be important for ascertaining the utility of this approach on a global level. Finally, beyond

wgMLST alleles, the use of WGS for molecular typing has the added benefit of capturing and storing other sequence information, such as foreign DNA, which may be informative in unusual cases or outbreaks of disease.

References

Achtman M. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. Annu Rev Microbiol. 2008; 62:53–70. [PubMed: 18785837]

Achtman M, Morelli G, Zhu P, Wirth T, Diehl I, Kusecek B, et al. Microevolution and history of the plague bacillus, Yersinia pestis. Proc Natl Acad Sci U S A. 2004; 101:17837–42. [PubMed: 15598742]

Antonation KS, Shury TK, Bollinger TK, Olson A, Mabon P, Van Domselaar G, et al. Sylvatic plague in a Canadian black-tailed prairie dog (Cynomys ludovicianus). J Wildl Dis. 2014; 50:699–702. [PubMed: 24807359]

Auerbach RK, Tuanyok A, Probert WS, Kenefic L, Vogler AJ, Bruce DC, et al. Yersinia pestis evolution on a small timescale: comparison of whole genome sequences from North America. PLoS One. 2007; 2:e770. [PubMed: 17712418]

Bearden SW, Perry RD. The Yfe system of Yersinia pestis transports iron and manganese and is required for full virulence of plague. Mol Microbiol. 1999; 32:403–14. [PubMed: 10231495]

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10:421. [PubMed: 20003500]

Cody AJ, McCarthy ND, Jansen van Rensburg M, Isinkaye T, Bentley SD, Parkhill J, et al. Real-time genomic epidemiological evaluation of human Campylobacter isolates by use of whole-genome multilocus sequence typing. J Clin Microbiol. 2013; 51:2526–34. [PubMed: 23698529]

Colman RE, Vogler AJ, Lowell JL, Gage KL, Morway C, Reynolds PJ, et al. Fine-scale identification of the most likely source of a human plague infection. Emerg Infect Dis. 2009; 15:1623–5. [PubMed: 19861057]

Drancourt M, Roux V, Dang LV, Tran-Hung L, Castex D, Chenal-Francisque V, et al. Genotyping, Orientalis-like Yersinia pestis, and plague pandemics. Emerg Infect Dis. 2004; 10:1585–92. [PubMed: 15498160]

Eisen RJ, Petersen JM, Higgins CL, Wong D, Levy CE, Mead PS, et al. Persistence of Yersinia pestis in soil under natural conditions. Emerg Infect Dis. 2008; 14:941–3. [PubMed: 18507908]

Gibbons HS, Krepps MD, Ouellette G, Karavis M, Onischuk L, Leonard P, et al. Comparative genomics of 2009 seasonal plague (Yersinia pestis) in New Mexico. PLoS One. 2012; 7:e31604. [PubMed: 22359605]

Huang XZ, Chu MC, Engelthaler DM, Lindler LE. Genotyping of a homogeneous group of Yersinia pestis strains isolated in the United States. J Clin Microbiol. 2002; 40:1164–73. [PubMed: 11923326]

Inglesby TV, Dennis DT, Henderson DA, Bartlett JG, Ascher MS, Eitzen E, et al. Plague as a biological weapon: medical and public health management. Working Group on Civilian Biodefense. JAMA. 2000; 283:2281–90. [PubMed: 10807389]

Johnson SL, Daligault HE, Davenport KW, Jaissle J, Frey KG, Ladner JT, et al. Thirty-two complete genome assemblies of nine Yersinia species, including Y. pestis, Y. pseudotuberculosis, and Y. enterocolitica. Genome Announc. 2015; 3

Jolley KA, Maiden MC. Automated extraction of typing information for bacterial pathogens from whole genome sequence data: Neisseria meningitidis as an exemplar. Euro Surveill. 2013; 18:20379. [PubMed: 23369391]

Jolley KA, Maiden MC. Using MLST to study bacterial variation: prospects in the genomic era. Future Microbiol. 2014; 9:623–30. [PubMed: 24957089]

Jolley KA, Hill DM, Bratcher HB, Harrison OB, Feavers IM, Parkhill J, et al. Resolution of a meningococcal disease outbreak from whole-genome sequence data with rapid Web-based analysis methods. J Clin Microbiol. 2012; 50:3046–53. [PubMed: 22785191]

Klevytska AM, Price LB, Schupp JM, Worsham PL, Wong J, Keim P. Identification and characterization of variable-number tandem repeats in the Yersinia pestis genome. J Clin Microbiol. 2001; 39:3179–85. [PubMed: 11526147]

