Reannotation of *Yersinia pestis* Strain 91001 Based on Omics Data

Yiqing Mao,^{1,2} Xianwei Yang,¹ Yang Liu,³ Yanfeng Yan,¹ Zongmin Du,¹ Yanping Han,¹ Yajun Song,¹ Lei Zhou,¹
Yujun Cui,^{1*} and Ruifu Yang^{1*}

¹State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, Beijing, People's Republic of China;

²Center of Information Technology, Beijing Institute of Health and Medical Information, Beijing, People's Republic of China;

³Department of Biotechnology, Beijing Institute of Radiation Medicine, Beijing, People's Republic of China

Abstract. *Yersinia pestis* is among the most dangerous human pathogens, and systematic research of this pathogen is important in bacterial pathogenomics research. To fully interpret the biological functions, physiological characteristics, and pathogenesis of *Y. pestis*, a comprehensive annotation of its entire genome is necessary. The emergence of omics-based research has brought new opportunities to better annotate the genome of this pathogen. Here, the complete genome of *Y. pestis* strain 91001 was reannotated using genomics and proteogenomics data. One hundred and thirty-seven unreliable coding sequences were removed, and 41 homologous genes were relocated with their translational initiation sites, while the functions of seven pseudogenes and 392 hypothetical genes were revised. Moreover, annotations of non-coding RNAs, repeat sequences, and transposable elements have also been incorporated. The reannotated results are freely available at <http://today.bmi.ac.cn>.

INTRODUCTION

Yersinia pestis, a gram-negative bacterium, is the causative agent of bubonic and pneumonic plague, which are systemic, invasive diseases. The pathogenic lifestyle of this microbe involves two distinct life stages: one in the flea vector and the other in mammalian reservoirs, primarily rodents.¹ This notorious pathogen has caused hundreds of millions of deaths in three major plague pandemics in human history. According to the World Health Organization, there have been 18 plague outbreaks since 2001, the latest of which occurred in Madagascar in September 2015.²

The first complete genome of *Y. pestis* strain CO92 was sequenced by the Wellcome Trust Sanger Institute (Cambridgeshire, United Kingdom) in 2001, and it consists of a 4.65-Mb chromosome and three plasmids of 96.2, 70.3, and 9.6 kb.³ The genome of strain 91001, an isolate that is avirulent to humans, was sequenced in our laboratory,⁴ and it consists of a circular chromosome and four plasmids, pCD1, pMT1, pPCP1, and pCRY. The size of its chromosome is slightly smaller than that of CO92 (4.60 Mb), and it contains 4,136 genes.

With the rapid development of experimental methods and techniques, especially advances in next-generation sequencing technology, there have been many extensive studies of *Y. pestis*, and numerous data have been accumulated in databases, such as GenBank of the National Center for Biotechnology Information (NCBI). Using transcriptomics and proteogenomics⁵ can alleviate many of the problematic areas in genome annotations. When we retrospectively analyzed genome annotation results, contradictions and even errors were inevitably found because of limited knowledge, and these errors could be amplified and incorporated in subsequent annotations. In addition, emerging proteomics data are dispersed in diverse databases and literatures, and they are not systematically classified and integrated. Our reanno-

tation study aims to generalize, summarize, and improve our existing knowledge of *Y. pestis*.

Genome annotation can elaborate cell functions, biological behaviors, and the pathogenesis of bacteria in a systematic way. However, this process has many limitations and, thus, reannotation is essential.⁶ Reannotation is a process of annotating a previously annotated genome using improved bioinformatics methods and more comprehensive databases.^{7–9} In addition to using better annotation tools, another effective method is to use data from multi-omics measurements, such as transcriptomics and proteogenomics, to increase the amount of information used in the annotations. Many online databases^{10–12} and platforms^{13–15} can facilitate this process for eukaryotes and prokaryotes. Although automatic annotation pipelines can save time and resources, they do not incorporate information from expert curators. Some studies have reannotated *Y. pestis* genomes using omics data, for example, using comparative methods. Schrimpe-Rutledge and others reannotated *Y. pestis* strains CO92 and PestoidesF using transcriptomics and proteomics data,¹⁶ and Payne and others revised the annotation of the *Y. pestis* KIM strain using proteomics data.¹⁷

By integrating published information, including information from various databases, proteogenomics data from tandem mass tag (TMT) mass spectrometry (MS), and small RNA (sRNA) information from RNA sequencing (RNA-seq),¹⁸ this study reannotated the complete genome of *Y. pestis* strain 91001, which was originally annotated and released in 2004. The coding sequences (CDSs), translational initiation sites (TISs), pseudogenes, function annotations, noncoding RNAs (ncRNAs), repeat sequences, and transposable elements have been updated by in-depth analyses. We also built a reannotation pipeline that is also suitable for other *Y. pestis* genomes. The pipeline analyzes and integrates *Y. pestis* omics data by deploying a series of gene prediction software, protein annotation tools, and information from public databases and published studies.¹⁹ All of the results are freely accessible at <http://today.bmi.ac.cn>.

RESULTS

CDS adjustment. To identify CDSs, we used BLASTX to align the genome of strain 91001 with the nonredundant

*Address correspondence to Yujun Cui or Ruifu Yang, State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, Beijing 100071, No. 20, Dongda Street, Fengtai District, Beijing, China. E-mails: ruifuyang@gmail.com or cuiyujun.new@gmail.com

TABLE 1
Number of coding sequences in 91001

Position	Previous annotation	Reannotation
Chromosome	3,889	3,774
pCD	85	77
pMT	122	110
pPCP	10	10
pCRY	30	28
Total number	4,136	3,999

(NR) reference database at the NCBI to generate an original dataset containing 4,136 genes. Then, we separately predicted gene positions using the GLIMMER, GeneMarkS, and Prodigal programs. A gene was only included in the final CDS set when its 5' and 3' positions were predicted consistently by at least two of the programs (Table 1).

