Data in Brief 38 (2021) 107402

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Dataset of single nucleotide polymorphisms and comprehensive proteomic analysis of *Streptococcus equi* subsp. *equi* ATCC 39506

Hayoung Lee[a,b,c], Sung Ho Yun[d], Ju-yong Hyon[a], Sang-Yeop Lee[a,b], Yoon-Sun Yi[d], Chi-Won Choi[e], Sangmi Jun[d], Edmond Changkyun Park[a,b,c], Seung Il Kim[a,b,c,*]

[a] *Research Center for Bioconvergence Analysis, Korea Basic Science Institute (KBSI), Ochang 28119, Republic of Korea*
[b] *Center for Convergent Research of Emerging Virus Infection, Korea Research Institute of Chemical Technology (KRICT), Daejeon 34114, Republic of Korea*
[c] *Department of Bio-Analytical Science, University of Science and Technology, Daejeon 34113, Republic of Korea*
[d] *Center for Research Equipment, Korea Basic Science Institute, Ochang 28119, Republic of Korea*
[e] *KBNP Technology Institute, KBNP, INC., Heungan-daero 415, Dongan-Gu, Anyang, Gyeonggi, Republic of Korea*

ARTICLE INFO

Article history:
Received 31 August 2021
Revised 13 September 2021
Accepted 16 September 2021
Available online 23 September 2021

Keywords:
Comprehensive proteomics
Extracellular vesicle purification
Immunoprecipitation–MS
Label-free proteomics
Shotgun proteomics
Subcellular fractionation
Vaccine candidate

ABSTRACT

Streptococcus equi subspecies *equi* (*S. equi*) is an opportunistic pathogen and a major causative agent of equine strangles, a contagious respiratory infection in horses and other equines. In this study, we provide the dataset associated with our research publication "*Streptococcus equi*-derived extracellular vesicles as a vaccine candidate against *Streptococcus equi* infections" [1]. We describe the genomic differences between *S. equi* 4047 and *S. equi* ATCC 39506 and outline the comprehensive proteome information of various fractions, including the whole cell lysate, membrane proteome, secretory proteome, and extracellular vesicle proteome. In addition, we included a dataset of highly immunoreactive proteins identified through immunoprecipitation. The specifications table provides a detailed summary of the gene annotation and quantitative information obtained for each proteome. The proteomics data were analyzed using shotgun proteomics with LTQ Velos and Q Exactive mass spectrometry in the data-

DOI of original article: 10.1016/j.vetmic.2021.109165
* Corresponding author.
E-mail address: ksi@kbsi.re.kr (S.I. Kim).

https://doi.org/10.1016/j.dib.2021.107402

dependent acquisition mode. We have deposited the acquired data, including the mass spectrometry raw files and exported MASCOT search results, in the PRIDE public repository under the accession numbers PXD025152 and PXD025527.

Specifications Table

Subject	Veterinary Science and Medicine
Specific subject area	Genomics and Proteomics
Type of data	Table
How data were acquired	Instrument: UltiMate 3000 RSLC nano HPLC System (Thermo Fisher Scientific, Waltham, MA, USA) coupled with LTQ Velos and Q Exactive mass spectrometer (Thermo Fisher Scientific) Software: kSNP3, MASCOT (version 2.4, Matrix Science, Boston, MA, USA)
Data format	Raw and analyzed
Parameters for data collection	Reference genome was used for comparative genomic analysis. *S. equi* protein samples were obtained from different fractions (including whole cell lysate [WCL], membrane proteome [MEM], secretory proteome [SEC], and extracellular vesicle proteome [EV]) or were immunoprecipitated with mock and convalescent mouse sera.
Description of data collection	The genome was annotated using the NCBI Prokaryotic Genome Annotation Pipeline. Single nucleotide polymorphisms were identified using the kSNP3 software. The subcellular fractionated and immunoprecipitated proteomes were profiled using LC–MS/MS. All cellular fractions were analyzed in triplicate. Proteins identified more than twice in those replicates were considered positive. The Percolator algorithm was used to recalculate the ion score to reduce the false positives in the IP samples.
Data source location	Institution: Research Center for Bioconvergence Analysis City/Town/Region: Korea Basic Science Institute, Ochang 28119 Country: Republic of Korea Primary data sources: The genomes of *S. equi* 4047 (NCBI reference sequence, NC_012471.1) and *S. equi* ATCC 39506 (GenBank, CP021972.1)
Data accessibility	Name of Repository: PRIDE [2] Data identification number: PXD025152 (Fractionated proteome) Direct URL to data: https://www.ebi.ac.uk/pride/archive/projects/PXD025152 Data identification number: PXD025527 (Immunoprecipitated proteome) Direct URL to data: https://www.ebi.ac.uk/pride/archive/projects/PXD025527
Related research article	Lee. H., and Kim. S. I. *Streptococcus equi*-derived extracellular vesicles as a vaccine candidate against *Streptococcus equi* infections, Vet. Microbiol. 259 (2021) 109165. https://doi.org/10.1016/j.vetmic.2021.109165