Kugeler KJ, Staples JE, Hinckley AF, Gage KL, Mead PS. Epidemiology of human plague in the United States, 1900–2012. Emerg Infect Dis. 2015; 21:16–22. [PubMed: 25529546]

Link VB. A history of plague in United States of America. Public Health Monogr. 1955; 26:1–120. [PubMed: 14371919]

Lowell JL, Wagner DM, Atshabar B, Antolin MF, Vogler AJ, Keim P, et al. Identifying sources of human exposure to plague. J Clin Microbiol. 2005; 43:650–6. [PubMed: 15695659]

Maiden MC. Multilocus sequence typingof bacteria. Annu Rev Microbiol. 2006; 60:561–88. [PubMed: 16774461]

Morelli G, Song Y, Mazzoni CJ, Eppinger M, Roumagnac P, Wagner DM, et al. Yersinia pestis genome sequencing identifies patterns of global phylogenetic diversity. Nat Genet. 2010; 42:1140–3. [PubMed: 21037571]

Motin VL, Georgescu AM, Elliott JM, Hu P, Worsham PL, Ott LL, et al. Genetic variability of Yersinia pestis isolates as predicted by PCR-based IS100 genotyping and analysis of structural genes encoding glycerol-3-phosphate dehydrogenase (glpD). J Bacteriol. 2002; 184:1019–27. [PubMed: 11807062]

Parkhill J, Wren BW, Thomson NR, Titball RW, Holden MT, Prentice MB, et al. Genome sequence of Yersinia pestis, the causative agent of plague. Nature. 2001; 413:523–7. [PubMed: 11586360]

Perez-Losada M, Cabezas P, Castro-Nallar E, Crandall KA. Pathogen typing in the genomics era: MLST and the future of molecular epidemiology. Infect Genet Evol. 2013; 16:38–53. [PubMed: 23357583]

Register F. Possession, use, and transfer of select agents and toxins. 2012; 77:61084–115.

Riehm JM, Vergnaud G, Kiefer D, Damdindorj T, Dashdavaa O, Khurelsukh T, et al. Yersinia pestis lineages in Mongolia. PLoS One. 2012; 7:e30624. [PubMed: 22363455]

Riehm JM, Projahn M, Vogler AJ, Rajerison M, Andersen G, Hall CM, et al. Diverse genotypes of Yersinia pestis caused plague in Madagascar in 2007. PLoS Negl Trop Dis. 2015; 9:e0003844. [PubMed: 26069964]

Sheppard SK, Jolley KA, Maiden MC. A gene-by-gene approach to bacterial population genomics: whole genome MLST of Campylobacter. Genes (Basel). 2012; 3:261–77. [PubMed: 24704917]

Stenseth NC, Atshabar BB, Begon M, Belmain SR, Bertherat E, Carniel E, et al. Plague: past, present, and future. PLoS Med. 2008; 5:e3. [PubMed: 18198939]

Touchman JW, Wagner DM, Hao J, Mastrian SD, Shah MK, Vogler AJ, et al. A North American Yersinia pestis draft genome sequence: SNPs and phylogenetic analysis. PLoS One. 2007; 2:e220. [PubMed: 17311096]

Vogler AJ, Keys CE, Allender C, Bailey I, Girard J, Pearson T, et al. Mutations, mutation rates, and evolution at the hypervariable VNTR loci of Yersinia pestis. Mutat Res. 2007; 616:145–58. [PubMed: 17161849]

Vogler AJ, Chan F, Nottingham R, Andersen G, Drees K, Beckstrom-Sternberg SM, et al. A decade of plague in Mahajanga, Madagascar: insights into the global maritime spread of pandemic plague. MBio. 2013; 4:e00623–12. [PubMed: 23404402]

Wang H, Cui Y, Wang Z, Wang X, Guo Z, Yan Y, et al. A dog-associated primary pneumonic plague in Qinghai Province, China. Clin Infect Dis. 2011; 52:185–90. [PubMed: 21288842]

Wong D, Wild MA, Walburger MA, Higgins CL, Callahan M, Czarnecki LA, et al. Primary pneumonic plague contracted from a mountain lion carcass. Clin Infect Dis. 2009; 49:e33–8. [PubMed: 19555287]