In this process, six protein annotation errors were corrected using the NR database, which has been updated since the original annotation in 2004 (Supplemental Table 1), and 131 CDSs in the original gene set were removed from the annotation, as they lacked consistent 5' or 3' positions (Supplemental Table 1).

Translation initiation sites. To verify the positions of TISs, we aligned the MS peptide sequences of strain 91001 with its genome sequences (TBLASTN threshold: $e\text{-value} < 10^{-3}$, identity $> 80\%$). If the predicted TIS of a gene was located in the middle of the MS peptide, we extended it to the last upstream initiation codon. If the extended sequence also matched this peptide, the gene was corrected by extending the TIS to the new position (Figure 1). The MS peptides were usually too short to acquire a reliable alignment of the gene sequences, which led to very few TISs being identified in the middle of the MS peptides; therefore, only two genes were corrected by this approach.

In addition to using MS data, we also integrated the results from commonly used gene prediction tools to further reannotate the positions of the TISs. We used the GLIMMER, GeneMarkS, and Prodigal programs to build a reliable gene set that included 2,302 fully consistent gene position prediction results (see section "Results"). The TISs of 39 genes were shown to differ between the previous annotation and our reliable gene set; thus, they were corrected based on their position in the reliable gene set.

Pseudogene reannotation. In the previous annotation, 143 genome fragments in strain 91001 were classified as pseudo-

genes. The majority of the pseudogenes resulted from the insertion or deletion of nucleotides within coding regions, which led to frame shifting.²⁰ Schrimpe-Rutledge and others reannotated the genome of *Y. pestis* strain CO92 using MS data, and 40 annotated pseudogenes were revised,¹⁶ and Payne and others reannotated the genome of the *Y. pestis* strain KIM using MS data, which led to the revision of only one pseudogene.¹⁷

Because of the high homology among known *Y. pestis* genomes, we combined the MS data and the reliable gene set of strain 91001 to exclude mistakenly annotated pseudogenes. First, the MS data from strain 91001 were aligned to all of the pseudogenes using TBLASTN (threshold: $e\text{-value} < 10^{-4}$, identity $> 80\%$), and the filtered results were termed dubious pseudogenes. Then, the reliable gene set was incorporated to identify the gene locations. If a dubious pseudogene contained a reliable gene, we considered this pseudogene to be a highly probable gene. Finally, seven pseudogenes were revised accordingly.

For instance, YP_1507 was a pseudogene in the previous annotation because it is similar to *Y. pestis* transposase Y1062, and it contains a frameshift mutation after codon 286. This pseudogene was split into two genes in the reliable prediction gene set. Four fragments of MS peptides could be separately matched to the two predicted genes (red box shown in Figure 2), which suggests that the genes could be successfully transcribed. Therefore, YP_1507 is no longer a pseudogene in our annotation, and it is reannotated as two putative genes.

Function reannotation. When a complete genome is submitted to a public database, functional annotations of this genome are generally enclosed. Most of the functional annotations are based on similarity information from homologous genes with experimentally verified functions, and they integrate information from various databases, such as the NR database at the NCBI.

Here, we aligned the sequences of the predicted CDSs of strain 91001 with the NR and European Bioinformatics Institute databases to acquire information concerning the corresponding protein products. To acquire further information, we also aligned the CDSs with 12 more databases, which provided annotation information regarding protein families and domains (Pfam, TIGRFAM, ProDom, SMART, and Prosite profiles), microbial proteomes (HAMAP), protein structures (SUPERFAMILY and Gene3D), motifs (PRINTS), and classification and prediction (PANTHER, PIRSF, and

FIGURE 1. Revision of translation initiation sites (TISs) using mass spectrometry (MS) data. YP_2457, YPO3875, and y2675 are homologous genes, but YPO3875 and y2675 are 84-bp longer than YP_2457 (as shown in the green box). A 26-amino acid MS peptide can be aligned to YP_2457 (position -45 to -33 bp, as shown in the red box). Thus, we extended YP_2457 to the latest TIS. After performing a correction using the MS data, the TIS of YP_2457 was consistent with the homologous genes of the other published *Yersinia pestis* genomes.

FIGURE 2. Reannotation of YP_1507. YP_1507 includes two coding sequences in the reliable prediction gene set. Four fragments of mass spectrometry peptides, marked with red boxes, matched the two predicted genes.

Coils) (Supplemental Table 2). We also classified the proteins according to their sequences or structural features by aligning them using the Clusters of Orthologous Groups, Pfam, and Gene Ontology databases. Finally, we identified the pathways to which the genes belonged using the Kyoto Encyclopedia of Genes and Genomes database.²¹

The predicted functions of 392 hypothetical or putative proteins in the previous annotation were revised, and 2,900 CDSs (73%) (thresholds: $e\text{-value} < 10^{-3}$ and aligned length $> 50\%$ of the total length) had at least one hit in the aforementioned databases, which will provide clues for further physiology and phenotype studies of *Y. pestis*.

ncRNA reannotation. ncRNAs, such as transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), and sRNAs, refer to RNAs that are not translated into proteins, but which play

important cellular roles.²² We identified and characterized the ncRNAs from strain 91001 using Rfam, a database of the RNA sequence families of structural RNAs, which includes ncRNA genes and *cis*-regulatory elements.²³ Seventy-three tRNAs, 35 *cis*-regulatory elements, 29 rRNAs, and 92 predicted RNAs were found in the database. We also incorporated information for the other 134 sRNAs of strain 91001 from an RNA-seq study.¹⁸ Together, we updated the information regarding two unknown RNAs, and we added 272 newly annotated ncRNAs using the literature and by comparing the results with the Rfam database. As a result, compared with the 102 previously annotated ncRNAs of strain 91001, 374 ncRNA elements are present in the current annotation.

