DATABASE
The Journal of Biological Databases and Curation

Database, 2021, 1–9
doi:10.1093/database/baab029
Original article

Original article

Application of beta and gamma carbonic anhydrase sequences as tools for identification of bacterial contamination in the whole genome sequence of inbred Wuzhishan minipig (*Sus scrofa*) annotated in databases

Reza Zolfaghari Emameh[1,*], Seyed Nezamedin Hosseini[2] and Seppo Parkkila[3,4]

[1]Department of Energy and Environmental Biotechnology, National Institute of Genetic Engineering and Biotechnology (NIGEB), 14965/161, Tehran, Iran, [2]Department of Recombinant Hepatitis B Vaccine, Production and Research Complex, Pasteur Institute of Iran, Tehran, Iran, [3]Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland and [4]Fimlab Ltd, Tampere University Hospital, Tampere, Finland

*Corresponding author: Email: zolfaghari@nigeb.ac.ir

Citation details: Zolfaghari Emameh, R., Hosseini, S.N., Parkkila, S. *et al.* Application of beta and gamma carbonic anhydrase sequences as tools for identification of bacterial contamination in the whole genome sequence of inbred Wuzhishan minipig (*Sus scrofa*) annotated in databases. *Database* (2021) Vol. 2021: article ID baab029; doi:10.1093/database/baab029

Received 16 March 2021; Revised 19 April 2021

Abstract

Sus scrofa or pig was domesticated thousands of years ago. Through various indigenous breeds, different phenotypes were produced such as Chinese inbred miniature minipig or Wuzhishan pig (WZSP), which is broadly used in the life and medical sciences. The whole genome of WZSP was sequenced in 2012. Through a bioinformatics study of pig carbonic anhydrase (CA) sequences, we detected some β- and γ-class CAs among the WZSP CAs annotated in databases, while β- or γ-CAs had not previously been described in vertebrates. This finding urged us to analyze the quality of whole genome sequence of WZSP for the possible bacterial contamination. In this study, we used bioinformatics methods and web tools such as UniProt, European Bioinformatics Institute, National Center for Biotechnology Information, Ensembl Genome Browser, Ensembl Bacteria, RSCB PDB and *Pseudomonas* Genome Database. Our analysis defined that pig has 12 classical α-CAs and 3 CA-related proteins. Meanwhile, it was approved that the detected CAs in WZSP are categorized in the β- and γ-CA families, which belong to *Pseudomonas* spp. and *Acinetobacter* spp. The protein structure study revealed that the identified β-CA sequence from WZSP belongs to *Pseudomonas aeruginosa* with PDB ID: 5JJ8, and the identified γ-CA sequence from WZSP belongs to *P. aeruginosa* with PDB ID: 3PMO.

Bioinformatics and computational methods accompanied with bacterial-specific markers, such as 16S rRNA and β- and γ-class CA sequences, can be used to identify bacterial contamination in mammalian DNA samples.

Introduction

Pigs (*Sus scrofa*) were domesticated in multiple geographic regions of Asia and Europe through artificial and natural selections about 10 000 years ago. Especially in China as one of the main centers, the domestication created a number of indigenous breeds with various phenotypes including Plateau, Lower Yangtze River Basin, Southwest and North China types (1–3). The whole genome sequences (WGS) of pig models and minipig varieties are important in biomedical studies, such as generation of porcine-induced pluripotent stem cells for the treatment of human diseases including diabetes and cancer as well as ophthalmic, neurodegenerative and cardiovascular diseases (4, 5).

Wuzhishan pig (WZSP) is a Chinese inbred miniature minipig, which is characterized by its small size, approximately weight of 30 kg, homozygosis, genetic stability and good predictability in *in vivo* studies (6). WZSP was developed in the Institute of Animal Science of the Chinese Academy of Agriculture Science in 1987. Fang *et al.* performed the WGS of WZSP in 2012, which defined a high-level derivation of transposons from transfer RNA with 2.2 million copies (12.4% of the genome) (7). In addition, many human gene and effective drug targets have been identified in the genome of WZSP. The WGS of WZSP, completed by the researchers from Beijing Genomics Institute, provided pivotal data for the use of this minipig model in biological, medical and veterinary medicine studies.

The genome of WZSP contains porcine endogenous retroviruses (PERVs), which can be transmitted in the germ lines and infect human cells, leading to severe combined immunodeficiency (8). Therefore, PERVs are considered a great potential risk of xenotransplantation of organs from transgenic pigs like WZSP to human.

Carbonic anhydrases (CAs) are ubiquitous enzymes with metal cofactors such as zinc, iron, cobalt or cadmium in the enzyme active sites catalyzing the hydration of CO_2 to HCO_3^- and H^+ for pH homeostasis and playing the crucial roles in many biochemical pathways and physiological functions (9, 10). CAs are classified into eight evolutionarily distinct families, including α, β, γ, δ, ζ, η, θ and ι (11–14). α-CAs are present in many prokaryotes and eukaryotes (15, 16). There are 13 α-CA isozymes in mammals, of which 12 are present in humans, including CA I–IV, CA VA and VB, CA VI, CA VII, CA IX and CA XII–XIV. CA XV can be found in several vertebrates with the exception of at least chimpanzee and human (17). In addition, the presence of three acatalytic CA-related proteins (CARPs), including CARP VIII, CARP X and CARP XI, has been reported, and these highly conserved proteins seem to play critical biological roles (18–22). Although β- and γ-CAs have been reported in several prokaryotes and eukaryotes, there is no report showing the presence of a β- or γ-CA in vertebrates (23, 24).