Zhang X, Hai R, Wei J, Cui Z, Zhang E, Song Z, et al. MLVA distribution characteristics of Yersinia pestis in China and the correlation analysis. BMC Microbiol. 2009; 9:205. [PubMed: 19775435]

Zhou D, Tong Z, Song Y, Han Y, Pei D, Pang X, et al. Genetics of metabolic variations between Yersinia pestis biovars and the proposal of a new biovar, microtus. J Bacteriol. 2004; 186:5147–52. [PubMed: 15262951]

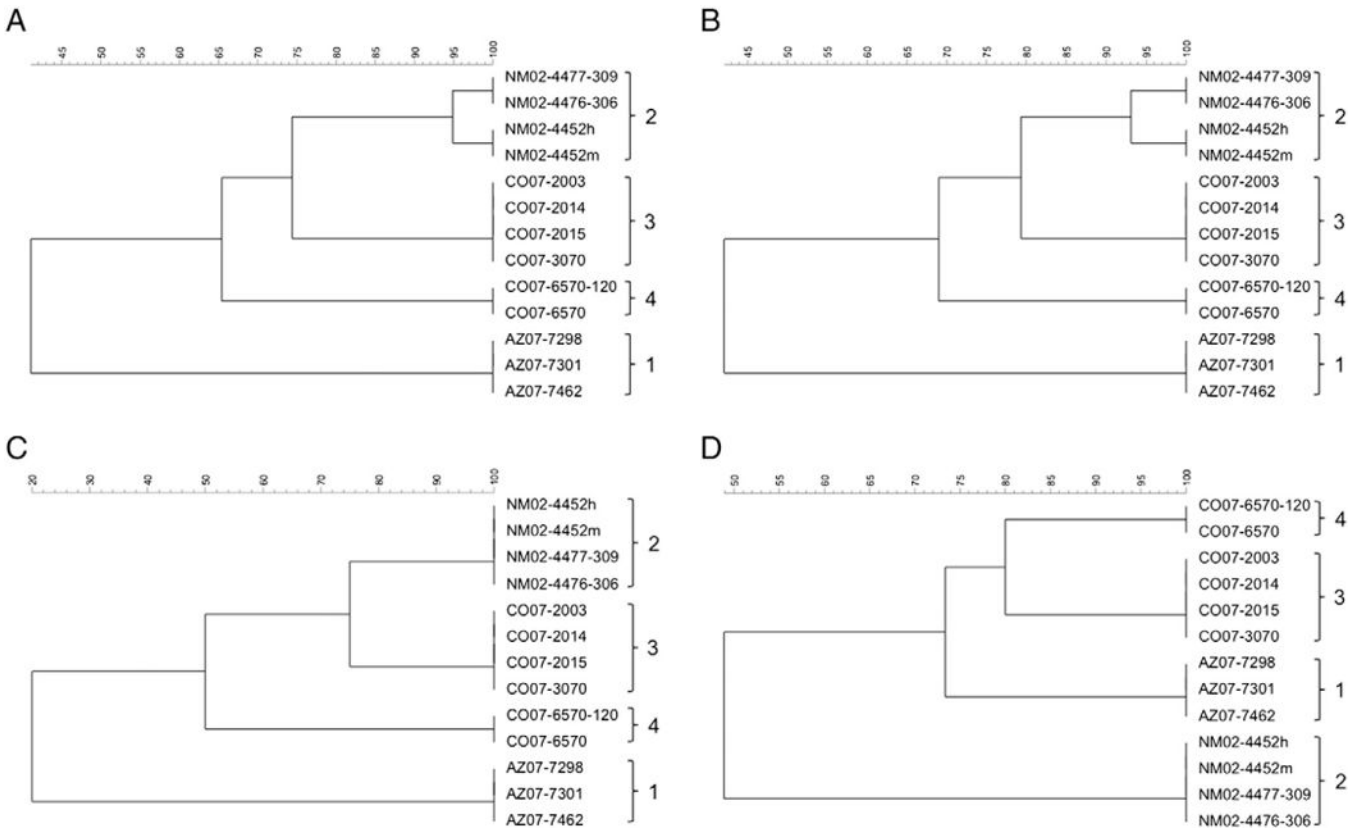

Fig. 1.

Cluster analysis of 4 groups (groups 1–4) of *Y. pestis* isolates with known epidemiological relationships using wgMLST allele data comprised of (A) SNPs, INDELs, and VNTRs (39 ORFs, 81 alleles). (B) Cluster analysis using SNPs alone (29 ORFs, 58 alleles), (C) VNTRs alone (5 ORFs, 13 alleles), and (D) INDELs alone (5 ORFs, 10 alleles). The 4 groups of linked isolates are indicated.

