Complete Genome Sequence of *Escherichia Coli* Strain S9922 Isolated from Meningitis in Calf and Comparative Genomic Analysis with Other *E. Coli* Pathotypes

Beibei Li
Shihezi University
Jingjing Ren
Shihezi University
Xun Ma
Shihezi University
Qian Qin
Shihezi University
Xinyu Wang
Shihezi University
Jianjun Jiang (jiangjianjun7788@163.com)
Shihezi University
Pengyan Wang
Shihezi University

Research Article

Keywords: Extraintestinal pathogenic Escherichia coli、Whole genome sequencing、Comparative genomic analysis

Posted Date: September 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-846112/v1

Abstract

Background: Extraintestinal pathogenic Escherichia coli (ExPEC) exists in the normal intestinal flora, but can invade and colonize extraintestinal sites and cause a wide range of infections. Genomic analysis of ExPEC has mainly focused on isolates of human, poultry and pig. In recent years, some large-scale dairy farms in Xinjiang broke out cases characterized by neurological symptoms and acute death in newborn calves. To better understand the genomic attributes underlying the pathogenicity of bovine-source ExPEC, a highly virulent strain, which named E. coli S9922 was isolated from cerebral effusion in a calf that died of meningitis, was sequenced and analyzed.

Results: Using single-molecule sequencing technology on PacBio and then assembled, the genes were predicted and annotated. The whole genome of E.coli S9922 was consisted of a chromosome and three plasmids containing 5055 genes, and the total length was 5269374 bp and the average G+C content was 50.82%. In addition, 291 host-, 204 virulence-, and 185 resistance-related genes, and 182 T3SS effector proteins were found by comparison with related databases. Comparison of this genome to 16 representative strains of pathogenic E.coli genomic sequences showed that E.coli S9922 had the greatest co-linearity with E.coli 90-9272. In addition, Core genes obtained by cluster analysis of E.coli S9922 homologous genes were classified, a total of 2570, 2780, and 2188 genes were obtained via COG, KEGG, and GO comparisons, respectively. The unique genes identified by homologous cluster analysis were classified 204, 550, 239 genes in COG, KEGG, and GO comparisons, respectively. Evolutionary tree analysis revealed a close evolutionary relationship between E.coli S9922 and E.coli 90-9272, and a distant relationship between E.coli S9922 and UTI89.

Conclusions: The study provide dgenomics of E.coli S9922 strain from the cattle that had died of meningitis. It enriched the genome data of E.coli and laid a theoretical foundation for further experimental study of ExPEC. Comparative genomics analysis showed that E.coli S9922 had a close evolutionary relationship with E.coli 90-9272, but far from that of UTI89.

Background

Escherichia coli (*E.coli*) is an important component of the intestinal flora of humans and animals. *E. coli*, which causes disease in humans and animals, is called pathogenic *E.coli*. According to the clinical pathogenic characteristics, pathogenic *E.coli* can be divided into intestinal pathogenic *E.coli* (IPEC) and extraintestinal pathogenic *E.coli* (ExPEC) [1]. ExPEC is a type of *E.coli* flora that can specifically colonize other tissues outside the host's intestine and cause serious pathogenesis in the host, including neonatal meningitis pathogenic *E. coli* (NMEC), urethra pathogenic *E.coli* (UPEC) avian pathogenic *E.coli* (APEC) [2]. In veterinary medicine, it can cause pneumonia, nephritis, sepsis, avian balloon inflammation, synovitis, peritonitis, and other diseases [3].

To date, domestic and foreign research on ExPEC focuses on human, poultry, and pig sources of ExPEC. In recent years, an outbreak in some large-scale dairy farms in Shihezi, Xinjiang in Northwest China was characterised by neurological symptoms and acute death in new-born calves. Most of the affected new-born calves aged 1–7 days, especially those aged within three days or 24 hours after birth, and mainly showed typical high fever, lethargy, gait stiffness, head and neck stiffness, and even some neurological symptoms, and a severe haemorrhage of the meninges was found by necropsy. The antibiotic treatment that was administered was not effective. The strain isolated from the brain tissue of the dead calves was initially identified as ExPEC and named *E.coli S9922*, the isolate has extremely high drug resistance, which brings challenges to the prevention and control of infection.

Some virulence-associated genes are shared among ExPEC independent of host species, suggesting that ExPEC may be zoonotically acquired [1]. The whole genome sequence of bovine-source ExPEC has not been published, which limits the study on the study on its functional genes, metabolite synthesis pathways, and comparative genomics. In order to further study of mechanisms of bovine-source ExPEC, we present here the genome sequence of *E.coli S9922* using single-molecule sequencing technology on PacBio. In addition, comparative genomic analyses with other characteristic *E.coli* strains to identify genomic differences. The study can lay a foundation for the exploration and utilisation of the pathogenic mechanism and functional genes of bovine-source ExPEC.

Results

General features of *E.coli S9922* isolates from calf encephalitis

ExPEC is a foodborne pathogen that can cause urethritis, infectious pneumonia, neonatal meningitis, cow mastitis and so on, which seriously threatens the health of human and animals. *E.coli S9922* was isolated from cerebral effusion in a calf that died of meningitis. However, the character of clinical isolate remains to be determined. We analyzed the isolate in this study by genomic sequencing using the PacBio platform (Novo Zhiyuan Technology, Beijing). The draft genome sequences had been deposited in the NCBI Sequence ReadArchive under the accession number: SRA8740874. The genome of *E.coli S9922* comprises a circular chromosome of 4,960,912bp with an average G+C content of 50.7%, and three circular plasmid (Plas1:110,372 bp, Plas2:73,649 bp, Plas3:77,569 bp) with an average G +C content of 49.62% , 52.87% and 51.82%, respectively (Fig.1). General genome information including Genome size, G+C content, Average gene size, Number of contings, Number of rRNA/tRNA/SRNA, Number of GIs were shown in Table 1.

Gene islands (GIs)

GIs comprise a horizontally acquired flexible gene pool that is a major driver in evolution and niche specialization of pathogenic bacteria. We used IslandPath-DIOMB (Version 0.2) to identify putative genomic islands in *E.coli S9922*. The results showed that there were 23 GIs in the *E.coli S9922* genome, 20 of which are located on chromosomes, two on plasmid 1, and one on plasmid 2. The total length of the GIs is 313608 bp, and that of the average length is 13635 bp, and they were shown in Fig. 2.

Virulence and pathogenicity analysis

T3SS of Gram-negative bacteria is commonly used to study pathogens, infection mechanisms, virulence effects at the molecular level. The results showed that there were 5055 encode genes identified in the T3SS of *E.coli S9922*, including 182 T3SS effector proteins and 4873 non-T3SS effector proteins.

