Novel small protein identification and quantitative proteomic analysis in Pseudomonas putida KT--2440

Yang, Xiaochen; Long, Katherine

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Novel small protein identification and quantitative proteomic analysis in *Pseudomonas putida* KT-2440

Ph.D. Thesis

Xiaochen Yang

June 2016

The Novo Nordisk Foundation Center for Biosustainability
The Technical University of Denmark

Supervisor:
Associate Professor Katherine S. Long

Novel small protein identification and quantitative proteomic analysis in *Pseudomonas putida* KT-2440

Ph.D. Thesis 2016 © Xiaochen Yang

The Novo Nordisk Foundation Center for Biosustainability

Department of Systems Biology

The Technical University of Denmark

Preface

The present Ph.D. study has been conducted at he Novo Nordisk Foundation Center for Biosustainability at the Technical University of Denmark from May 2013 to June 2016. The work is under the supervision of Associate Professor Katherine S. Long from the Technical University of Denmark.

The Ph.D. project was funded by he Novo Nordisk Foundation and the Ph.D. grant was from the People Programme (Marie Curie Actions) of the European Union`s Seventh Framework Programme (FP7-People-2012-ITN), under grant agreement no. 317058, "BACTORY".

Xiaochen Yang
Hørsholm, June 2016

Abstract

Bacterial cell factories offer an environmental-friendly and sustainable alternative for traditional chemical and fuel-based industry. Better understanding of industrial bacteria in systematic way is meaningful for the application of bacterial cell factories on broader spectrum of industry.

This thesis investigated an industrial bacterium, *Pseudomonas putida* KT-2440*,* in two aspects. First, the research focused on discovering novel small proteins (s-proteins) in the bacterium. With large-scale approaches for gene identification, groups of novel s-proteins were identified and validated from the genome, transcriptome and proteome of the bacterium. The application of new research approach, ribosome profiling, enabled us to analysis novel open reading frames (ORFs) from different standpoint. Second, by quantitative proteomic approach, the differential expressions of genes were analyzed at proteome level under different environmental conditions. The results yield insights into the adaptation of *P.putida* KT-2440 in different environments.

Based on bioinformatic, proteomic and transcriptomic approaches, global gene expression was analyzed on both transcriptional and translational levels. Our research for the first time validates novel s-proteins in *P.putida* KT-2440, and significantly increases the knowledge of *P.putida* protein expression and environmental adaptation.

Dansk resumé

Bakterielle cellefabrikker udgør et miljøvenligt og bæredygtigt alternativ til traditionel kemi- og oliebaseret industri. En bedre systematisk forståelse af industrielle bakterier er afgørende for brugen af bakterielle cellefabrikker i det brede spektrum af industrien.

Denne afhandling undersøgte en industriel bakterie, *Pseudomonas putida* KT-2440, på to måder. Første del fokuserede på opdagelsen af nye små proteiner (s-proteiner) i bakterien. Ved brug af storskalatilgange til genidentificering blev grupper af nye s-proteiner identificeret og valideret fra genomet, transkriptomet og proteomet af bakterien. Anvendelsen af en anden metode, ribosomprofilering, lod os analysere nye åbne læserammer (ORFs) fra en anden synsvinkel. Anden del, ved en kvantitativ proteomtilgang, lod os analysere differentielt udtryk af gener under forskellige vækstbetingelser. Resultaterne giver indsigt i adapteringen af *P. putida* KT-2440 til forskellige miljøer.

Baseret på bioinformatiske, proteom- og transkriptomtilgange, blev globalt genudtryk analyseret på både transkriptom- og proteomniveau. Vores forskningen validerer for første gang nye s-proteiner i *P. putida* KT-2440 og øger betragteligt videnen om *P. putida* proteinudtryk og miljømæssig adaptering.

Acknowledgements

Three years of Ph.D. study filled with a lot of challenges and joys. I would like to express my gratitude to all the people that supported and helped me during these three years.