Value of the Data

- This proteomics study performed on *S. equi* subsp. *equi* reveals a comprehensive baseline map of the bacterial proteome under different conditions, which is a valuable resource for future studies.
- Protein expression of bacterial components and gene annotation data obtained from various databases can be useful to investigate the bacterial pathogenesis or specific cellular proteome.
- The antigens isolated from mouse serum through immunoprecipitation can be attractive targets and facilitate further discovery of vaccine candidates.
- These datasets would be a valuable baseline for further proteogenomic studies.

1. Data Description

The data presented in this paper are the results of a comparative genomic analysis of *Streptococcus. equi* (*S. equi*) 4047 and ATCC 39506 and proteomic analysis of four different fractions (WCL, MEM, SEC, and EV) and EV samples treated through immunoprecipitation (IP). For comparative genomics analysis, the reference genomes of *S. equi* 4047 (NCBI reference sequence, NC_012471.1) and *S. equi* ATCC 39506 (GenBank, CP021972.1) were used. We identified 246 single nucleotide polymorphism (SNP) sites. Twenty-one of the identified SNP sites caused synonymous or non-synonymous mutations (Table 1). Other SNP sites were involved in the mutation of either phages or pseudogenes. The fractionated proteome dataset obtained from LC–MS/MS analysis included locus tags, UniProt accessions, description of the identified proteins, official gene symbols, enriched cluster information, predicted localization, the results of reverse vaccinology, and normalized exponentially modified protein abundance index (emPAI) values. Each accession of *S. equi* ATCC 39506 was matched with that of *S. equi* 4047 using BLASTp. Differentially expressed proteins (DEPs) underwent Gene Set Enrichment Analysis for prokaryotes (GSEA-Pro) (http://gseapro.molgenrug.nl/). The raw data were submitted to the ProteomeXChange database under the accession numbers PXD025152 and PXD025527.

Table 1
Comparison of coding region mutations between *Streptococcus equi* subsp. *equi* 4047 and *Streptococcus equi* subsp. *equi* ATCC 39506.

Locus_tag	Description	Coding region change(s)	Amino acid change(s)
SE071780_00014[1]	cell division protein FtsH	A->G	H366R
SE071780_00045	putative phosphoribosylformylglycinamidine synthase protein	A->G	-
SE071780_00193	putative autolysin regulatory protein	A->C	P53H
SE071780_00367	type I glyceraldehyde-3-phosphate dehydrogenase	T->G	C65F
SE071780_00369	phosphoglycerate kinase	G->A	K234E
SE071780_00390	putative exported protein	C->T	-
SE071780_00420	endonuclease MutS2	T->C	T453I
SE071780_00530	transcription termination/antitermination protein NusA	A->G	-
SE071780_00836	ArpU family transcriptional regulator	A->C	P53H
SE071780_00878	hypothetical protein	C->T	I66L, V81A, S95P
SE071780_00900	PTS fructose transporter subunit IID	G->A	I147V
SE071780_00908	voltage-gated chloride channel protein	A->C	H91N
SE071780_00925	putative exported protein	G->A	I31V
SE071780_00944	glycogen phosphorylase	T->C	-
SE071780_00966	collagen binding putative ancillary pilus subunit Cne	G->A	T574A
SE071780_00967	fimbrial protein	T->C	-
SE071780_01063	ABC transporter permease	A->C	-
SE071780_01237	triphosphoribosyl-dephospho-CoA synthase CitG	C->A	N271H
SE071780_01600	DNA adenine methylase	G->A	-
SE071780_01920	collagen-like surface-anchored protein SclI	G->C	D332E
SE071780_02280	cell surface protein	A->G	R217K

The genome sequence of *Streptococcus equi* subsp. *equi* 4047 was used as the reference genome.
[1] Proteins identified by proteomic analyses are highlighted in bold.