The *E.coli S9922* genome had a total of 291 genes annotated in PHI, of which 200 are attenuated virulence genes (Fig.3). Using Diamond software, the amino acid sequence of *E. coli S9922* was compared with the VFDB database, and 204 genes related to the virulence were found, mainly including: lipopolysaccharide (LPS) (10 genes), flagella (40 genes), type IV pili (12 genes), type I pili (19 genes), TTSS (Virulence-related factors, such as SPI-1) (eight genes), enterobacteria (14 genes), and VAS cluster (eight genes)

The ARDB database contains the names of drug resistance-related genes which can be found through the annotation of the database. The amino acid sequence of *E.coli S9922* was compared with the ARDB database. *E.coli S9922* was resistant to adriamycin, erythromycin, vancomycin, teicoplanin, penicillin, bacitracin, aminoglycosides, glycylcycline, macrolides, β-lactams, acridine yellow, tetracycline, streptomycin, isepamicin, netilmicin, tobramycin, kanamycin, sisomicin, dibekacin, chloramphenicol, monoamimycin, cephalosporin, gentamicin, phosphomycin, polymyxin, enrofloxacin, norfloxacin, deoxycholic acid, tigecycline, fluoroquinolone, phosphoamimycin, glycyl, kasugamycin, chloride, puromycin, trimethoprim, and sulphonamide.

Comparative genomics analysis

E.coli S9922 genome was compared to 16 reference *E.coli* strains, the main features of the above *E.coli* strains are shown in Table 2. *E.coli S9922* was similar to the other *E.coli* with G+C content between 50.5–50.8%, The majority of ExPEC belong to phylogenetic group B2 including UPEC, APEC and NMEC.

Collinearity refers to the linkage relationship between genes, which is a phenomenon where homologous genes are arranged in the same order within the genomes of different species. The degree of collinearity between two species can be used as a measure of the evolutionary distance between them, helping to understand the relationship between species. In order to

further determine the genomic differentiation and similarities between *E.coli S9922* and the other strains, we performed the Whole-genome collinearity analysis, these results were displayed in Fig. 4. It can be seen that *E.coli S9922* and *E.coli 90-9272* had the highest number of collinear fragments, indicating that *E.coli S9922* and *E.coli 90-9272* had the least evolutionary distance and closest genetic relationship. PCN061 was second most closely related to *E.coli S9922*, and *E.coli S9922* had the least collinear fragment and the most farthest related to UTI89.

Population evolution analysis

Core genes and specific genes are likely to correspond to the commonality and characteristics of *E.coli S9922*, thus serving as the basis for the study of the functional differences between samples. Compared with the genes of reference strains, *E.coli S9922* obtained 2780 core genes, 6586 non-genes, and 669 specific genes by cluster analysis. Among them, by comparing COG, KEGG and GO, core genes homologous to *E.coli S9922* were obtained, which were 2570, 2780 and 2188 genes, respectively. The unique genes were classified as follows: 204, 550, 239 genes in COG, KEGG, and GO comparisons, respectively, in shown in Additional file 1.

In genomic evolution, a gene family is a group of genes with a common ancestor. Different genes in a family often have a similar structure and function, which can serve as the basis for identifying unknown gene function and provide clues regarding gene evolution history. Pairwise comparison was made between *E.coli S9922* and other strains to obtain gene family clustering results, as shown in Table 3 and Fig. 5.

Among the gene family clustering results obtained by pairwise comparisons between *E.coli S9922* and other strains, the number of unclustered genes was the highest in the *E.coli* S9922 strain. The number of *E.coli S9922* specific gene families was 35.

By comparing *E.coli S9922* with other strains, the evolutionary tree was obtained as shown in Fig. 6. It can be seen that the evolutionary relationship between *E.coli S9922* and *90-9272* is relatively close, while the evolutionary relationship between *E.coli S9922* and UTI89 is distant.

Discussion

Pathogenic *Escherichia coli*, as an important zoonotic pathogen, seriously threatens the health of human and animals. Among them, ExPEC can cause extraintestinal infection in a wide range of hosts, inducing meningitis, arthritis, pneumonia, endocarditis and so on [4]. Xinjiang is a large province of cattle and sheep farming in China and diseases caused by ExPEC have also occurred from time to time in recent years. Nowadays, most researchers mainly study the ExPEC of poultry source and human source, and few reports on bovine-source ExPEC were reported. Therefore, a better understanding of bovine-source ExPEC genomics is needed to improve our understanding of its mechanisms of disease. In this study, the whole genome sequence of *E.coli S9922* and the comparison with other characteristic *E.coli* strains were analysed to lays a foundation for the next study of its pathogenic mechanism.

Whole genome sequencing analysis

In this study, the genomic DNA of the isolate was sequenced using PacBio platform. Through the genome-wide sequencing of *E.coli S9922*, a large amount of data were generated and assembled. The N50 value is an index for evaluating the assembly of the sequence. In general, the larger the N50 value, the better the assembly [5]. The contig N50 value of the *E.coli S9922* genome was 4,948,664 bp, indicating that the assembly data was good. In this study, the sequencing depth was greater than 150X which could ensure the accuracy of genomic analysis. The genome of *E.coli S9922* comprises a circular chromosome and three circular plasmid, *E.coli S9922* genome was compared to other *E.coli* (Table 2), *E.coli S9922* was similar to the other *E.coli* with G+C content between 50.5–50.8%.

Genomic islands (GIs) comprise a horizontally acquired flexible gene pool that is a major driver in evolution and niche specialization of pathogenic bacteria [6]. To identify putative genomic islands in *E.coli S9922*, we used IslandPath-DIMOB for

island prediction, we identified 23 genomic islands on the *E.coli S9922*, 20 of which are located on chromosomes, two on plasmid 1, and one on plasmid 2. It is speculated that the virulence gene of *E.coli S9922* may be obtained through horizontal transfer. There are 10 pre phagosomes in the *E.coli S9922* genome, one on plasmid 2 and nine on the chromosome.