Databases such as Ensembl Genome Browser contain huge data resources of vertebrate genomes to support the related studies in various fields, such as evolutionary and computational biology, associated with the WGS, gene expression studies and encoded protein analyses in vertebrates (25). Due to the bacterial contamination of eukaryotic nucleic acid samples with environmental microbiome and normal flora of the eukaryotic hosts, some contaminant gene and protein sequences from prokaryotes have been erroneously annotated for eukaryotes in databases (26).

In this study, we performed a quality control analysis of the WGS results of WZSP annotated in databases using *β- and γ-CA* gene sequences as markers through bioinformatics and data mining approaches.

Methods

Identification of CAs from *S. scrofa*

To identify genomics and proteomics information of the CA isozymes from *S. scrofa*, the National Center for Biotechnology Information (NCBI) database (https://www.ncbi.nlm.nih.gov/) (27) was used to define the chromosome location and exon counts of the corresponding genes. In addition, data from the UniProt database (https://www.uniprot.org/) (28) were used to define the subcellular localization of CA isozymes from *S. scrofa*.

Analysis of β- and γ-CA sequences

In this analysis, β-CA protein sequence from *Acetobacter aceti* (UniProt ID: A0A1U9KGA1) and γ-CA protein sequence from *Shigella flexneri* (UniProt ID: P0A9X0) were used as the query sequences. Basic Local Alignment Search Tool (BLAST) analysis was performed on both β- and γ-CA query sequences using BLAST algorithm of Ensembl Genome Browser (https://asia.ensembl.org/index.html) (25). To find similar sequences in the BLAST analysis, Pig-Wuzhishan (assembly: minipig_v1.0; accession: GCA_002844635.1; genebuild released: September 2019) was selected by species selector section, and

Table 1. α-CAs from *S. scrofa*

α-CAs	UniProt ID	NCBI ID	Gene location	Exon count	Subcellular localization
CA I	A0A287AI92	XP_001924218.1	Chromosome 4	7	Cytoplasmic
CA II	A0A287B6M0	XP_001927840.1	Chromosome 4	7	Cytoplasmic
CA III	A0A4X1UEH4	NP_001008688.1	Chromosome 4	7	Cytoplasmic
CA IV	F1S1C3	NP_001230849.1	Chromosome 12	8	Membrane-bound
CA VA	A0A5G2QRM5	XP_020949335.1	Chromosome 6	13	Mitochondrial
CA VB	F1SQS9	XP_005673507.1	Chromosome X	9	Mitochondrial
CA VI	F1RIH8	NP_001137588.1	Chromosome 6	8	Secretory
CA VII	A0A286ZZG4	XP_020949678.1	Chromosome 6	8	Cytoplasmic
CA IX	A0A5G2QGY0	XP_001925555.2	Chromosome 1	12	Transmembrane
CA XII	F1S092	XP_020949824.1	Chromosome 1	11	Transmembrane
CA XIII	A0A287ASJ5	XP_001924497.3	Chromosome 4	9	Cytoplasmicc
CA XIV	A0A287B0I5	XP_020945576.1	Chromosome 4	9	Transmembrane
CARP VIII	A0A287BFY8	XP_020944998.1	Chromosome 4	10	Cytoplasmic
CARP X	A0A480LJN7	XP_020922898.1	Chromosome 12	11	Secretory
CARP XI	A0A4X1VZX6	XP_005664726.1	Chromosome 6	9	Secretory