2. Experimental Design, Materials and Methods

2.1. S. equi culture conditions

S. equi ATCC 39506 was procured from the Korean Culture Center of Microorganisms (Seoul, Republic of Korea). The bacteria were cultivated overnight in Todd–Hewitt broth supplemented with 1% yeast extract at 37.5 °C in an atmosphere with 5% CO_2. Bacteria were further inoculated in fresh broth for 7–8 h until the mid-logarithmic growth phase (optical density at 600 nm = 0.4–0.6) was achieved for protein sample preparation.

2.2. Comparative genomic analysis

SNPs in the genomes of *S. equi* subsp. *equi* ATCC 39506 and *S. equi* subsp. *equi* 4047 were identified using kSNP3 with a k-mer length of 31 [3]. SNPs on the coding sequence were selected, and the amino acid variants were identified using BLASTp for comparison with the protein sequences of the two variants.

2.3. Preparation of fractionated protein samples of S. equi ATCC 39506

The four subcellular fractions (including WCL, MEM, SEC, and EV) from *S. equi* ATCC 39506 cultures were purified as previously described [4], with slight modifications. Briefly, 6 L of the cultured bacteria were harvested through centrifugation at 8000 × *g* for 30 min at 4 °C to remove the cell debris. The supernatant and pellet fractions were retained separately for further processing. For the WCL fraction, the pellet was washed thrice with 20 mM Tris-HCl buffer (pH 8.0) and disrupted using a French pressure cell (SLM Aminco, Urbana, IL, USA) at 138 MPa until a transparent solution was obtained. For MEM fraction, the resulting supernatant was collected through centrifugation at 15000 × *g* for 20 min and precipitated using sodium carbonate precipitation method. The supernatant was mixed in a 10 mM sodium carbonate solution and precipitated through ultracentrifugation at 150000 × *g* for 1 h at 4 °C. The pellet was resuspended in 20 mM Tris-HCl buffer (pH 8.0). For SEC fraction, the supernatant of the harvested bacteria was filtered using a 0.45-μm bottle-top vacuum filter (Corning, NY, USA). The resulting flow-through was dissolved in 80% ammonium sulfate at 4 °C for 2 h, and then precipitated at 5000 × *g* for 20 min at 4 °C. Thereafter, the pellet was resuspended in 20 mM Tris-HCl buffer (pH 8.0). Ammonium sulfate was removed through buffer exchange on Vivaspin 20 (Sartorius, Göttingen, Germany) using over 10 volumes of 20 mM Tris-HCl (pH 8.0). For EV fraction, the supernatant of the harvested bacteria was ultrafiltered using a 0.22 μm vacuum filter and concentrated on a 500 kDa hollow fiber membrane mounted on a QuixStand benchtop system (GE Healthcare, Little Chalfont, UK). The collected supernatant was then precipitated through differential centrifugation methods. The sample was centrifugated at 8000 × *g* for 1 h at 4 °C and the resulting supernatant was further centrifugated at 150000 × *g* for 3 h at 4 °C. The pellet was resuspended in 20 mM Tris-HCl buffer (pH 8.0). The protein concentration of each sample was determined using a Micro BCA Protein Assay Kit (Thermo Fisher Scientific, Waltham, MA, USA) according to the manufacturer's instructions. Each subcellular fraction sample was aliquoted and stored at −80 °C until further use.

2.4. Identification of the proteome of each fraction using LC–MS/MS

In-gel digestion was performed using a previously described method [5] with slight modifications. Briefly, an equal amount of protein of each subcellular fraction was separated through sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE). Each lane was cut into

nine slices based on the molecular size. The sliced gels were destained using a destaining solution comprising 50% acetonitrile and 10 mM ammonium bicarbonate and washed thrice with distilled water. The gels were sequentially treated with a reducing solution (10 mM dithiothreitol and 100 mM ammonium bicarbonate) and an alkylating solution (55 mM iodoacetamide). Trypsin digestion was then performed using 50 mM ammonium bicarbonate at 37 °C for 16 h.

The resulting peptides were dissolved in 0.5% trifluoroacetic acid and further fractionated using a 10 cm × 75 μm ID C18 reverse-phase column (PROXEON) at a flow rate of 300 nL/min. Peptides were eluted using a gradient of 0–65% acetonitrile for 100 min. All MS and MS/MS spectral data were acquired in a data-dependent mode in triplicate for each experimental condition using LTQ Velos (Thermo Fisher Scientific). The full MS spectra were acquired in positive mode within a range of 300–2000 m/z. The top six most intense ions in the acquired scan were selected for followed collision-induced dissociation. The maximum ion injection time was 100 ms for the MS/MS scans. The normalized collision energy was 35 and the isolation window employed was 2 m/z. The dynamic exclusion settings utilized were repeat count 1, exclusion duration 30 s, exclusion list size 50, exclusion mass width low and high 1.5, respectively.