Pathogenicity analysis

Secretion systems are necessary for the transport of proteins across the cell envelope, mediating interactions between bacteria, their hosts, and the surrounding micro-environment [7]. Secretory proteins play an important role in the pathogenesis of bacteria [8]. The analysis and prediction of secretory proteins of *E.coli S9922* can lay a foundation for the complete and systematic study of the molecular pathogenesis of *E. coli*. The type III secretory system (T3SS) is the main transport pathway from the secretion protein of Gram-negative bacteria to the extracellular environment. Therefore, the T3SS effector protein is related to the pathogenesis of Gram-negative bacteria. In this study, 323 secretory proteins and 182 T3SS effector proteins were predicted. Through the comparison of the PHI database, 291 genes related to host action were annotated, including 22 genes with super virulence. The virulence-related factors, such as lipopolysaccharide (LPS), flagella, type IV pili, type I pili, TTSS (SPI-1 encode), enterobacteria, eosinophil cationic protein, membrane, vas cluster, and T2SS were found through VFDB annotation. These virulence genes have a strong correlation with the pathogenicity of *E.coli*.

At present, the prevention and treatment of *E. coli* is based on antibiotics. However, with the emergence of antibiotic-resistant strains, it has brought difficulties to clinical treatment, and antibiotic-resistant strains can also be spread through various ways to affect the public health safety of humans and animals [9-10]. *E.coli* that produces extended-spectrum β-lactamase is widely spread, and the isolation ratio of Ampicillin-resistant ExPEC strains from poultry and pigs is increasing. In addition, its resistance to quinolone and aminoglycoside drugs also showed an upward trend, and the isolation of super bacteria with multi-drug resistance is gradually increasing [11-12]. Therefore, it is of great significance to understand the drug-resistant phenotypes of local pathogenic bacteria and to guide clinical medication to prevent and control related bacterial diseases. *E.coli S9922* has 185 resistance-related genes and was found to be resistant to quinolones, β-lactams, macrolides, aminoglycosides, and so on, and was also found to be a very resistant strain. It reported that several ExPEC strains isolated from Xinjiang sheep were multi-drug resistant. And the resistance rate to first-generation cephalosporins, aminoglycosides, tetracyclines and quinolones is generally high [13]. Six strains of pathogenic *E.coli*, isolated from infected calves, were multi-drug resistant, the drug resistance rates of penicilli, amoxicillin, cefazolin, cefradine, streptomycin, gentamicin, florfenicol, teicoplanin, clindamycin, meropenem, tetracycline and tobramycin were 100% [14].

Comparative genomics analysis

E.coli S9922 was compared with their pathogenic *E. coli* strains published in NCBI gene database. The hosts of these other reference strains included swine, avian, and human sources. The isolation sites included urethra, intestinal tract, mammary gland, brain, and several other locations. The results of genome-wide collinearity analysis showed that the evolutionary relationship between *E.coli S9922* and *E.coli* 90-9272 was close, and that between *E.coli S9922* and *E. coli* UTI89 was distant, which was consistent with the results of the evolutionary tree analysis. It was determined from the distance of genetic relationship between *E.coli S9922* and other strains that the evolution of *E.coli* S9922 was not significantly related to the host, considering that most of the genes of *E.coli S9922* were horizontally transferred.

The *E.coli* phylogenetic group is divided into 6 groups: A, B1, B2, C, D, E and F, the commensal *E.coli* and animal-derived ExEPC are fell in phylogenetic groups A and B1. But ExPEC, which is highly pathogenic to humans, mostly belongs to the phylogenetic group B2, followed by the group D. In this study, *E.coli S9922* belongs to phylogenetic group A. Guo *et al*. discovered that the predominant phylogenetic groups of ExPEC infection in cattle and sheep in parts of Xinjiang were A and B1 from 2015 to 2019[14], the result was consistent with the results of this study. In addition, Porcine ExPEC were mostly fell in phylogenetic groups A, B1 and D [15]. The ESBLs-producing *E.coli* isolated from chicken and pork were mainly group A by Zeng *et al*. [16]. Guo *et al*. found that the main phylogenetic groups of *E.coli* in the feces of sick pigs in Guangdong were A1 and B [17]. It can be seen that Phylogenetic groups may have different distributions in different animals, different regions, and even different organs of the same animal.

Population evolution analysis

The core genes obtained by cluster analysis of *E.coli S9922* homologous genes were classified, there were 2570 genes in COG comparison, 2780 genes in KEGG comparison, and 2188 genes in GO comparison. The specific genes obtained by homologous cluster analysis were classified, there were 204 genes in COG comparison, 550 genes in KEGG comparison, and 239 genes in GO comparison. It was found that the functions of proteins, encoded by core gene clusters, were closely related to the most basic growth-related metabolic processes of cells, reflecting the importance of the presence of core genes for cell survival. The specific gene of *E.coli S9922* is mostly like a prophage genomic sequence that carries a transposon element. Transposon is a removable element in the DNA sequence, which, through the processes of cutting, integration, and a series of "jumps", can move from one position within a genome to another location. Statistical analysis found that *E.coli 90-9272* and *E.coli S9922* were two strains that experienced deletion and insertion of sequences, just prior to the predicted phage-like genomic sequence area. We suspected that they were as phage large mobile components through genetic level transfer, may contain some new pathogenic factors. In addition, the functions of proteins encoded by specific genes were mostly related to specific site recombination, repair, energy generation, and transformation. Some genes are also related to signal recognition, defence mechanisms, and outer membrane, pili, and flagellin protein production, which may play a role in the pathogenic mechanism of *E. coli S9922*.

Conclusion

In this study the whole genome of *E.coli S9922* was sequenced, which enriched the data of bovine ExPEC and provided the basis for the related research. Comparative genomics analysis showed that *E.coli S9922* had a close evolutionary relationship with 90-9272, but far from that of UTI89. In short, the first completed genomic sequence for bovine ExPEC and the genomic differences identified by comparative analyses provide a baseline understanding of bovine ExPEC genetics and lay the foundation for their further study.

Methods

Strains, culture conditions and DNA isolation

Strain *E.coli S9922* was isolated from the brain effusion of calves that had died of meningitis in a dairy farm in Xinjiang province and was identified as extraintestinal pathogenic Escherichia coli (ExPEC) by biology. The strain of *E.coli S9922* was inoculated on solid Luria-Bertani (LB) agar at 37°C for overnight. Next, a single colony was selected and cultured into LB liquid medium at 37°C for 12 h under shaking conditions. Finally, the DNA was extracted using TIANamp Bacteria DNA Kit (TIANGEN Biotech CO., Ltd. Beijing, China) according to the manufacturer's instructions.