Repeat sequence reannotation. Repeat sequences consist of three categories: local repeats (tandem repeats (TRs), interspersed repeat families (mainly transposable elements and retrotransposons), and large segmental duplications (fragments of genomic amplifications), which all could be identified using pattern recognition tools. From the 91001 genome, we identified 289 TRs, five long terminal repeat (LTR) retrotransposons, three clustered regularly interspaced short palindromic repeats (CRISPRs), and 147 large segmental duplications. Insertion sequences (ISs), TRs, and large segmental duplications are the main types of repeat sequences in strain 91001. ISs comprised *IS100* ($N = 30$),

FIGURE 3. Length distribution of tandem repeats (TRs), insertion sequences (ISs), large segmental duplications, and long terminal repeat (LTR) retrotransposons in strain 91001. Most TRs are shorter than 100 bp. ISs mostly comprised *IS285*, *IS100*, *IS200*, and *IS1661*. Three highly frequent, large segmental duplications carry genes encoding transposases for *IS1541*, *IS100*, and *IS285*. The lengths of the LTRs range from 5,000 to 18,000 bp.

TABLE 2
Length of repeats in 91001

Item	Tandem repeat	Insert sequence	Large segmental duplication	Long terminal repeat	Clustered regularly interspaced short palindromic repeat
Total number	289	145	147	5	3
Total length (bp)	37,147	169,793	253,081	58,080	1,162

IS1541 ($N = 51$), IS285 ($N = 22$), and IS1661 ($N = 7$), which are consistent with the findings of Eppinger and others.^{24,25} Most of the TRs were less than 100 bp, which is consistent with our previous analysis.²⁶ Three large segmental duplications revealed a high repeat frequency, as they separately carried genes encoding transposases for IS1541, IS100, and IS285 (Figure 3). In total, 235 TRs, five LTRs, three CRISPRs, and 147 large segmental duplications were added to the reannotation results (Tables 2 and 3).

Mobile element reannotation. Prophage annotation. A prophage is a bacteriophage that remains in a noninfectious state within bacterial cells. It is not just a parasite in the bacteria, but an active participant in the physical activities of its host, and it plays an important role in the bacterial life cycle. One prophage that carried seven genes was identified previously in strain 91001, and we identified three additional genome fragments that we classified as prophages. The intact one, named Enterobacteria_phage_SfV, carries 16 genes and has a

TABLE 3
Long terminal repeats in 91001

Position	Length (bp)	Function of gene
c2324869..2342875	18,007	YP_2087, pyridine nucleotide transhydrogenase YP_2088, NAD(P) transhydrogenase subunit alpha YP_2089, hypothetical protein YP_2090, hypothetical protein YP_2091, amino acid antiporter YP_2092, hypothetical protein YP_2093, DNA-binding transcriptional regulator RstA YP_2095, transposase for IS100 YP_2096, transposase/IS protein YP_2097, carboxypeptidase YP_2098, transposase for the IS1541 insertion element YP_2099, insecticidal toxin complex protein YP_2100, DNA gyrase (topoisomerase II) B subunit
3248621..3254234	5,614	YP_2920, hypothetical protein YP_2921, hypothetical protein YP_2922, Immunity protein 38 YP_2923, hemagglutinin/hemolysin
4000815..4013575	12,761	YP_3515, Zn-dependent protease with chaperone function YP_3516, transketolase YP_3517, transposase for the IS1541 insertion element YP_3518, erythrose-4-phosphate dehydrogenase YP_3519, phosphoglycerate kinase YP_3520, fructose-bisphosphate aldolase YP_3521, mechanosensitive ion channel YP_3523, transposase/IS protein YP_3524, transposase for insertion sequence IS100 YP_3525, transposase for the IS1541 insertion element
4013424..4025651	12,228	YP_3526, sulfatase YP_3527, DeoR family regulatory protein YP_3528, tagatose 6-phosphate kinase YP_3529, phosphosugar isomerase YP_3530, PTS transport protein YP_3531, PTS permease YP_3532, PTS permease YP_3533, PTS, mannose-/fructose-specific component IIA YP_3534, 2-deoxy-D-gluconate 3-dehydrogenase YP_3535, 2-deoxy-D-gluconate 3-dehydrogenase YP_3536, glycosyl hydrolase family 88
c4479494..4488963	9,470	YP_3936, ImpA domain protein YP_3937, transcriptional regulator YP_3938, toxin SymE, type I toxin-antitoxin system YP_3939, toxin SymE, type I toxin-antitoxin system YP_3940, SMI1/KNR4 family (SUKH-1) YP_3941, Rhs family protein YP_3942, Rhs repeat YP_3943, hypothetical protein YP_3944, Rhs accessory genetic element

IS = Insertion sequence; NAD(P) = nicotinamide adenine dinucleotide phosphate; PTS = phosphotransferase system.