The MS/MS spectra were analyzed through the database search strategy mounted on MASCOT version 2.4 (Matrix Science) for the identification of proteins. The genome of *S. equi* ATCC 39506 (GenBank, CP021972.1, 2208 sequences, date of release: 20-Jun-2017) was used as a reference genome for this analysis. The following parameters were used for protein identification: missed cleavages, 2; peptide mass tolerance, ±0.8 Da; peptide fragment tolerance, ±0.8 Da; peptide charge, 2+, 3+, and 4+; static modifications, carbamidomethyl; and dynamic modification, oxidation (Met). A target-decoy search was used to filter out the low-confidence peptides and proteins with a false discovery rate (FDR) of at least 1%. To reduce the number of false positives in the entire result, proteins identified more than twice in the three replicates were considered positive in the fractionation dataset. For comparative analysis, data visualization and statistical analysis were conducted using the Perseus platform [6]. The log2 emPAI value was used as a normalized value. The missing values after filtration were imputed under a normal distribution [7,8]. The DEPs were selected using the significance analysis of the microarray method [9] (FDR < 0.05) through modifying the gene-specific *t*-statistic calculated by repeated permutation with the background. Heatmap analysis was conducted to visualize the differences in the profiles of each resulting proteome (WCL, SEC, EV, and MEM). Next, the potential vaccine candidates (PVCs) were filtered out using the results of reverse vaccinology and are highlighted in three different colors in the heatmap.

2.5. IP and protein identification using MS/MS

IP experiment was conducted to sort the immunogenic antigens present in the EVs. The EV sample (30 μg) was dissolved in IP lysis buffer (Thermo Fisher Scientific) for 1 h at 4 °C and centrifuged at 4000 × g for 3 min at 4 °C, and the supernatant was collected. To reduce false positive reactions, the supernatant was incubated with anti-IgG conjugated with magnetic beads (Dynabeads; Invitrogen, Carlsbad, CA, USA) for 1 h at 4 °C and then separated. Twenty microliter aliquots of pooled mock and convalescent sera ($n = 10$ each) from the vaccination experiments were incubated with magnetic beads for 30 min at 25 °C as previously described [1]. Each supernatant was incubated with the antibody-bead complex at 4 °C overnight. The beads were vigorously washed with phosphate-buffered saline five times. The bound proteins were eluted with SDS sample buffer and heated at 95 °C for 10 min.

The eluents were cut into five slices based on molecular weight, followed by in-gel digestion, and analyzed through LC–MS/MS using Orbitrap Q Exactive Plus (Thermo Fisher Scientific). The full MS spectra were acquired in positive mode within a range of 150–2000 m/z in 120 min. The top 20 most intense ions in the acquired full mass scan were selected for followed higher-energy collisional dissociation. The maximum ion injection times used were 100 ms for the MS scan and 50 ms for the MS/MS scans. The automatic gain control target settings were 1.0×10^6 for the MS scan mode and 5.0×10^4 for the MS/MS scan mode. The normalized collision energy was

27 and the isolation window employed was 2 m/z. The dynamic exclusion duration was set by 20s. The following parameters were used for protein identification: missed cleavages, 2; peptide mass tolerance, ±10 ppm; peptide fragment tolerance, ±0.8 Da; The other parameters used were as described above. The Peptide score was recalculated using Percolator [10]. Peptide ions with a score above 39, indicating identical or extensive homology matches, were selected. Proteins with more than one independent significant peptide match were selected using the IP dataset.

2.6. Deposition of the MS dataset

The MS proteomics data obtained in this study were deposited in the ProteomeXchange Consortium [11] via the PRIDE partner repository with the identifiers PXD025152 and PXD025527.

Ethics Statements

All animal experiments were reviewed and approved by the Animal Ethics Committee of the Korea Basic Science Institute (approval number KBSI-20-18).