Genomic sequencing and assembly

The genomic DNA of the isolate was sequenced at the Novo Zhiyuan Technology Co., Ltd. Beijing using PacBio platform. Qualified DNA samples were broken with Covaris g-TUBE into the desired fragment size to construct the library. After DNA damage repair and end repair, the hairpin connector was ligated to both ends of the DNA fragment using DNA adhesive enzymes. The DNA fragment was purified with AMpure PB magnetic beads to construct the SMRT Bell library. The constructed library was quantified by Qubit concentration, and the size of the inserted fragment was detected using Agilent 2100, and then sequenced using the PacBio platform. By controlling the quality of the original subordinate data and removing the low-quality sequences, the Cleandata was available for analysis. Cleandata was statistically analysed to obtain information, such as data volume, reads length, and quality value distribution. SMRT Link v.5.1.0 software [18-19] was used to assemble the genome of reads, and preliminary assembly results reflecting the basic situation of the *E.coli S9922* genome were obtained.

Genome component analysis

The software GeneMarkS (Version: 4.17) was used to predict the coding genes of the *E.coli* S9922 genome. tRNA and rRNA in genomes were predicted with tRNAscan-SE software (Version: 1.3.1), rRNAmmer software (Version: 1.2) , respectively. sRNA was first compared and annotated using the Rfam database , and then CM search program (Version: 1.1rc4) was used to determine the final sRNA. PhiSpy software (Version: 2.3) was used to predict prophages. IslandPath-DIOMB software (Version: 0.2) was used to predict the gene island (GI) by detecting the phylogenetic bias and mobility genes (such as transposase or integrase) in the sequence to determine the island and the potential horizontal gene transfer. CRISPRdigger (Version: 1.0) was used to predict CRISPR of the sample genome. All predicted open reading frames (ORFs) were automatically forced into similarity searches against GO (version:1.419), KEGG (https://www.genome.jp/kegg/), COG (http://www.ncbi.nlm.nih.gov/COG/), NR (version: 20121005), TCDB (http://www.tcdb.org/), Pfam (http://pfam.xfam.org/), Swiss-Prot (http://www.ebi.ac.uk/uniprot/) and CAZy databases (evaluate ≤ 1e-5). The alignment results were filtered, and the comparison results with the highest score (default identity > = 40%, coverage > = 40%) were selected for annotation.

Virulence and pathogenicity analysis

SignalP (Version 4.1) and TMHMM (Version 2.0c) were used to detect whether signal peptides and transmembrane structures were contained, and to comprehensively predict whether the protein sequences were for secretory proteins. For the TNSS system, proteins related to the secretion system were extracted from the annotation results of the protein sequence function database for annotation and the effective T3 software (version 1.0.1) was used to predict the T3SS effector protein. The genes of *E.coli S9922* were annotated using the Pathogen and Host Interaction Database (PHI). By using Diamond software, the amino acid sequence of *E.coli S9922* was compared with the VFDB database, and the annotation results were obtained by combining the genes of *E.coli S9922* and the corresponding virulence factor function annotation information. Using Diamond software, the amino acid sequence of *E.coli S9922* was compared with the database of ARDB, and the genes of *E.coli S9922* and its corresponding annotation information of drug resistance function were combined to get the annotation results.

Comparative genomics analysis

General characteristics of *E.coli* genome and 16 reference *E.coli* genomes were compared: PCN061 [GenBank:NZ_CP006636] [1],ECC-1470 [GenBank:NZ_CP010344] [20], CE10 [GenBank:NC_017646] [21], IHE3034 [GenBank: NC_017628] [22], NMEC_O18[GenBank: NZ_CP007275] [23], CFT073 [GenBank:NC_004431] [24], UTI89 [GenBank: NC_007946] [25], APEC_O78 [GenBank: NC_020163] [26], APECO1 [GenBank: NC_008563] [27], 042 [GenBank: NC_017626] [28], CB9615 [GenBank: NC_013941] [29], 92-9272 [GenBank: NZ_CP024239] [30], 12009 [GenBank: NC_013353] [31], APEC_O2-211 [GenBank: NZ_CP006834] [32], UMN026 [GenBank: NZ_CP006834] [33] and NRG_857C [GenBank: NC_017634] [34]. MUMmer software (Version 3.23) and LASTZ (Version 1.03.54) were used to compare the whole genomes of *E. coli S9922* and twelve other representative *E.coli* strains to determine the large-scale collinearity relationship and search for translocation/trans, inversion/inv, and trans+ inv areas.

Population evolution analysis

Cd-hit software was used to cluster the protein sequences of multiple samples to be analysed. Screening parameters for identity and comparison length were set, and the clustering data of all protein sequences were obtained according to the software analysis results.

BLAST (Version 2.2.26) was used to compare the protein sequences of multiple target genomes, in pairs, and the untrusted results were filtered. Further, the redundancy was removed using Solar, and the proteins were clustered by using Hcluster-SG according to the similarity ratio. Thus, the results for gene family clustering were obtained.

The single-copy core gene identified based on core-pan analysis was used to align multiple protein sequences using MUSCLE software, and the evolutionary tree was constructed with NJ (neighbor-joining) using TreeBeST software.

Abbreviations

APEC: avian pathogenic E. coli

ARDB: Antibiotic Resistance Genes Database

COG: Cluster of Orthologous Group

E.coli: Escherichia coli

ExPEC Extraintestinal pathogenic Escherichia coli

GIs: gene islands

GO: Gene Ontology;

IPEC: intestinal pathogenic E. coli

KEGG: Kyotoencyclopedia of Genes and Genomes;

NMEC: neonatal meningitis pathogenic E. coli

NR:Non-Redundant Protein

PHI:Pathogen and Host Interaction

SignalP: Signal peptide

TMHMM: Trans Membrane prediction using Hidden Markov Models

TCDB: The Transporter Classification Database

UPEC: urethra pathogenic E. coli

VFDB: The virulence factor database

Declarations

Acknowledgements

Not applicable.

Nucleotide sequence accession number

The draft genome sequences have been deposited in the NCBI Sequence Read Archive under the accession number: SRA8740874.

Availability of data and material

All relevant data in this study are available from the corresponding author upon reasonable request.

Authors' contributions

PY W, JJ J designed the study, JJ R, BB L and M X analyzed the sequence, JJ R, BB L, QQ and XY W prepared the manuscription. All authors read and approved the final manuscript.

Funding

This work was supported by a grant from the NSFC-XinJiang Joint Fund (U1803109, NSFC, China).

The funding body did not have any role in the design of the study and collection, analysis, or interpretation of data or in the writing of the manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no conflicts of interest concerning this work.