FIGURE 4. Number of genes in genomic islands (GIs), and lengths of GIs in strain 91001. There were 16 GIs, which carried fewer than 10 genes, and 17 GIs were shorter than 10 kb.

total length of 12.4 kb. The other two were incomplete prophages, named *Enterobacteria_phage_PsP3* and *Lactococcus_phage_bIL312*, and they carried 15 and 14 genes with lengths of 8.8 and 12.3 kb, respectively. Notably, one proto-spacer that is associated with CRISPR elements was located in the *Enterobacteria_phage_PsP3* prophage,^{27,28} and the information regarding the proto-spacer was included in the annotation.

Genomic island annotation. A genomic island (GI) is a large genomic region that contains multiple genes probably acquired via horizontal transfer. A GI may be associated with a variety of biological functions, symbiotic or pathogenic mechanisms, or adaptation. We used the IslandViewer database,²⁹ which integrates information from IslandPick, IslandPath-DIMOB, and SIGI-HMM, to find GIs in the 91001 genome. Ten genomic fragments were listed in the pre-

vious annotation, and we identified 24 new GIs, the majority of which contained fewer than 20 genes; the average length of the GIs was 10,208 bp (Figure 4). There are 16 GIs in strain CO92, six of which had been found in the previous annotation of strain 91001. Now, the other 10 GIs in strain CO92 have been annotated in strain 91001.

Trans-omics database system. Using the same reannotation pipeline in strain 91001, we revised the annotation of the other 11 complete genome maps of *Y. pestis* (Supplemental Table 3). To deposit the revised annotations and facilitate the application of these results, we constructed an online trans-omics database system of *Y. pestis* (TODY) using MySQL and Python. The system provides data browsing and downloads, as well as tools for homology fragment analyses. TODY also contains expression profile data from microarray

TABLE 4
Tools and databases

Reannotation items	Tools/data/database	Reference/links	Localization/online
CDS adjustment	GLIMMER 3.02	http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi	Localization
	Prodigal 2.6	http://compbio.ornl.gov/prodigal/	Localization
	GeneMarkS 4.1	http://topaz.gatech.edu/GeneMark/	Localization
Translation initiation sites	MS data		Localization
Pseudogene	MS data		Localization
Homology analysis	BLAST 2.2.28	http://blast.ncbi.nlm.nih.gov/Blast.cgi	Localization
Function	Gene Ontology	http://geneontology.org/	Localization
	NR (Feb 3.5.0)	ftp://ftp.ncbi.nlm.nih.gov/refseq/	Localization
	Pfam 27.0 (May 2015)	http://pfam.xfam.org/	Localization
	PANTHER 9.0	http://www.pantherdb.org/	Localization
	TIGRFAM 15.0	http://www.jcvi.org/cgi-bin/tigrfams/index.cgi	Localization
	HapMap 2015-02-04	http://hapmap.ncbi.nlm.nih.gov/	Localization
	Prosite patterns 20.113	http://prosite.expasy.org/	Localization
	SUPERFAMILY 1.75	http://www.supfam.org/SUPERFAMILY/	Localization
	PRINTS 42.0	http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/index.php	Localization
	Gene3D 3.5.0	http://gene3d.biochem.ucl.ac.uk/	Localization
	Prosite profiles 20.113	http://prosite.expasy.org/	Localization
	COILS	http://www.ch.embnet.org/software/COILS_form.html	Localization
	PIRSF 3.01	http://www.uniprot.org/	Localization
	SMART 6.2	http://smart.embl-heidelberg.de/	Localization
ncRNA	Rfam 12.0 (July 2014)	http://rfam.xfam.org/	Localization
Repeat sequence	RepeatMasker 4.0.6	http://www.repeatmasker.org/	Localization
	Tandem Repeat Finder 4.0.7	http://tandem.bu.edu/trf/trf.html	Localization
	LTR_Finder 1.0.5	http://tlife.fudan.edu.cn/ltr_finder/	Online
	ISFinder (September 2015)	https://www-is.biotoul.fr/	Online
	CRISPRfinder	http://crispr.u-psud.fr/Server/	Online
Prophage	PHAST (November 2014)	http://phast.wishartlab.com/	Online
Genomic island	IslandViewer 3.0	http://www.pathogenomics.sfu.ca/islandviewer/	Online

Localization refers to the tool that can be run locally, and online refers to the tool that can only be run on the web.

hybridization experiments, which were performed under various conditions, and the MS data for strain 91001. More experimental data from the literature will be included in the database in the future.

MATERIALS AND METHODS

Data source. Complete genomes of *Y. pestis* (Supplemental Table 3) and the NR database were downloaded from the NCBI. Strain 91001, which is avirulent to humans,⁴ was isolated from a Brandt's vole (*Microtus brandti*) in Inner Mongolia, China. MS data from this strain were acquired by the TMT

method³⁰ in 2015. The 91001 sample was labeled by stable isotope labeling by amino acids in cell culture (SILAC) method.³¹ Heavy isotope labeled amino acids were Lys (CNLM-291-H-0.25) and Arg (CNLM-539-H-0.25). The expression profile data were obtained from microarray hybridization experiments,³² and the sRNA data were obtained from RNA-seq results.¹⁸ In addition, data from related literatures, including MS data from the KIM, CO92, and PestoidesF strains, and sRNA data from the 91001 and KIM strains,^{33,34} were collected manually.

Tools and databases. Gene prediction was conducted with GLIMMER 3.02,³⁵ Prodigal 2.6,³⁶ and GeneMarkS 4.1.³⁷

FIGURE 5. Reannotation pipeline for *Yersinia pestis*. The first step is to generate reannotation data. The second step is to structure these results and construct a database.