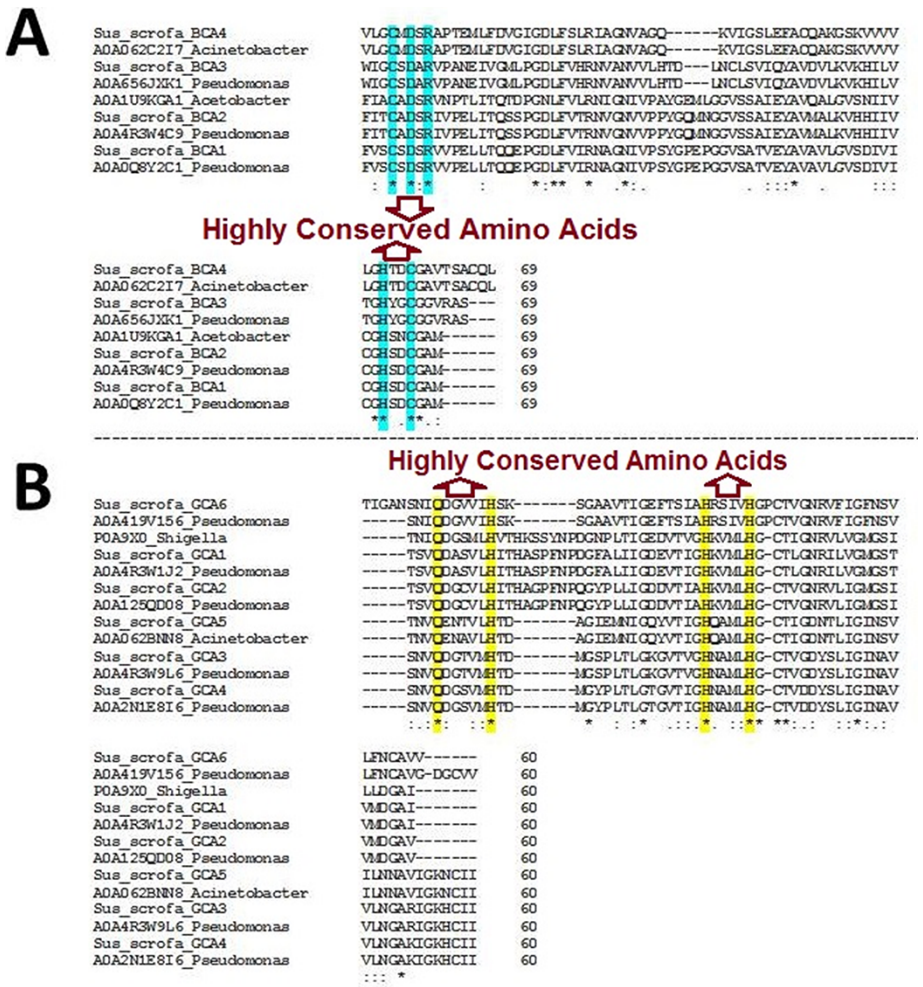

Figure 1. Multiple sequence alignment (MSA) of β- and γ-CA sequences. (A) MSA of β-CA sequences shows highly conserved amino acids in cyan color; (B) MSA of γ-CA sequences shows highly conserved amino acids in yellow color.

Table 2. List of β- and γ-CA sequences from WZSP with 100% identity to counterpart sequences from bacteria

				TBLASTN results					
CA family	CA query (UniProt ID)	WZSP CA (Ensmbl genomic location)		Length (amino acids)	ID (%)	Bacteria (UniProt ID)	*E*-value	ID (%)	RSCB PDB 3D model
β-CA	*Acetobacter aceti* (A0A1U9KGA1)	BCA1	AJKK01119664: 532–1149	222	46.40	*Pseudomonas* sp. (A0A0Q8Y2C1)	7e-59	100	5JJ8
		BCA2	KQ002894: 52 809–53 450	203	52.22	*Pseudomonas* sp. (A0A4R3W4C9)	2e-64	100	
		BCA3	AJKK01121845: 27–380	109	36.70	*Pseudomonas syringae* (A0A656JXK1)	5e-12	100	
		BCA4	AJKK01117230: 2023–2607	176	26.14	*Acinetobacter* sp. (A0A062C2I7)	1e-09	100	
γ-CA	*Shigella flexneri* (P0A9X0)	GCA1	KQ002894: 61 481–62 005	175	60.57	*Pseudomonas* sp. (A0A4R3W1J2)	6e-71	100	3PMO
		GCA2	AJKK01118454: 663–1190	176	61.36	*Pseudomonas fluorescens* (A0A125QD08)	1e-71	100	
		GCA3	KQ002836: 4671–5114	155	38.06	*Pseudomonas* sp. (A0A4R3W9L6)	2e-28	100	
		GCA4	AJKK01180312: 124–558	152	37.50	*Pseudomonas fluorescens* (A0A2N1E8I6)	9e-27	100	
		GCA5	AJKK01118286: 1328–1756	150	35.33	*Acinetobacter* sp. (A0A062BNN8)	3e-25	100	
		GCA6	AJKK01161219: 1382–1714	119	34.45	*Pseudomonas synxantha* (A0A419V156)	2e-13	100	