References

1. Canying L, Huajun Z, Minjun Y, *et al*. Genome analysis and in vivo virulence of porcine extraintestinal pathogenic Escherichia coli strain PCN033. *BMC Genomics*.2015; 16:717.
2. Kaper J B . Pathogenic Escherichia coli. *Int Med J Med Mcrobiol*. 2005; 295(6-7):355-356.
3. Yindi, X. Research progress on parenteral pathogenic Escherichia coli. *Animal Husb Vet Sci Technol Info*. 2011; 7: 21-23.
4. Stromberg Z R, Johnson J R, Fairbrother J M, *et al*. Evaluation of Escherichia coli isolates from healthy chickens to determine their potential risk to poultry and human health. *Plos One*. 2017;12 (7):e0180599.
5. MillerJ R□Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. *Genomics*.2010;95 (6): 315-327.
6. Ochman H, Lawrence J G, Groisman E A. Lateral gene transfer and the nature of bacterial innovation. *Nature*. 2000; 405:299-304.
7. Tseng T T, Tyler B M, Setubal J C. Protein secretion systems in bacterial-host associations, and their description in the Gene Ontology. *BMC Microbiol*.2009;9 Suppl 1:S2.
8. Yang Y, Wu Q. Bioinformatics analysis of secreted proteins of 16M genome of Brucella serrata. *Journal of Animal Husb Vet Med*. 2009; 40: 1059-1062.
9. Cantas L, Shah S, Cavaco L M, et al. A brief multi-disciplinary review on antimicrobial resistance in medicine and its linkage to the global environmental microbiota. *Front Microbiol*. 2013; 4:96.
10. Thaller M C, Migliore L, Marquez C, *et al*. Tracking acquired antibiotic resistance in commensal bacteria of Galápagos Land Iguanas: no man, no resistance. *Plos One*. 2013; 5.
11. Smith J L, Fratamico P M, Gunther N W. Extraintestinal pathogenic *Escherichia coli*. *Foodborne Pathog Dis.*2007; 4 (2): 134-163.
12. Chen T, Tang Xi-biao, Zhang Xuan, *et al*. Serotypes and virulence genes of extraintestinal pathogenic Escherichia coli isolatesfrom diseased pigs in China. Vet J, 2012; 192 (3): 483-488.
13. Qiangqiang G, Mengli H, xia Z *et al*. Isolation, identification and drug resistance of pathogenic Escherichia coli strains from diseased sheep lung. *Chin J Anim Infect Dis Sci*. 2020; 28 (3): 29-34.
14. Xiaoxiao G, Qin W, Qiaoxiaoci T *et al*. Isolation, identification and detection of drug resistance and drug resistance gene of pulmonary pathogenic Escherichia coli from calves. *Chin J Anim Husb &Vet Med*. 2020; 47(1):240-248.
15. Yi D, Xibiao T, Ping L, et al. Clonal analysis and virulent traits of pathogenic extraintestinal Escherichia coli isolates from swine in China. *BMC Vet Res*. 2012; 8:140.
16. Li Z. Molecular Characteristics of plasmid-mediated ESBLs genes in Escherichia coli Isolated from Animals, Retail Meat and Humans. Master's thesis, South China Agricultural University. China: Sichuan; 2016.

17. Antimicrobial resistance and phylogenetic analysis of *E.coli* isolated from swine in Guang dong province. *Chin J Anim Husb &Vet Med*. 2014;41 (01): 182-186.
18. Simon A, Adam A, Vermeesch J R, *et al*. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res*. 2018;46(5): 2159–2168.
19. Reiner J, Pisani , Qiao W, *et al.* Cytogenomic identification and long-read single molecule real-time (SMRT) sequencing of a bardet–biedl syndrome 9 (BBS9) deletion. *Npj Genom Med*.2018;3(1):3.
20. Leimbach A, Poehlein A, Witten A, et al. Complete genome sequences of Escherichia coli strains 1303 and ECC-1470 isolated from bovine mastitis. *Genome Announc*. 2015;3(2);e00182-15.
21. Shuting L, Xiaobing Z, Yafang Z, *et al.* Complete genome sequence of the neonatal-meningitis-associated *Escherichia coli* Strain CE10. *J Bacteriol*. 2011;193(24):7005.
22. Moriel D G, Bertoldi I, Spagnuolo A, et al. Identification of protective and broadly conserved vaccine antigens from the genome of extraintestinal pathogenic *Escherichia coli*. *Proc Natl Acad Sci US A*, 2010;107(20):9072-9077.
23. Nicholson B A, Wannemuehler Y M, Logue C M, *et al*. Complete genome sequence of the neonatal meningitis-causing *Escherichia coli* strain NMEC O18. *Genome Announc*.2016;4(6): e01239-16.
24. Welch R A, Burland V, Plunkett G, *et al*. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc Natl Acad Sci USA*, 2002;99(26):17020-17024.
25. Chen S, Hung C, Xu J, *et al*. Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: A comparative genomics approach. P Natl Acad SCI USA. 2006;103(15): 5977-5982.
26. Mangiamele P, Nicholson B, *et al*. Complete genome sequence of the avian pathogenic *Escherichia coli* Strain APEC O78. *Genome Announc*, 2013; 1 (2): e00026-13.
27. Johnson T J, Kariyawasam S, Wannemuehler Y, *et al.* The genome sequence of avian pathogenic *Escherichia coli* strain O1:K1:H7 shares strong similarities with human extraintestinal pathogenic *E. coli* genomes. *J Bacteriol*. 2007; 189 (12):4554-4554.
28. Chaudhuri R R, Mohammed S, Hobman J L, *et al*. Complete genome sequence and comparative metabolic Profiling of the prototypical enteroaggregative *Escherichia coli* strain 042. *Plos One*, 2010;5 (1):e8801.
29. Zhemin Z, Xiaomin L, Bin L *et al.* Derivation of *Escherichia coli* O157:H7 from its O55:H7 precursor. *Plos One*. 2010; 5(1):e8700.
30. Smith P, Lindsey R L, Rowe L A, *et al*. High-quality whole-genome sequences for 21 enterotoxigenic Escherichia coli strains generated with PacBio sequencing. *Genome Announc*. 2018; 6(2):e01311-17.
31. Ogura Y, Ooka T, Lguchi A, *et al*. Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic *Escherichia coli*. *P Natl Acad SCI USA*.2009; 106(42):17939-17944.
32. Nielsen D W, Mangiamele A P, Ricker B N, *et al*. Complete genome sequence of avian pathogenic *Escherichia coli* Strain APEC O2-211. *Microbiol Res Announc*. 2018; 7(12):e01046-18.
33. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, *et al.* Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *Plos Genet*. 2009;5: e1000344.
34. Nash J H, Villegas A, Kropinski A M, *et al.* Genome sequence of adherent-invasive *Escherichia coli* and comparative genomic analysis with other *E. coli* pathotypes. *BMC Genomics*. 2010;11 (1):667.