Protein functional annotation was conducted with the NCBI NR database, gene ontology database³⁸ and the InterProScan³⁹ platform, which integrates information from other databases, including Pfam,⁴⁰ PANTHER,⁴¹ TIGRFAM,⁴² HAMAP,⁴³ Prosite patterns,⁴⁴ SUPERFAMILY,⁴⁵ PRINTS,⁴⁶ Gene3D,⁴⁷ Prosite profiles,⁴⁴ ProDom,⁴⁸ SMART⁴⁹ and Coils.⁵⁰ ncRNAs were annotated with Rfam.⁵¹ Repeat sequences were annotated with NCBI BLASTN,⁵² RepeatMasker,⁵³ Tandem Repeat Finder,⁵⁴ LTR-finder,⁵⁵ ISFinder,⁵⁶ and CRISPRfinder.⁵⁷ Prophages and GIs were annotated with the PHAST database⁵⁸ and IslandViewer,²⁹ respectively (Table 4).

The MS spectra from 91001 samples were searched with the Sorcerer-SEQUEST (version 4.0.4 build; Sage-N Research, Inc., Milpitas, CA) against the composite target/decoy database to estimate false discovery rate.^{59,60} The spectral interpretation rate was 75% (the number of identified spectra/total spectra = 28,089/37,558), protein identification rate was 68% (the number of identified protein/total protein count = 2,825/4,136) (<http://tody.bmi.ac.cn/download/> identified by MS data).

Identification of homologous genes. There are arguments over the criteria that are used to identify homologous genes. By loosening sequence consistence criteria (lower identity and coverage values), more potential homologous genes could be identified, and the rate of detecting false negatives could be decreased; however, this would increase the number of possible false-positive results. As the genome of *Y. pestis* is highly conserved, the number of homologous genes aligned by BLASTN only decreased by approximately 2% when the nucleotide acid identity threshold increased from 80% to 95%, and very few homologous alleles were missed during detection. Hence, we tended to select strict criteria to increase the reliability of the results, that is, homologous genes across different strains of *Y. pestis* were identified through BLASTN using the following thresholds: e-value < 10⁻⁵, identity > 95%, and aligned length > 80% of the total length. The method of core gene set construction had been put forward along with Eppinger and others' method.²⁴

Definition of a reliable gene set by integrating the results from different gene prediction tools. Gene prediction is fundamental to genome annotation, but different tools usually generate different results because they use different algorithms. The widely used gene prediction software GLIMMER is based on an interpolated Markov model. GeneMarkS uses a hidden Markov model and an iterative self-learning algorithm for gene prediction, whereas Prodigal is based on a dynamic programming algorithm. Different versions of the same software may generate different prediction results. For example, GLIMMER 1.0⁶¹ was first released in 1998. In 2007, the GLIMMER 3.0⁶² update contained some major improvements compared with the original version, such as supporting longer open reading frames, as well as ribosome binding site and overlapping gene predictions. In addition, the accuracy of the genome annotation results is affected by the parameter settings and database updates.

The prediction results for *Y. pestis* genomes using GLIMMER (threshold: overlap number = 1, gene length > 100, score > 30), GeneMarkS (parameter: prok, combine mode), and Prodigal varied, especially regarding the positions of TISs (Supplemental Table 4). Therefore, a predicted gene was defined as highly reliable only when the three prediction

tools presented fully consistent results (i.e., both the predicted 3' and 5' ends of the gene were consistent). In addition, this reliable gene set was used to adjust the positions of the TISs of 2,302 genes.

***Y. pestis* genome reannotation pipeline.** The de novo annotation of a new genome generally begins with gene prediction using prediction tools.⁶³ Then, predicted genes are aligned to a reference genome, and protein databases based on nucleotide or amino acid sequence homologies are used to infer gene products and functions. Finally, the annotations are complemented with experimental data or manual revisions. To reannotate strain 91001, we generated a pipeline that could be applied to other *Y. pestis* genomes (Figure 5). Our reannotation pipeline is divided into two major steps. The first step is data preprocessing. In this step, CDSs and the reliable gene set were determined based on the gene prediction results. Then, the allele genotype set was built after performing the homology and functional analyses. Finally, TISs and pseudogenes were reannotated according to the screened results obtained from the homologous gene data, MS data, and reliable gene prediction sets. In addition, repeat sequences, mobile elements, prophages, and GIs were reannotated across the whole genome, and ncRNAs were identified in noncoding regions. The second step is to structure the reannotation results. Because a variety of computational tools and public databases are incorporated in this process, preprocessed data must be screened, reduced, corrected, classified, and standardized. Then, processed, structured data are integrated to generate the final reannotation results. The pipeline was applied to 12 complete genome maps of *Y. pestis* strains, and the results were imported into the TODY database. In addition to the TIS and pseudogene reannotations, other preprocesses were automatic, and some computing processes occurred in parallel; manual work was indispensable for the reannotations.

DISCUSSION

In this study, we integrated all of the available genomic data, as well as data from public databases, experimental proteomic data, and data from the literature to systematically reannotate the CDSs, TISs, pseudogenes, ncRNAs, repeat sequences, and mobile elements in strain 91001. Compared with the previous annotation, the updated version removed 137 hypothetical genes, corrected the positions of 41 TISs, and identified seven pseudogenes and 392 genes with hypothetical functions. It added 272 ncRNAs, 230 repeat sequences, three prophages, and 24 GIs, and it improved protein annotation by integrating information from multiple databases. Totally 89% of the whole genome sequence was annotated as possible coding regions in strain 91001, which will facilitate biological studies of *Y. pestis*. We also built a reannotation pipeline to analyze other publicly released *Y. pestis* genomes and to create online tools to use this reannotation information.