CRediT Author Statement

Hayoung Lee: Conceptualization, data curation, Writing – original draft, Writing – review & editing; **Sung Ho Yun:** Investigation, and validation; **Ju-yong Hyon:** Conceptualization, and validation; **Sang-Yeop Lee:** Conceptualization, Data curation; **Yoon-Sun Yi:** Investigation; **Chi-Won Choi:** Investigation; **Sangmi Jun:** Validation; **Edmond Changkyun Park:** Resources, Writing – review & editing; **Seung Il Kim:** Supervision, Writing –review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the grants received from the Korea Basic Science Institute research program (grant number K39402) and a National Council of Science & Technology (NST) grant from the Korean government (MSIP) (grant number CRC-16- 01-KRICT). The funding sources were not involved in any which way in the research project or in the preparation and submission of this article for publication.

References

[1] H. Lee, S.H. Yun, J. Hyon, S.Y. Lee, Y.S. Yi, C.W. Choi, S. Jun, E.C. Park, S.I Kim, *Streptococcus equi*-derived extracellular vesicles as a vaccine candidate against *Streptococcus equi* infection, Vet. Microbiol. 259 (2021) 109165, doi:10.1016/j.vetmic.2021.109165.

[2] Y. Perez-Riverol, A. Csordas, J. Bai, M. Bernal-Llinares, S. Hewapathirana, D.J. Kundu, A. Inuganti, J. Griss, G. Mayer, M. Eisenacher, E. Pérez, J. Uszkoreit, J. Pfeuffer, T. Sachsenberg, Ş. Yilmaz, S. Tiwary, J. Cox, E. Audain, M. Walzer, A.F. Jarnuczak, T. Ternent, A. Brazma, J.A. Vizcaíno, The PRIDE database and related tools and resources in 2019: Improving support for quantification data, Nucleic Acids Res. 47 (2019) D442–D450, doi:10.1093/nar/gky1106.

[3] S.N. Gardner, T. Slezak, B.G. Hall, kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome, Bioinformatics 31 (2015) 2877–2878, doi:10.1093/bioinformatics/btv271.

[4] C. Choi, E.C. Park, S.H. Yun, S.Y. Lee, Y.G. Lee, Y. Hong, K.R. Park, S.H. Kim, G.H. Kim, S. Il Kim, Proteomic Characterization of the Outer Membrane Vesicle of *Pseudomonas putida* KT2440, J. Proteome Res. 13 (2014) 4298–4309, doi:10.1021/pr500411d.
[5] S.-Y. Lee, S.H. Yun, Y.G. Lee, C.W. Choi, S.H. Leem, E.C. Park, G.H. Kim, J.C. Lee, S. Il Kim, Proteogenomic characterization of antimicrobial resistance in extensively drug-resistant *Acinetobacter baumannii* DU202, J. Antimicrob. Chemother. 69 (2014) 1483–1491, doi:10.1093/jac/dku008.
[6] S. Tyanova, T. Temu, P. Sinitcyn, A. Carlson, M.Y. Hein, T. Geiger, M. Mann, J. Cox, The Perseus computational platform for comprehensive analysis of (prote)omics data, Nat. Methods. (2016), doi:10.1038/nmeth.3901.
[7] G.I. Bertin, A. Sabbagh, F. Guillonneau, S. Jafari-Guemouri, S. Ezinmegnon, C. Federici, B. Hounkpatin, N. Fievet, P. Deloron, Differential protein expression profiles between *Plasmodium falciparum* parasites isolated from subjects presenting with pregnancy-associated malaria and uncomplicated malaria in Benin, J. Infect. Dis. 208 (2013) 1987–1997, doi:10.1093/infdis/jit377.
[8] T. Välikangas, T. Suomi, L.L. Elo, A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation, Brief. Bioinform. 19 (2017) 1344–1355, doi:10.1093/bib/bbx054.
[9] V.G. Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 5116–5121, doi:10.1073/pnas.091062498.
[10] M. Brosch, L. Yu, T. Hubbard, J. Choudhary, Accurate and Sensitive Peptide Identification with Mascot Percolator, J. Proteome Res. 8 (2009) 3176–3181, doi:10.1021/pr800982s.
[11] E.W. Deutsch, N. Bandeira, V. Sharma, Y. Perez-Riverol, J.J. Carver, D.J. Kundu, D. García-Seisdedos, A.F. Jarnuczak, S. Hewapathirana, B.S. Pullman, J. Wertz, Z. Sun, S. Kawano, S. Okuda, Y. Watanabe, H. Hermjakob, B. Maclean, M.J. Maccoss, Y. Zhu, Y. Ishihama, J.A. Vizcaíno, The ProteomeXchange consortium in 2020: Enabling “big data” approaches in proteomics, Nucleic Acids Res. 48 (2020) D1145–D1152, doi:10.1093/nar/gkz984.