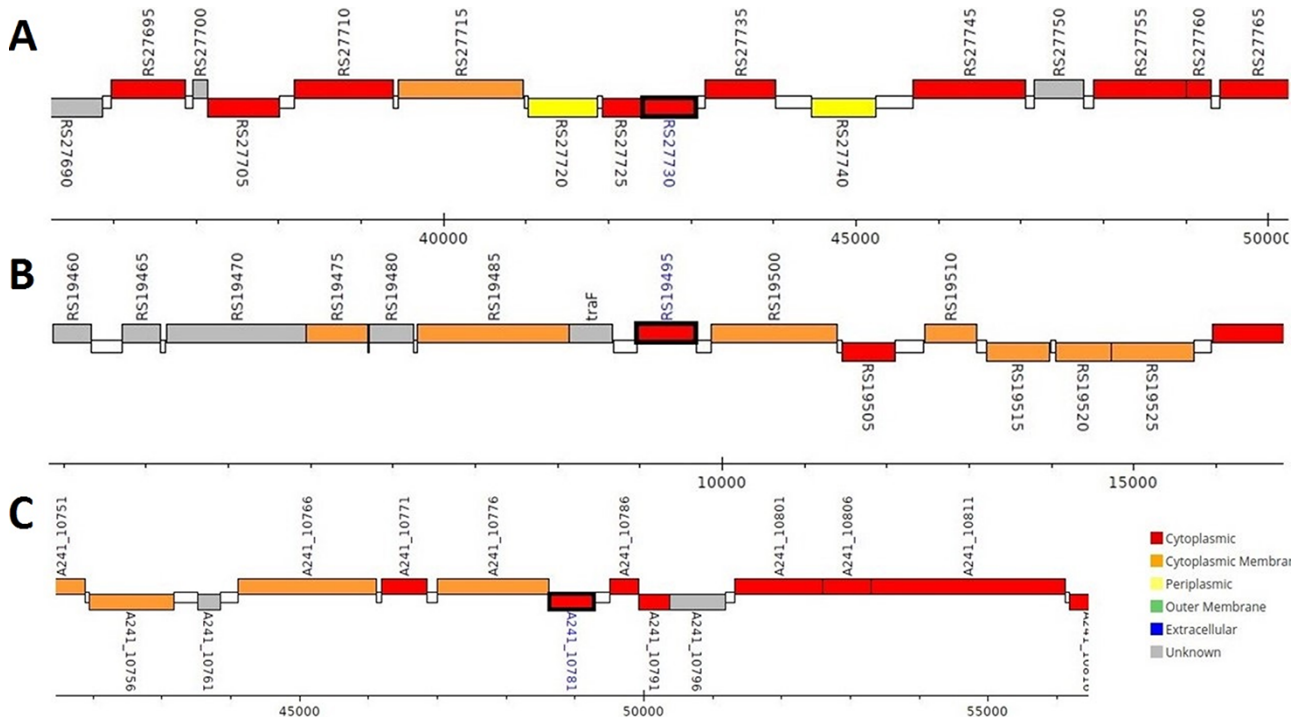

Figure 2. Genomic analysis of β-CA sequences from putative contaminants associated with *Pseudomonas* spp. The analysis shows the presence of coding genes for β-CA from (A) *Pseudomonas* sp. (UniProt ID: A0A0Q8Y2C1), (B) *Pseudomonas* sp. LP_8_YM (UniProt ID: A0A4R3W4C9) and (C) *Pseudomonas syringae* pv. actinidiae ICMP 19096 (UniProt ID: A0A656JXK1).

TBLASTN search tool with normal sensitivity was applied to search for the translated nucleotide databases using a protein query. In the next step, the defined β- and γ-CA protein sequences of WZSP were analyzed by the BLAST homology search tool of the UniProt database. In the final step, multiple sequence alignment (MSA) analysis was performed on all β- and γ-CA protein sequences involved in this evaluation using Clustal Omega algorithm of the European Bioinformatics Institute database (https://www.ebi.ac.uk/Tools/msa/clustalo/) (29). To reduce the size of protein sequences and output figures from MSA analysis, just 69 and 60 amino acid sequences of β- and γ-CA protein sequences containing the enzyme active sites were selected, respectively.

Genomic analysis of β- and γ-CA sequences from putative bacterial contaminants

The coding genes for β- and γ-CAs from *Pseudomonas* spp. as one of the putative contaminants in WGS of WZSP were evaluated using the BLASTP search tool in the *Pseudomonas* Genome Database, version 20.2 (https://www.pseudomonas.com/) (30) by using 1e-4 as the default value cutoff. In addition, the coding genes for β- and γ-CAs from *Acinetobacter* spp. as another potential contaminant were analyzed by the Ensembl Bacteria database (http://bacteria.ensembl.org/index.html) (31).

Protein structure analysis

Four β-CA protein sequences from bacterial contaminates including UniProt IDs: A0A0Q8Y2C1, A0A4R3W4C9, A0A656JXK1 and A0A062C2I7 and six γ-CA protein sequences from bacterial contaminants including UniProt IDs: A0A4R3W1J2, A0A125QD08, A0A4R3W9L6, A0A2N1E8I6, A0A062BNN8 and A0A419V156 were analyzed by RCSB Protein Data Bank (PDB) (https://www.rcsb.org/) (32) to identify the most similar crystallized and 3D model proteins to the query β- and γ-CA protein sequences of bacterial contaminants.

Results

Identification of α-CAs from *S. scrofa*

This analysis defined 12 α-CA isozymes including CA I–IV, CA VA and VB, CA VI, CA VII, CA IX and CA XII–XIV and three CARPs including CARP VIII, CARP X and CARP XI in *S. scrofa*. The results revealed that chromosome 1 contains the coding genes for CA IX and CA XII; chromosome 4 contains the coding genes for CA I–III, CA XIII, CAXIV and CARP VIII; chromosome 6 contains the coding genes for CA VA, CA VI, CA VII and CARP XI; chromosome 12 contains the coding genes for CA IV and CARP X and chromosome X contains the coding gene for CA VB. Our study on the subcellular localization of α-CAs from *S.*