Notably, the functions of around one-seventh (619/3,999) of the genes, occupying 10.7% of the genome, are still unknown, and these genes were annotated as "hypothetical" or "putative." Although some genes may have resulted from false-positive prediction results caused by overprediction of the bioinformatics tools,²⁷ many of them were predicted to perform essential functions, although this will need to be

confirmed by experimental studies. Therefore, this reannotation will be a persistent work in progress, and it will increase our knowledge of bacteria. In addition, this reannotation was not just an in silico experiment that was conducted using bioinformatics tools, because it required a large amount of experimental data to increase its credibility. However, because much of the experimental data have not been fully digitized, formatted, structured, and standardized, the development of artificial intelligence to collect and process experimental data should be used to improve the automatic reannotation of *Y. pestis*.

Received March 16, 2016. Accepted for publication May 17, 2016.

Published online July 5, 2016.

Note: Supplemental tables appear at www.ajtmh.org.

Acknowledgments: We thank Xianwei Yang and Yang Liu for bioinformatics instructions, Yanfeng Yan for providing genomic data, Zongmin Du and Lei Zhou for providing proteogenomics data, Yanping Han for providing transcriptomics data. We also thank all of our colleagues from the State Key laboratory of Pathogen and Biosecurity, and the Information Technology Center of Beijing Institute of Health and Medical Information.

Authors' addresses: Yiqing Mao, Xianwei Yang, Yanfeng Yan, Zongmin Du, Yanping Han, Yajun Song, Lei Zhou, Yujun Cui, and Ruifu Yang, State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, Beijing, P. R. China, E-mails: polo_simon@sina.com, 715693659@qq.com, yanfengyan@yahoo.com.cn, zongmindu@gmail.com, yanpinghan@gmail.com, songyajun88@gmail.com, zhoulei@bmi.ac.cn, cuiyujun.new@gmail.com, and ruifuyang@gmail.com. Yang Liu, Department of Biotechnology, Beijing Institute of Radiation Medicine, Beijing, P. R. China, E-mail: liuyang@bmi.ac.cn.

REFERENCES

- Brubaker RR, 2002. *Yersinia pestis*. Sussman M, ed. *Molecular Medical Microbiology*. London, United Kingdom: Academic Press.
- WHO, 2015. *Plague: Disease Outbreak News*. Available at: <http://www.who.int/csr/don/archive/disease/plague/en/>. Accessed May, 2015.
- Parkhill JWB, Thomson NR, Titball RW, Holden MT, Prentice MB, Sebaihia M, James KD, Churcher C, Mungall KL, Baker S, Basham D, Bentley SD, Brooks K, Cerdeño-Tarraga AM, Chillingworth T, Cronin A, Davies RM, Davis P, Dougan G, Feltwell T, Hamlin N, Holroyd S, Jagels K, Karlyshev AV, Leather S, Moule S, Oyston PC, Quail M, Rutherford K, Simmonds M, Skelton J, Stevens K, Whitehead S, Barrell BG, 2001. Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* 413: 523–527.
- Song Y, Tong Z, Wang J, Wang L, Guo Z, Han Y, Zhang J, Pei D, Zhou D, Qin H, Pang X, Han Y, Zhai J, Li M, Cui B, Qi Z, Jin L, Dai R, Chen F, Li S, Ye C, Du Z, Lin W, Wang J, Yu J, Yang H, Wang J, Huang P, Yang R. 2004. Complete genome sequence of *Yersinia pestis* strain 91001, an isolate avirulent to humans. *DNA Res* 11: 179–197.
- Jaffe JD, Berg HC, Church GM, 2004. Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* 4: 59–77.
- Ouzounis CA, Karp PD, 2002. The past, present and future of genome-wide re-annotation. *Genome Biol* 3: comment2001.1–comment2001.6.
- Camus JC, Pryor MJ, Médigue C, Cole ST, 2002. Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology* 148: 2967–2973.
- Gundogdu O, Bentley SD, Holden MT, Parkhill J, Dorrell N, Wren BW, 2007. Re-annotation and re-analysis of the *Campylobacter jejuni* NCTC11168 genome sequence. *BMC Genomics* 8: 162.