Tables

Table 1 Overall genome features of *E. coli S9922*

Characteristic	*E. coli S9922*
Genome size (bp)	5,269,374
G+C content(%)	50.82
Contings	4
Average gene size(bp)	903
rRNA	87
tRNA	22
sRNA	63
Genomic island(GIs)	23
prophage	10
CRISPR	36

Table 2 General features of *E.coli* S9922 genome and other *E.coli* strains

Strain	Serotype	Patholtype	Phylogroup	tRNA	rRNA	Total genes	Chromosome Size(kb)	Plasmid Size(kb)	G+C(%)
S9922		ExPEC	A	87	22	5,055	4,961	110,74,78	50.8
PCN061	O9	ExPEC	A	86	22	4,808	4,604	2,6,6,35, 104,146	50.8
ECC-1470		ExPEC	B1	90	22	4,699	4,804	100	
APEC_O78	O78	APEC	B2	88	19	4,652	4,798	218,113	50.7
APEC_ O1	O1:K1:H7	APEC	B2	95	22	5,392	5,082	66	50.6
042	O44:H18	EAEC	D	93	22	5,191	5,242	113	50.6
CB9615	O55:H7	EPEC	E	101	7	5,328	5,386	66	50.5
92-9272	015:H11	ETEC		89	22	5,097	4,907	274	
12009	O103:H2	EHEC	B2	98	22	5,472	5,449	72	
APEC_O2-211	O2	EHEC		86	22	5,164	5,114	198,4,2	50.6
CE10	O7:K1:NM	NMEC	F	92	22	5,239	5,314		50.6
IHE3034	O18:K1:H7	NMEC	B2	96	21	4,969	5,108	0	50.7
NMEC_O18	O18:K1	NMEC	B2	89	22	4,819	5,003	153	50.8
CFT073	O6:K1:H?	UPEC	B2	89	22	5,020	5,231	0	50.5
UTI89		UPEC	B2	88	22	4,998	5,066	114	50.6
UMN026	O7:K1	UPEC	D	87	22	5,174	5,202	122,34	50.7
NRG_857C	O83:H1	AIEC	B2	84	22	4,701	4,748	147	50.7

Table 3: List of statistical results of gene family identification

Species	Total genes	Genes in families	Unclustered genes	Families	Unique families
S9922	5,055	4,659	396	3,260	35
042	5,095	4,947	148	3,633	2
APEC_ O1	4,952	4,913	39	3,607	0
APEC_O2-211	4,936	4,818	118	3,526	3
APEC_O78	4,641	4,568	73	3,391	0
CFT073	4,989	4,905	84	3,522	1
ECC-1470	4,610	4,538	72	3,420	0
IHE3034	4,988	4,955	33	3,633	0
NMEC_O18	4,832	4,807	25	3,525	1
NRG857C	4,425	4,282	143	3,125	1
O103H2str.12009	5,380	5,207	173	3,618	13
90-9272	4,542	4,445	97	3,304	4
O55H7str.CB9615	5,238	5,091	147	3,633	0
O7K1str.CE10	5,066	4,947	119	3,575	11
PCN061	4,343	4,270	73	3,228	1
UMN026	4,825	4,663	162	3,363	2
UTI89	4,895	4,849	46	3,588	0

Note: Total genes: Total number of genes encoded by the species; Genes in families: The number of Genes in a family; Unclustered genes: the number of genes not clustered; Families: The number of genes in each family; Unique Family: The number of unique genetic families (those that contain only genes of the species) that are unique to the species.

Figures

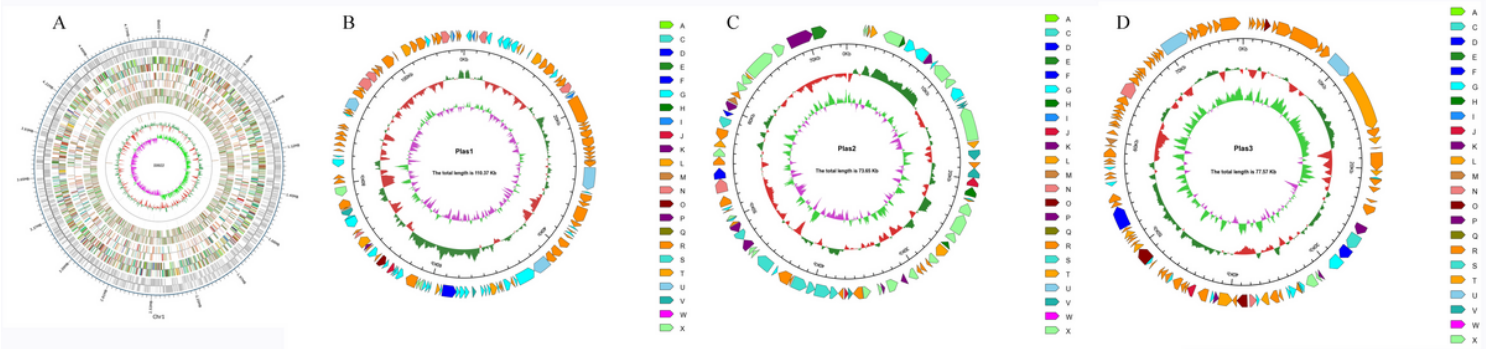

Figure 1

Circular maps of E.coli S9922. Circular maps of E.coli S9922 chromosome (A) Note: The outermost circle is the position coordinates of the genomic sequence. From the outside to the inside, they are the coding gene, the gene function annotation result, the ncRNA, the genomic GC content, and the genomic GC skew value distribution. Circular maps of E.coli S9922-Plas1 (B), E.coli S9922-Plas2 (C) and E.coli S9922-Plas3 (D). Note: From the outside to the inside, the images are COG functional annotation classification genes (arrows indicate positive-chain coding clockwise), genomic sequence position coordinates, genomic GC content, and genomic GC skew value distribution.