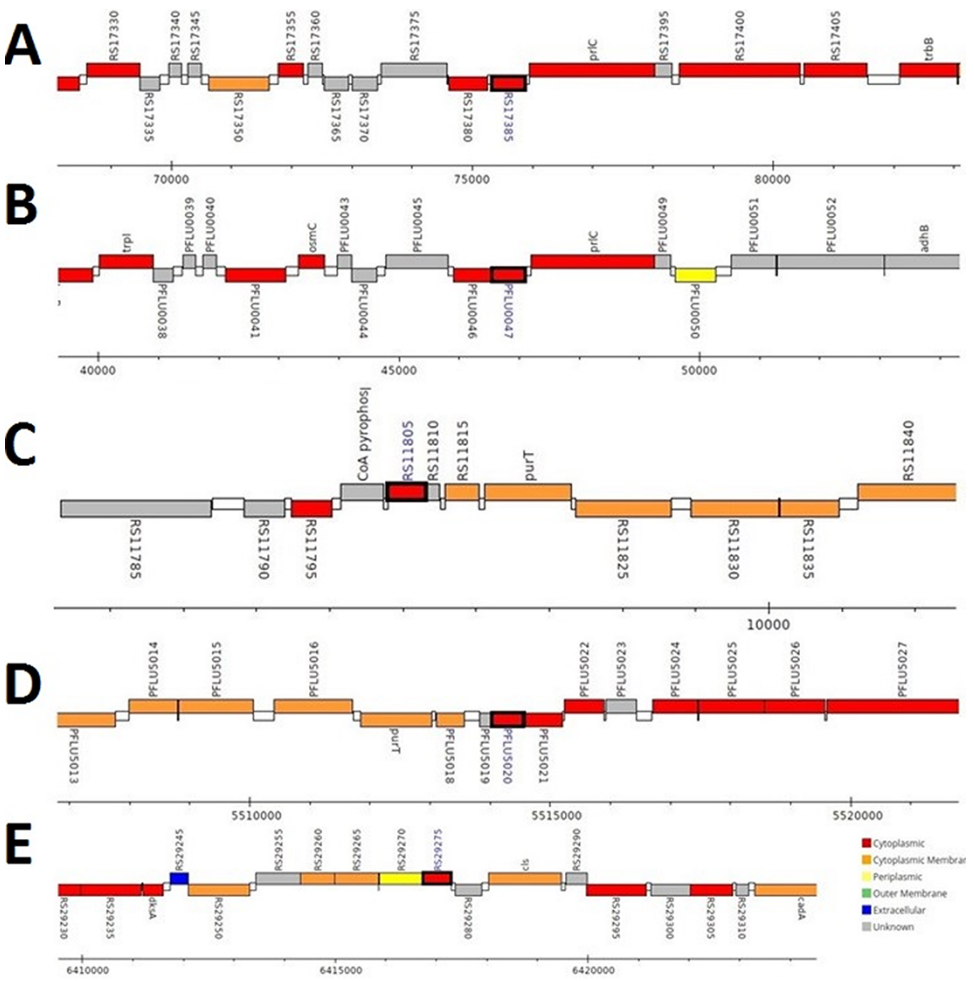

Figure 3. Genomic analysis of γ-CA sequences from putative contaminants associated with *Pseudomonas* spp. The analysis shows the presence of coding genes for γ-CA from (A) *Pseudomonas* sp. LP_8_YM (UniProt ID: A0A4R3W1J2), (B) *Pseudomonas fluorescens* (UniProt ID: A0A125QD08), (C) *Pseudomonas* sp. LP_8_YM (UniProt ID: A0A4R3W9L6), (D) *Pseudomonas fluorescens* (UniProt ID: A0A2N1E8I6) and (E) *Pseudomonas synxantha* (UniProt ID: A0A419V156).

scrofa predicted that CA I–III, CA VII, CA XIII and CARP VIII are cytoplasmic; CA VA and CA VB are mitochondrial; CA VI, CARP X and CARP XI are secretory; CA IX, CA XII, and CA XIV are transmembrane and CA IV is membrane-bound (Table 1).

Analysis of β- and γ-CA sequences

The BLAST homology analysis of the predicted WZSP CA sequences first identified a β-CA sequence from *A. aceti* and a γ-CA sequence from *S. flexneri*. A more detailed BLAST homology analysis of β-CA and γ-CA sequences from WZSP showed 100% similarity with bacterial β- and γ-CA sequences from *Pseudomonas* spp. and *Acinetobacter* spp. To confirm the identity of the defined sequences, MSA of the β-CA sequences showed the five highly conserved amino acids, including cysteine, aspartic acid, arginine (CXDXR) and histidine and cysteine (HXXC), which are known to be characteristic features of β-CA enzymes. Similarly, the predicted γ-CA sequences showed the four highly conserved amino acids characteristic of γ-CAs, including glutamine and histidine (QXXXXXH) as well as two histidines (HXXXXH) (Table 2; Figure 1).

Genomic analysis of β- and γ-CA sequences from putative bacterial contaminants

The analysis revealed that the *β*- and *γ-CA* genes from putative bacterial contaminants are located in the genomes of *Pseudomonas* spp. and *Acinetobacter* spp. Further evaluation revealed that all the encoded β- and γ-CAs from the putative bacterial contaminants are probably cytoplasmic proteins (Figures 2–4).

Protein structure analysis

The 3D models of crystallized β- and γ-CA protein structures, most similar to the bacterial contaminant proteins described in this study, were visualized in NGL (WebGL) viewer of the RSCB PDB database (accession codes 5JJ8 and 3PMO) (Figure 5). The visualized images of the bacterial β- and γ-CA proteins show homodimeric