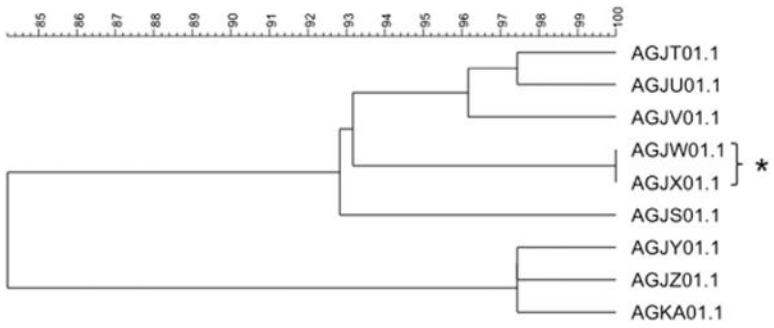

Fig. 2.

Cluster analysis using wgMLST allele data composed of SNPs, INDELs, and VNTRs (29 ORFs, 54 alleles) derived from publicly available sequence data for 9 *Y. pestis* isolates from New Mexico in 2009. The single set of isolates with a known epidemiological link is indicated by the asterisk.

Table 1

Epidemiologically linked sets of *Y. pestis* isolates genome sequenced in this study.

Group	Strain ID	Source	Isolation method	Date isolated	State	County	GenBank accession no.
Group 1	AZ07-7301	Mountain lion	Direct culture	November 2007	Arizona	Coconino	SAMN03322694
Group 1	AZ07-7462	Soil	Mouse passage	November 2007	Arizona	Coconino	SAMN03322695
Group 1	AZ07-7298	Human	Direct culture	November 2007	Arizona	Coconino	SAMN03322696
Group 2	NM02-4452h	Human	Direct culture	November 2002	New Mexico	Santa Fe	SAMN03322697
Group 2	NM02-4452m	Human	Mouse passage	November 2002	New Mexico	Santa Fe	SAMN03322698
Group 2	NM02-4476-306	Flea	Mouse passage	November 2002	New Mexico	Santa Fe	SAMN03322699
Group 2	NM02-4477-309	Flea	Mouse passage	November 2002	New Mexico	Santa Fe	SAMN03322700
Group 3	CO07-2014	Squirrel	Direct culture	April 2007	Colorado	Denver	SAMN03322701
Group 3	CO07-2003	Squirrel	Direct culture	April 2007	Colorado	Denver	SAMN03322702
Group 3	CO07-2015	Squirrel	Direct culture	April 2007	Colorado	Denver	SAMN03322703
Group 3	CO07-3070	Squirrel	Mouse Passage	June 2007	Colorado	Denver	SAMN03322704
Group 4	CO07-6570	Rabbit	Mouse Passage	September 2007	Colorado	Moffat	SAMN03322705
Group 4	CO07-6570-120	Flea	Mouse passage	September 2007	Colorado	Moffat	SAMN03322706

Table 2

List of mutations identified in 13 US *Y. pestis* isolates by wgMLST.