- Guo FB, Xiong L, Teng JL, Yuen KY, Lau SK, Woo PC. 2013. Re-annotation of protein-coding genes in 10 complete genomes of Neisseriaceae family by combining similarity-based and composition-based methods. *DNA Res* 20: 273–286.
- Powell S, Forslund K, Szklarczyk D, Trachana K, Roth A, Huerta-Cepas J, Gabaldon T, Rattei T, Creevey C, Kuhn M, Jensen LJ, von Mering C, Bork P, 2014. eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res* 42: D231–D239.
- Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muniz-Rascado L, Garcia-Sotelo JS, Weiss V, Solano-Lira H, Martinez-Flores I, Medina-Rivera A, Salgado-Osorio G, Alquicira-Hernandez S, Alquicira-Hernandez K, Lopez-Fuentes A, Porron-Sotelo L, Huerta AM, Bonavides-Martinez C, Balderas-Martinez YI, Pannier L, Olvera M, Labastida A, Jimenez-Jacinto V, Vega-Alvarado L, Del Moral-Chavez V, Hernandez-Alvarez A, Morett E, Collado-Vides J, 2013. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res* 41: D203–D213.
- Narsai R, Devenish J, Castleden I, Narsai K, Xu L, Shou H, Whelan J. 2013. Rice DB: an Oryza Information Portal linking annotation, subcellular location, function, expression, regulation, and evolutionary information for rice and Arabidopsis. *Plant J* 76: 1057–1073.
- Sass S, Buettner F, Mueller NS, Theis FJ, 2015. RAMONA: a Web application for gene set analysis on multilevel omics data. *Bioinformatics* 31: 128–130.
- Fisch KM, Meissner T, Gioia L, Ducom JC, Carland TM, Loguerio S, Su AI, 2015. Omics Pipe: a community-based framework for reproducible multi-omics data analysis. *Bioinformatics* 31: 1724–1728.
- Peterson ES, McCue LA, Schrimpe-Rutledge AC, Jensen JL, Walker H, Kobold MA, Webb SR, Payne SH, Ansong C, Adkins JN, Cannon WR, Webb-Robertson BJ, 2012. VESPA: software to facilitate genomic annotation of prokaryotic organisms through integration of proteomic and transcriptomic data. *BMC Genomics* 13: 131.
- Schrimpe-Rutledge AC, Jones MB, Chauhan S, Purvine SO, Sanford JA, Monroe ME, Brewer HM, Payne SH, Ansong C, Frank BC, Smith RD, Peterson SN, Motin VL, Adkins JN, 2012. Comparative omics-driven genome annotation refinement: application across Yersiniaceae. *PLoS One* 7: e33903.
- Payne SH, Huang ST, Pieper R, 2010. A proteogenomic update to *Yersinia*: enhancing genome annotation. *BMC Genomics* 11: 460.
- Yan Y, Su S, Meng X, Ji X, Qu Y, Liu Z, Wang X, Cui Y, Deng Z, Zhou D, Jiang W, Yang R, Han Y, 2013. Determination of sRNA expressions by RNA-seq in *Yersinia pestis* grown in vitro and during infection. *PLoS One* 8: e74495.
- Zhou L, Ying W, Han Y, Chen M, Yan Y, Li L, Zhu Z, Zheng Z, Jia W, Yang R, Qian X, 2012. A proteome reference map and virulence factors analysis of *Yersinia pestis* 91001. *J Proteomics* 75: 894–907.
- Lerat E, Ochman H, 2005. Recognizing the pseudogenes in bacterial genomes. *Nucleic Acids Res* 33: 3125–3132.
- Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M, 2016. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 44: D457–D462.
- Eddy SR, 2001. Non-coding RNA genes and the modern RNA world. *Nat Rev Genet* 2: 919–929.
- Wittkopp PJ, Kalay G, 2012. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet* 13: 59–69.
- Eppinger M, Worsham PL, Nikolich MP, Riley DR, Sebastian Y, Mou S, Achtman M, Lindler LE, Ravel J, 2010. Genome sequence of the deep-rooted *Yersinia pestis* strain Angola reveals new insights into the evolution and pangenome of the plague bacterium. *J Bacteriol* 192: 1685–1699.
- Eppinger M, Rosovitz MJ, Fricke WF, Rasko D, Kokorina G, Fayolle C, Lindler LE, Carniel E, Ravel J. 2007. The complete genome sequence of *Yersinia pseudotuberculosis* IP31758, the causative agent of Far East scarlet like fever. *PLoS Genet* 3: e142.
- Li Y, Cui Y, Cui B, Yan Y, Yang X, Wang H, Qi Z, Zhang Q, Xiao X, Guo Z, Ma C, Wang J, Song Y, Yang R, 2013. Features of variable number of tandem repeats in *Yersinia pestis*