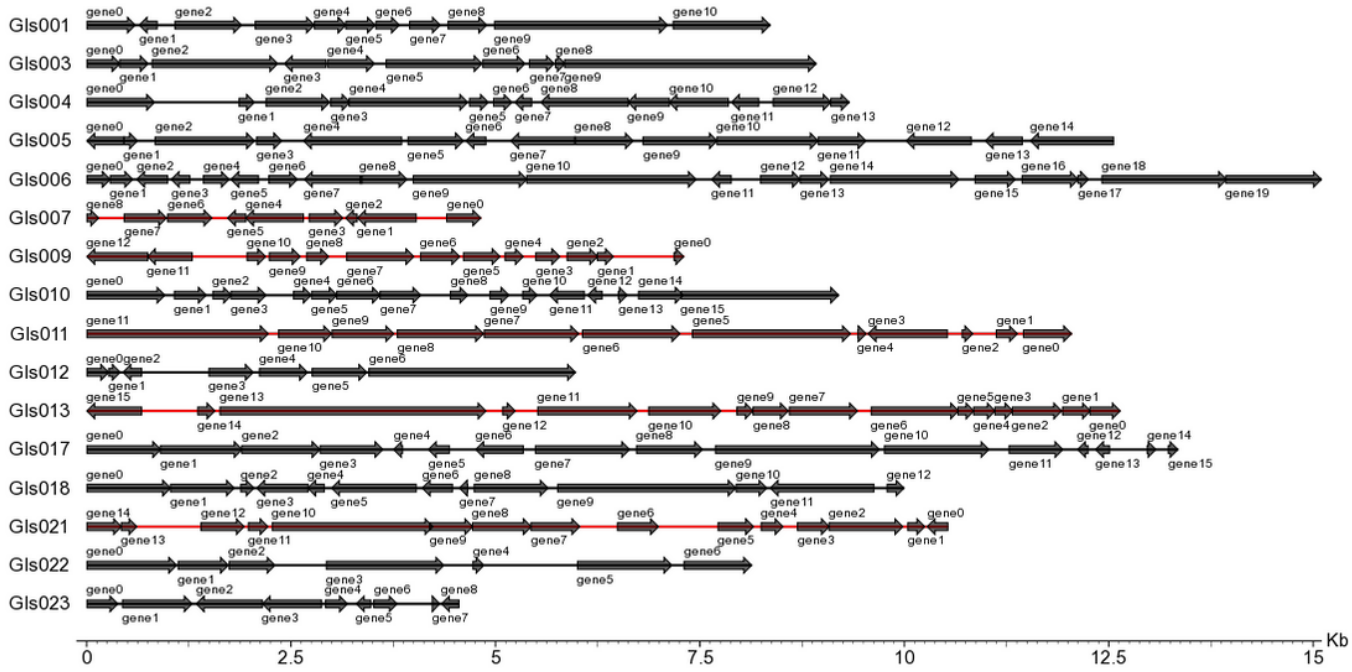

Figure 2

E.coli S9922 Statistical map of gene distribution in Gis

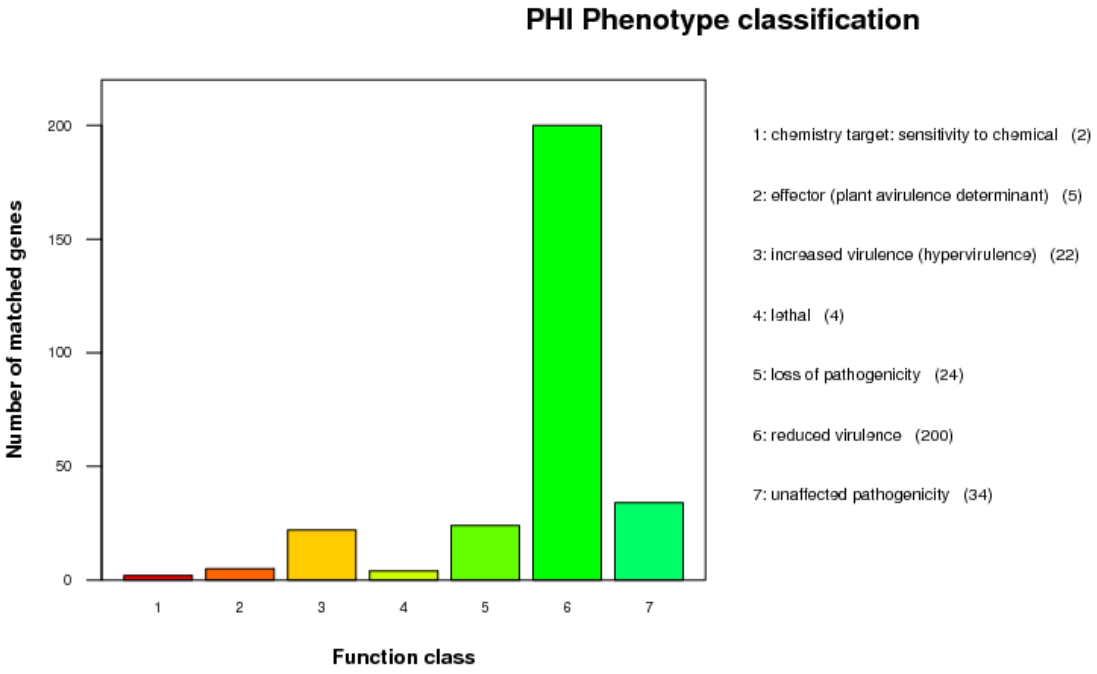

Figure 3

E. coli S9922 PHI phenotype mutation type distribution map

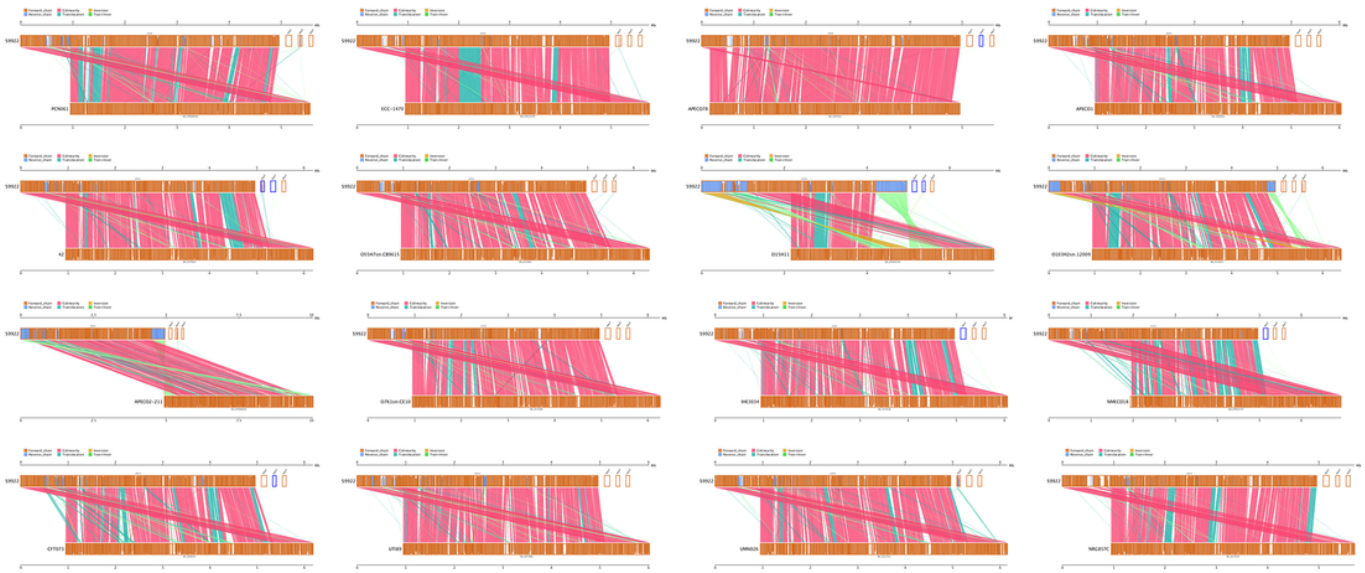

Figure 4

Parallel collinearity of the whole genome of E.coli S9922 and each reference strain Note: The upper axis represents the measured genome and the lower axis represents the reference sequence genome. The yellow box in the upper and lower axes represent the forward chain of the genome, the blue box represents the reverse chain of the genome, the height of the fill colour in the box represents the similarity of the alignment, and the complete fill represents the similarity of 100%. Collinear:

With linear alignment; Translocation: Translocation; -Rufus: Well, Inversion. Tran+Inver: Comparison of translocation and inversion.

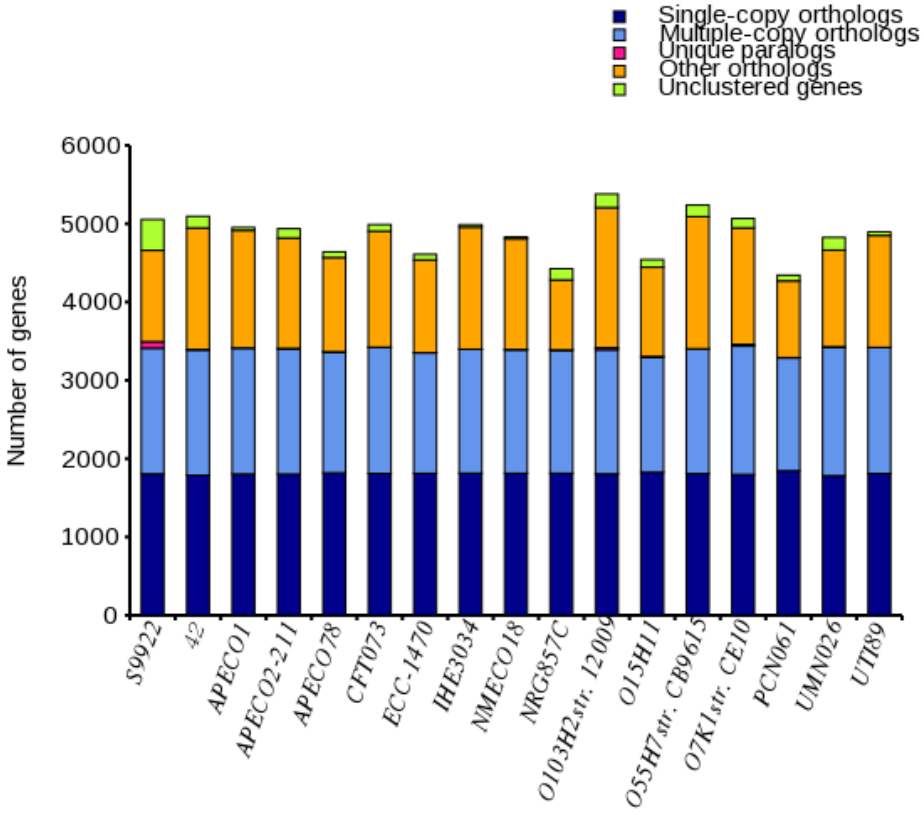

Figure 5

Bar chart for the number of homologous genes Note: Single-copy orthologs: The number of single-copy homologous genes in a family of common genes of species; Polycopy orthologs: The number of polycopy homologous genes in a family of common genes of a species; Unique Paralogs: a gene in a family of species-specific genes; Other orthologs: All the other genes; Unclustered genes: Genes not clustered into any family.

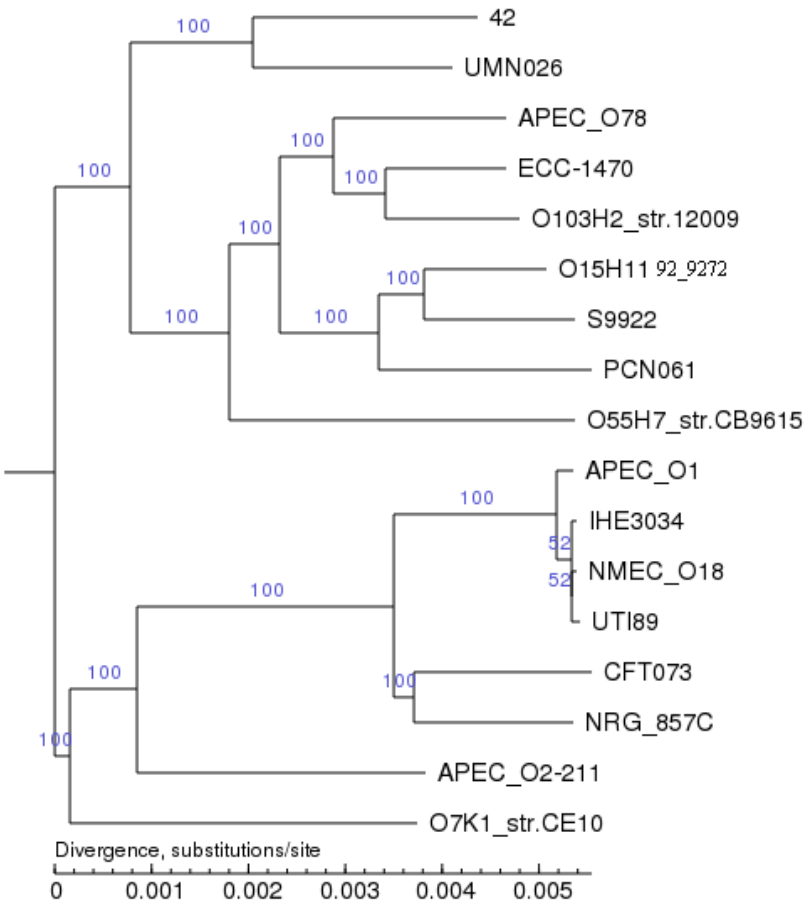

Figure 6

Diagram of evolutionary relationships among species

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.xls