ID	Gene symbol	Annotation	CO92 genome position	Mutation compared to CO92	Substitution type	Effect on A.A. sequence	Group	Previously described
YPO0005	ravA	Regulatory ATPase	4032	15 bp loss	VNTR	QQFQI deleted	4	
YPO0015	malQ	4-alpha-glucanotransferase	138177	T/C	SNP, Syn		1	Touchman et al., 2007
YPO0139	hslO	Heat shock protein 33	153268	C/A	SNP, Non-syn	G/C	4	
YPO0193	slyD	FKBP-type peptidylprolyl isomerase	211446	6 bp loss	VNTR	HE deleted	1	
YPO0283	hmuR	Hemin receptor	284552	G/-	INDEL	Frameshift	2	Gibbons et al., 2012
YPO0442	serB	Phosphoserine phosphatase	463235	C/T	SNP, Non-syn	Q/Stop	4	
YPO0452	slt	Lytic murein transglycosylase	477895	C/A	SNP, Syn		1	
YPO0734	YPO0734	Hypothetical protein	792941	C/A	SNP, Syn		1	
YPO0776	YPO0776	Siderphore biosynthesis protein	Multiple	Gain/loss	VNTR		1/3/04	Klevytska et al., 2001
YPO0776	YPO0776	Siderophore biosynthesis protein	845495	T/G	SNP, Non-syn	Q/P	3	
YPO1002	YPO1002	Enterotoxin-like protein	1116888	G/A	SNP, Non-syn	P/L	4	
YPO1020	recB	Exonuclease V subunit beta	1156992	C/T	SNP, Syn		1	
YPO1055	fabZ	(3R)-hydroxymyristoyl-ACP dehydratase	1198069	15 bp deletion	VNTR	VVKPD deleted	1	Klevytska et al., 2001
YPO1236	YPO1236	Aldolase	1395944	18 bp insertion	INDEL	EPPPSL inserted	1/2/3/4	
YPO1356	YPO1356	Hypothetical protein	1519950	C/A	SNP, Non-syn	G/W	1	
YPO1383	pfl	Formate acetyltransferase 1	1558606	C/T	SNP, Non-syn	A/T	2	Gibbons et al., 2012
YPO1386	ansB	L-asparaginase II	1564799	C/T	SNP, Non-syn	A/V	2	Gibbons et al., 2012
YPO1397	YPO1397	Hypothetical protein	1580108	Gain/Loss	VNTR		1/3	
YPO1491	ybiT	ABC transporter ATB-binding protein	1692374	A/T	SNP, Non-syn	D/V	1	
YPO1638	mnmA	tRNA-specific 2-thiouridylase	1864049	A/G	SNP, Syn		2 (4452 only)	
YPO1813	YPO1813	Sugar binding protein	2060795	C/A	SNP, Non-syn	A/E	2	Gibbons et al., 2012
YPO1910	irp 1	Yersiniabactin biosynthetic protein	2151056	G/T	SNP, Non-syn	H/N	1	
YPO1989	YPO1989	Hypothetical protein	2260679	9 bp deletion	INDEL	DIN deleted	4	
YPO2005	YPO2005	Hypothetical protein	2278317	A/G	SNP, Non-syn	V/A	1/2	Auerbach et al., 2007; Touchman et al., 2007

ID	Gene symbol	Annotation	CO92 genome position	Mutation compared to CO92	Substitution type	Effect on A.A. sequence	Group	Previously described
YPO2029	YPO2029	Hypothetical protein	2300659	T/G	SNP, Non-syn	D/A	1/2/3/4	Auerbach et al., 2007; Touchman et al., 2007
YPO2045	YPO2045	Hemolysin	2320415	G/T	SNP, Non-syn	G/C	1	Touchman et al., 2007
YPO2527	menD	2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylate synthase	2836287	G/A	SNP, Non-syn	A/V	1	
YPO2679	celB	PTS system N,N′-diacetylchitobiose–specific transporter subunit IIC	3004948	T/C	SNP, Non-syn	V/A	4	
YPO2884	YPO2884	Hypothetical protein	3221741	C/T	SNP, Non-syn	R/H	1	
YPO2948	YPO2948	Hypothetical protein	3294996	G/T	SNP, Non-syn	E/Stop	4	
YPO2998	YPO2998	Two-component response-regulatory protein	3348820	A/-	INDEL	Frameshift	2	
YPO3049	YPO3049	Binding protein-dependent transporter	3406528	G/A	SNP, Non-syn	A/V	2	Gibbons et al., 2012
YPO3287	YPO3287	Two-component response-regulatory protein	3667446	G/A	SNP, Non-syn	E/K	1	Touchman et al., 2007
YPO3352	ydjL	Zinc-binding dehydrogenase	3739401	C/A	SNP, Syn		1/2/3/4	Auerbach et al., 2007; Touchman et al., 2007
YPO3598	YPO3598	Hypothetical protein	4006482	-/C	INDEL	Frameshift	1	
YPO3859	rffA	TDP-4-oxo-6-deoxy-D-glucose transaminase	4332558	C/A	SNP, Non-syn	G/C	4	
YPO3941	glgX	Glycogen debranching protein	4428477	C/T	SNP, Non-syn	G/E	1	Touchman et al., 2007
YPO4000	dppD	Dipeptide transporter ATP-binding protein	4509262	C/T	SNP, Syn		1	
YPO4060	fdhD	Formate dehydrogenase accessory protein	4579183	A/G	SNP, Non-syn	S/G	1/2/3/4	Auerbach et al., 2007; Touchman et al., 2007
YPPCP1.07	pla	Outer membrane protein	7431	C/T	SNP, Non-syn	T/I	2 (4452 only)	