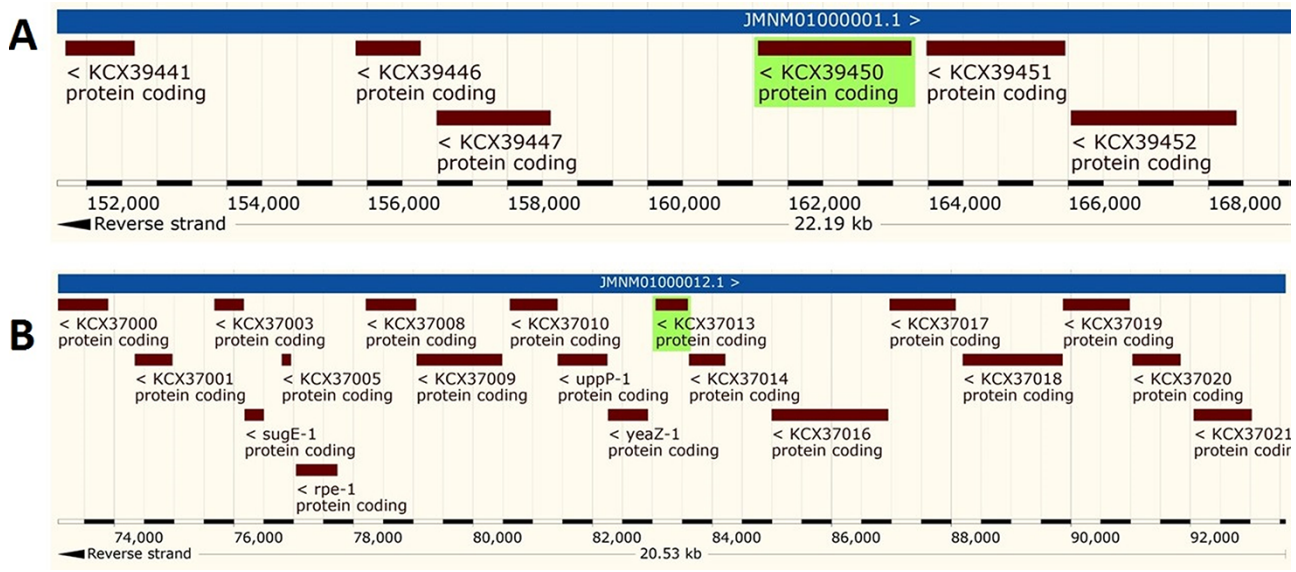

Figure 4. Genomic analysis of β- and γ-CA sequences from putative contaminants associated with *Acinetobacter* spp. The analysis shows the presence of coding genes for (A) β-CA from *Acinetobacter* sp. 263903-1 (UniProt ID: A0A062C2I7) and (B) γ-CA from *Acinetobacter* sp. 263903-1 (UniProt ID: A0A062BNN8).

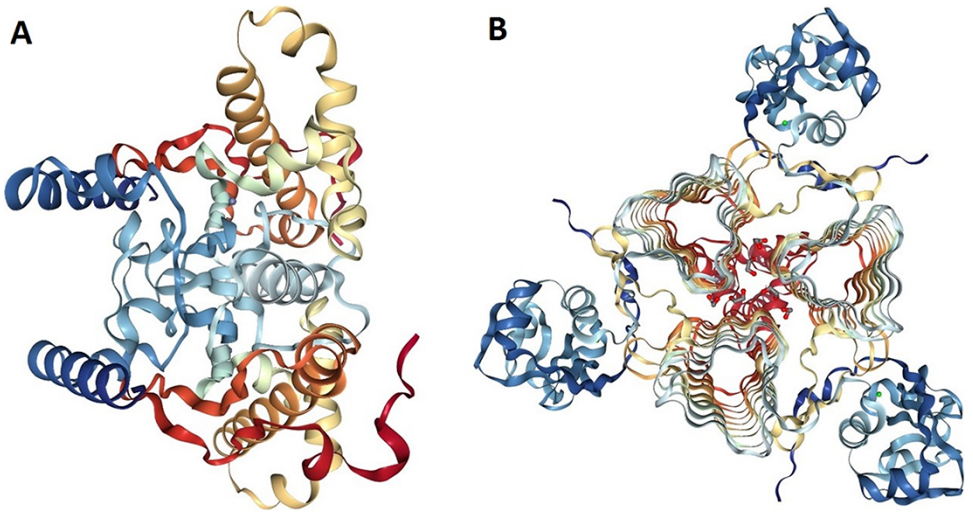

Figure 5. Protein structure analysis of β- and γ-CA protein sequences from bacterial contaminants. (A) Accession ID: 5JJ8 crystal structure belongs to β-CA from *P. aeruginosa*, and (B) Accession ID: 3PMO crystal structure belongs to γ-CA from *P. aeruginosa*. A and B were obtained from the PDB database, which are the most similar crystalized structures to β- and γ-CAs from bacterial contaminants, respectively.

and homotrimeric structures typical for the β- and γ-CA proteins, respectively (33).

Discussion

α-CAs have been classically considered the only CA family that is present in vertebrates. In line with those observations, our study revealed that *S. scrofa* has 12 α-CA isozymes and 3 CARPs similar to human (26). These α-CAs have subcellular localizations that are concordant with human enzymes, including cytoplasmic CA I–III, CA VII, CARP VIII and CA XIII; membrane-bound CA IV; mitochondrial CA VA and CA VB; secretory CA VI, CARP X and CARP XI; and transmembrane CA IX, CA XII and CA XIV (15).

Surprisingly, the first analyses of our study using the query bacterial β- and γ-CA sequences detected counterpart CA sequences in WZSP, and indeed, the MSA analysis approved that these sequences belong to the β- and γ-CA families. The BLAST search homology analyses of the identified β- and γ-CAs from WZSP displayed 100% identity to β- and γ-CA sequences from *Pseudomonas* spp. and *Acinetobacter* spp. In addition, genomic characterization of the detected β- and γ-CA sequences by the *Pseudomonas* Genome Database and Ensembl Bacteria database showed

the presence of corresponding β- and γ-*CA* genes in the genomes of *Pseudomonas* spp. and *Acinetobacter* spp., with cytoplasmic subcellular localization of the encoded CAs.