- and the development of a hierarchical genotyping scheme. *PLoS One* 8: e66567.
27. Pourcel C, Salvignol G, Vergnaud G, 2005. CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* 151: 653–663.
 28. Cui Y, Li Y, Gorge O, Platonov ME, Yan Y, Guo Z, Pourcel C, Dentovskaya SV, Balakhonov SV, Wang X, Song Y, Anisimov AP, Vergnaud G, Yang R, 2008. Insight into microevolution of *Yersinia pestis* by clustered regularly interspaced short palindromic repeats. *PLoS One* 3: e2652.
 29. Langille MG, Brinkman FS, 2009. IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics* 25: 664–665.
 30. Aebersold R, Mann M, 2003. Mass spectrometry-based proteomics. *Nature* 422: 198–207.
 31. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M, 2002. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* 1: 376–386.
 32. Zhou D, Han Y, Qiu J, Qin L, Guo Z, Wang X, Song Y, Tan Y, Du Z, Yang R, 2006. Genome-wide transcriptional response of *Yersinia pestis* to stressful conditions simulating phagolysosomal environments. *Microbes Infect* 8: 2669–2678.
 33. Beauregard A, Smith EA, Petrone BL, Singh N, Karch C, McDonough KA, Wade JT, 2013. Identification and characterization of small RNAs in *Yersinia pestis*. *RNA Biol* 10: 397–405.
 34. Koo JT, Alleyne TM, Schiano CA, Jafari N, Lathem WW, 2011. Global discovery of small RNAs in *Yersinia pseudotuberculosis* identifies *Yersinia*-specific small, noncoding RNAs required for virulence. *Proc Natl Acad Sci USA* 108: E709–E717.
 35. Delcher AL, Bratke KA, Powers EC, Salzberg SL, 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 23: 673–679.
 36. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ, 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11: 119.
 37. Besemer JLA, Borodovsky M, 2001. GeneMarkS—a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* 29: 2607–2618.
 38. Gene Ontology Consortium, Blake JA, Dolan M, Drabkin H, Hill DP, Li N, Sitnikov D, Bridges S, Burgess S, Buza T, McCarthy F, Peddinti D, Pillai L, Carbon S, Dietze H, Ireland A, Lewis SE, Mungall CJ, Gaudet P, Chrisholm RL, Fey P, Kibbe WA, Basu S, Siegle DA, McIntosh BK, Renfro DP, Zweifel AE, Hu JC, Brown NH, Tweedie S, Alam-Faruque Y, Apweiler R, Auchincloss A, Axelsen K, Bely B, Blatter M, Bonilla C, Bouguerlet L, Boutet E, Breuza L, Bridge A, Chan WM, Chavali G, Coudert E, Dimmer E, Estreicher A, Famiglietti L, Feuermann M, Gos A, Gruaz-Gumowski N, Hieta R, Hinz C, Hulo C, Huntley R, James J, Jungo F, Keller G, Laiho K, Legge D, Lemerrier P, Lieberherr D, Magrane M, Martin MJ, Masson P, Mutowo-Muilenet P, O'Donovan C, Pedruzzi I, Pichler K, Poggioli D, Porras Millan P, Poux S, Rivoire C, Roechert B, Sawford T, Schneider M, Stutz A, Sundaram S, Tognolli M, Xenarios I, Foulgar R, Lomax J, Roncaglia P, Khodiyar VK, Lovering RC, Talmud PJ, Chibucos M, Giglio MG, Chang H, Hunter S, McAnulla C, Mitchell A, Sangrador A, Stephan R, Harris MA, Oliver SG, Rutherford K, Wood V, Bahler J, Lock A, Kersey PJ, McDowall DM, Staines DM, Dwinell M, Shimoyama M, Lalederkind S, Hayman T, Wang S, Petri V, Lowry T, D'Eustachio P, Matthews L, Balakrishnan R, Binkley G, Cherry JM, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hitz BC, Hong EL, Karra K, Miyasato SR, Nash RS, Park J, Skrzypek MS, Weng S, Wong ED, Berardini TZ, Huala E, Mi H, Thomas PD, Chan J, Kishore R, Sternberg P, Van Auken K, Howe D, Westerfield M, 2013. Gene Ontology annotations and resources. *Nucleic Acids Res* 41: D530–D535.
 39. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R, 2005. InterProScan: protein domains identifier. *Nucleic Acids Res* 33: 116–120.
 40. Finn RD, Clements ABJ, Punta M, 2014. The Pfam protein families database. *Nucleic Acids Res* 40: D222–D230.
 41. Mi H, Poudel S, Muruganujan A, Casagrande JT, Thomas PD, 2016. PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res* 44: D336–D342.
 42. Haft DH, Selengut JD, Richter RA, Harkins D, Basu MK, Beck E, 2013. TIGRFAMs and genome properties in 2013. *Nucleic Acids Res* 41: D387–D395.
 43. Pedruzzi I, Rivoire C, Auchincloss AH, Coudert E, Keller G, de Castro E, Baratin D, Cuche BA, Bouguerlet L, Poux S, Redaschi N, Xenarios I, Bridge A, 2015. HAMAP in 2015: updates to the protein family classification and annotation system. *Nucleic Acids Res* 43: D1064–D1070.
 44. Sigrist CJ, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bouguerlet L, Xenarios I, 2013. New and continuing developments at PROSITE. *Nucleic Acids Res* 41: D344–D347.
 45. Wilson D, Madera M, Vogel C, Chothia C, Gough J, 2007. The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Res* 35: D308–D313.
 46. Attwood TK, Coletta A, Muirhead G, Pavlopoulou A, Philippou PB, Popov I, Roma-Mateo C, Theodosiou A, Mitchell AL, 2012. The PRINTS database: a fine-grained protein sequence annotation and analysis resource—its status in 2012. *Database* 2012: bas019.
 47. Lees JG, Lee D, Studer RA, Dawson NL, Sillitoe I, Das S, Yeats C, Dessailly BH, Rentzsch R, Orengo CA, 2014. Gene3D: multi-domain annotations for protein sequence and comparative genome analysis. *Nucleic Acids Res* 42: D240–D245.
 48. Bru C, Courcelle E, Carrere S, Beausse Y, Dalmar S, Kahn D, 2005. The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res* 33: D212–D215.
 49. Letunic I, Doerks T, Bork P, 2009. SMART 6: recent updates and new developments. *Nucleic Acids Res* 37: D229–D232.
 50. Lupas AVDM, Stock J, 1991. Predicting coiled coils from protein sequences. *Science* 252: 1162–1164.
 51. Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, Eddy SR, Gardner PP, Bateman A, 2013. Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res* 41: D226–D232.
 52. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL, 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10: 421.
 53. Temple S, 2012. Using and understanding RepeatMasker. *Methods Mol Biol* 859: 29–51.
 54. Benson G, 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27: 573–580.
 55. Xu Z, Wang H, 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* 35: W265–W268.
 56. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M, 2006. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res* 34: D32–D36.
 57. Grissa I, Vergnaud G, Pourcel C, 2007. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* 35: W52–W57.
 58. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS, 2011. PHAST: a fast phage search tool. *Nucleic Acids Res* 39: W347–W352.
 59. Ping L, Zhang H, Zhai Dammer EB, Duong DM, Li N, Yan Z, Wu J, Xu P, 2013. Quantitative proteomics reveals significant changes in cell shape and an energy shift after IPTG induction via an optimized SILAC. *J Proteome Res* 12: 5978–5988.
 60. Elias JE, Gygi SP, 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* 4: 207–214.
 61. Salzberg SL, Delcher AL, Kasif S, White O, 1998. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res* 26: 544–548.
 62. Delcher AL, Bratke KA, Powers EC, Salzberg SL, 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 23: 673–679.
 63. Richardson EJ, Watson M, 2012. The automatic annotation of bacterial genomes. *Brief Bioinform* 14: 1–12.