Previous studies have revealed that both host gut-associated flora and environmental microbiome, such as airborne microbes as well as bacterial contamination of equipment and solutions used for DNA isolation, can represent potentially interfering substances and contamination sources of the shotgun metagenomic sequencing samples, leading to false-positive results (34–36). For similar reasons, it would be highly possible that the isolated DNA samples from WZSP for WGS project had been contaminated with bacterial members of the Pseudomonadales order including *Pseudomonas* spp. and *Acinetobacter* spp., resulting in the detection of β- and γ-CAs from these bacterial species in the Ensembl assembly (minipig_v1.0) of *S. scrofa*. In addition, further analysis with protein structure modeling of β- and γ-CA sequences from bacterial contaminants revealed that β-CA sequences from contaminants were similar to 5JJ8 crystal structure from *P. aeruginosa*, and γ-CA sequences from contaminants were similar to 3PMO crystal structure from *P. aeruginosa*, which both approve the membership of β- and γ-CA sequences of bacterial contaminants to Pseudomonadales order.

There are different pipelines for decontamination of genomic reads in DNA-Seq and RNA-Seq projects, such as hierarchical clustering algorithm (37), RapMap (38), DecontaMiner (39), Sequencing Quality Assessment Tool or SQUAT (40), map-guided scaffolding or MaGuS (41), and Kraken 2 (42), which can improve the quality of genomic samples. DNA-free reagents and kits are used to reduce the bacterial contamination in the sequencing projects (43). Internal controls of every step in the sequencing protocols can detect the trace fragments of foreign DNA or RNA to reduce the risk of bacterial contamination (44). Nevertheless, our results demonstrate that the sequences present in genomic databases do contain incorrect sequences due to microbial contamination, underlining the need for high-quality internal controls and biocuration.

Conclusions

In addition to aforementioned methods for detection of bacterial contamination in the WGS projects of animals, the bioinformatics and computational approaches accompanied with bacterial-specific markers, such as CA sequences, can be employed to detect and reduce the risk of microbial contamination in the WGS projects through implementation of biocuration in databases. It is important to control the quality of short-size libraries, contigs and scaffolds as well as to perform internal checks of solutions, reagents and equipment during the shotgun genomic projects. This can be led to reducing the risk of annotation of false DNA and protein sequences in databases.

Acknowledgements

We thank the National Institute of Genetic Engineering and Biotechnology (NIGEB) of the Islamic Republic of Iran for preparing the condition to perform this study. No funding organizations had any role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; nor in the decision to publish the results.

Funding

National Institute of Genetic Engineering and Biotechnology (NIGEB) of the Islamic Republic of Iran (to R.Z.E.).

Author contributions

All authors participated in the design of the study. R.Z.E. and S.P. designed the study. R.Z.E. carried out the search to detect α-, β- and γ-CA sequences, performed bioinformatics and computational biology studies and drafted the first version of the manuscript. S.N.H. contributed to artwork preparation of the figures and preparing the manuscript for submission to the journal. All authors participated in writing further versions and read and approved the final manuscript.

Conflict of interest. The authors declare that they have no conflicts interests.

References

1. Larson,G. *et al.* (2010) Patterns of East Asian pig domestication, migration, and turnover revealed by modern and ancient DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 7686–7691.
2. Tong,X. *et al.* (2020) Whole genome sequence analysis reveals genetic structure and X-chromosome haplotype structure in indigenous Chinese pigs. *Sci. Rep.*, **10**, 9433.
3. Harbers,H. *et al.* (2020) Investigating the impact of captivity and domestication on limb bone cortical morphology: an experimental approach using a wild boar model. *Sci. Rep.*, **10**, 19070.
4. Esteban,M.A. *et al.* (2009) Generation of induced pluripotent stem cell lines from Tibetan miniature pig. *J. Biol. Chem.*, **284**, 17634–17640.
5. Gun,G. and Kues,W.A. (2014) Current progress of genetically engineered pig models for biomedical research. *Biores. Open Access*, **3**, 255–264.
6. Wang,L. *et al.* (2019) Genomic analysis reveals specific patterns of homozygosity and heterozygosity in inbred pigs. *Animals (Basel)*, **9**.
7. Fang,X. *et al.* (2012) The sequence and analysis of a Chinese pig genome. *Gigascience*, **1**, 16.
8. Ma,Y. *et al.* (2010) Identification of full-length proviral DNA of porcine endogenous retrovirus from Chinese Wuzhishan miniature pigs inbred. *Comp. Immunol. Microbiol. Infect. Dis.*, **33**, 323–331.
9. Zolfaghari Emameh,R. *et al.* (2016) Innovative molecular diagnosis of Trichinella species based on beta-carbonic anhydrase genomic sequence. *Microb. Biotechnol.*, **9**, 172–179.

10. Zolfaghari Emameh,R. *et al.* (2016) Identification and inhibition of carbonic anhydrases from nematodes. *J. Enzyme. Inhib. Med. Chem.*, **31**, 176–184.
11. Del Prete,S. *et al.* (2014) Discovery of a new family of carbonic anhydrases in the malaria pathogen Plasmodium falciparum—the eta-carbonic anhydrases. *Bioorg. Med. Chem. Lett.*, **24**, 4389–4396.
12. Kikutani,S. *et al.* (2016) Thylakoid luminal theta-carbonic anhydrase critical for growth and photosynthesis in the marine diatom Phaeodactylum tricornutum. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 9828–9833.
13. Jensen,E.L. *et al.* (2019) A new widespread subclass of carbonic anhydrase in marine phytoplankton. *ISME J.*, **13**, 2094–2106.
14. Del Prete,S. *et al.* (2020) Bacterial iota-carbonic anhydrase: a new active class of carbonic anhydrase identified in the genome of the Gram-negative bacterium Burkholderia territorii. *J. Enzyme. Inhib. Med. Chem.*, **35**, 1060–1068.
15. Zolfaghari Emameh,R. *et al.* (2014) Bioinformatic analysis of beta carbonic anhydrase sequences from protozoans and metazoans. *Parasit. Vectors*, **7**, 38.
16. Zolfaghari Emameh,R. *et al.* (2014) Beta carbonic anhydrases: novel targets for pesticides and anti-parasitic agents in agriculture and livestock husbandry. *Parasit. Vectors*, **7**, 403.
17. Hilvo,M. *et al.* (2005) Characterization of CA XV, a new GPI-anchored form of carbonic anhydrase. *Biochem. J.*, **392**, 83–92.
18. Aspatwar,A. *et al.* (2015) Inactivation of ca10a and ca10b genes leads to abnormal embryonic development and alters movement pattern in zebrafish. *PLoS One*, **10**, e0134263.
19. Sterky,F.H. *et al.* (2017) Carbonic anhydrase-related protein CA10 is an evolutionarily conserved pan-neurexin ligand. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, E1253–E1262.
20. Karjalainen,S.L. *et al.* (2018) Carbonic anhydrase related protein expression in astrocytomas and oligodendroglial tumors. *BMC Cancer*, **18**, 584.
21. Aspatwar,A., Tolvanen,M.E. and Parkkila,S. (2013) An update on carbonic anhydrase-related proteins VIII, X and XI. *J. Enzyme. Inhib. Med. Chem.*, **28**, 1129–1142.
22. Juozapaitiene,V. *et al.* (2016) Purification, enzymatic activity and inhibitor discovery for recombinant human carbonic anhydrase XIV. *J. Biotechnol.*, **240**, 31–42.
23. Zolfaghari Emameh,R. *et al.* (2016) Horizontal transfer of beta-carbonic anhydrase genes from prokaryotes to protozoans, insects, and nematodes. *Parasit Vectors*, **9**, 152.
24. Zolfaghari Emameh,R. *et al.* (2018) Involvement of beta-carbonic anhydrase genes in bacterial genomic islands and their horizontal transfer to protists. *Appl. Environ. Microbiol.*, **84**.
25. Yates,A.D. *et al.* (2020) Ensembl 2020. *Nucleic Acids Res.*, **48**, D682–D688.
26. Zolfaghari Emameh,R. *et al.* (2020) Assessment of databases to determine the validity of beta- and gamma-carbonic anhydrase sequences from vertebrates. *BMC Genomics*, **21**, 352.
27. Sayers,E.W. *et al.* (2020) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **48**, D9–D16.
28. UniProt,C. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
29. Sievers,F. and Higgins,D.G. (2014) Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol. Biol.*, **1079**, 105–116.
30. Winsor,G.L. *et al.* (2011) Pseudomonas Genome Database: improved comparative analysis and population genomics capability for Pseudomonas genomes. *Nucleic Acids Res.*, **39**, D596–D600.
31. Kersey,P.J. *et al.* (2012) Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Res.*, **40**, D91–D97.
32. Goodsell,D.S. *et al.* (2020) RCSB Protein Data Bank: enabling biomedical research and drug discovery. *Protein Sci.*, **29**, 52–65.
33. Ferry,J.G. (2010) The gamma class of carbonic anhydrases. *Biochim. Biophys. Acta*, **1804**, 374–381.
34. Fouladi,F. *et al.* (2020) Air pollution exposure is associated with the gut microbiome as revealed by shotgun metagenomic sequencing. *Environ. Int.*, **138**, 105604.
35. Fricker,A.M., Podlesny,D. and Fricke,W.F. (2019) What is new and relevant for sequencing-based microbiome research? A mini-review. *J. Adv. Res.*, **19**, 105–112.
36. Eisenhofer,R. *et al.* (2019) Contamination in low microbial biomass microbiome studies: issues and recommendations. *Trends Microbiol.*, **27**, 105–117.
37. Lafond-Lapalme,J. *et al.* (2017) A new method for decontamination of de novo transcriptomes using a hierarchical clustering algorithm. *Bioinformatics*, **33**, 1293–1300.
38. Srivastava,A. *et al.* (2016) RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. *Bioinformatics*, **32**, i192–i200.
39. Sangiovanni,M. *et al.* (2019) From trash to treasure: detecting unexpected contamination in unmapped NGS data. *BMC Bioinform.*, **20**, 168.
40. Yang,L.A. *et al.* (2019) SQUAT: a Sequencing Quality Assessment Tool for data quality assessments of genome assemblies. *BMC Genomics*, **19**, 238.
41. Madoui,M.A. *et al.* (2016) MaGuS: a tool for quality assessment and scaffolding of genome assemblies with Whole Genome Profiling Data. *BMC Bioinform.*, **17**, 115.
42. Wood,D.E., Lu,J. and Langmead,B. (2019) Improved metagenomic analysis with Kraken 2. *Genome Biol.*, **20**, 257.
43. Salter,S.J. *et al.* (2014) Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.*, **12**, 87.
44. Wurm,P. *et al.* (2018) Qualitative and quantitative DNA- and RNA-based analysis of the bacterial stomach microbiota in humans, mice, and gerbils. *mSystems*, **3**.