First of all, I would like to thank my supervisor Katherine S. Long for the responsible supervision. You are a precise supervisor that take care of details of work and help me to make thorough plans for experiments. You respected students` idea and give me freedom to try with my idea. You were responsible for my project and all the challenges I met during the experiments can be discussed in time through our formal and spontaneous meetings. You worked a lot on my internal and external cooperation, which improved the progress of the project.

I would also like to thank Christine Finnie, my former co-supervisor that taught me a lot about proteomics and mass spectrometry, and corrected the introduction part of my thesis.

I would like to thank Lasse Ebdrup Pedersen, Anna Koza and Sheila I. Jensen for your help on ribosome profiling and small protein expression. Thanks to Stefen Kol, Tune Wulff and Scott Harrison for helping me with my work on mass spectrometry and western blot. I would also thank Marta Matos and Emre Ozdemir for the help on bioinformatics analysis.

A big thank to colleagues in my group, Holger Døssing, Martin Holm Rau, Klara Bojanovic, Isotta D'Arrigo and Mikkel Lindegaard, as well as our 14 BACTORY Ph. D. students. Thanks to all the people in my office and neighbor office, Se Hyeuk, Dario, Particia, Sofie, Malfalda, Maja, Ida, Virginia, Rosa. You guys provide a wonderful working environment, and I really enjoyed the time with all of you guys, not only during the work, but also leisure time.

I would also thank Tian Zhan, Jiangfeng Zhu, Yaojun Tong, Xinglin Jiang, Mingji Li, Songyuan Li and other Chinese friends in Denmark while you guys always made me homesick.

I really thank Thomas Kallehauge, Christian Bille Jendresen, Ufuk Celebioglu, and many other colleagues in DTU biosustain and DTU system biology for discussion and helping with my work. My life will be hundred times harder without support and help from your guys. I would also like to thank Richard Szubin and Justin Tan from University of California, San Diego, for sharing their working experience.

I would like to express my thanks of gratitude to Søren Mollin and Mette Munk. Thanks for establishing this amazing programme and taking care of our Bactory students during the last three years.

I would like to express my special gratitude to our overseas cooperators. Many thank Daniela Zühlke and Katharina Maria Riedel from Ernst-Moritz-Arndt-University, Greifswald, Germany. Thank you for training me quantitative proteomic approaches and helping me to finish the experiment. *Herzlichen Dank*. I would also thank Stefano Donadio, Margherita Sosio and other people from the company Naicons, Milan, Italy. Thank you for training me in the company and warm reception. *La ringrazio tanto.*

At last, I would thank my parents, grandparents, my uncles and cousins. You always believe me, support me and respect my choice. While it was thousands miles away, I can still feel the support and love from you. I love you forever. I would also thank my friends in Asia, Europe and America. While you guys live in different places all over the world, our friendship will never end.

Xiaochen Yang
June 2016
Hørsholm

Publications

Included in the thesis

Xiaochen Yang, Sheila I. Jensen, Tune Wulff, Scott J, Harrison, Katherine S. Long. (2016) Identification and validation of novel small proteins in *Pseudomonas putida*. Manuscript in submission.

Xiaochen Yang, Daniela Zühlke, Katharina Maria Riedel, Katherine Long. (2016) Quantitative analysis of the cytosolic proteome of *Pseudomonas putida* under stress conditions and in different nutrient environments. Manuscript in preparation.

Xiaochen Yang, Lasse E. Pedersen, Anna Koza, Katherine S. Long. (2016) Ribosome profiling analysis and novel small ORF identification in *Pseudomonas putida*. Manuscript in preparation.

Not included in the thesis

D'Arrigo, I., Bojanovič, K., **Yang, X**., Rau, M. H., & Long, K. S. (2016). Genome-wide mapping of transcription start sites yields novel insights into the primary transcriptome of *Pseudomonas putida*. *Environmental Microbiology*. doi:10.1111/1462-2920.13326

Table of Contents

Chapter 1. Introduction

Proteins are major functional molecules in living organisms and are involved in most biological processes. In bacteria, protein expression not only maintains the basic cellular functions but also affects specific characteristics of the cell, such as stress resistance, compound production, etc. Two strategies are often used for analyzing protein function and its effect on organism characteristics, including exploration of novel proteins and analyzing the behavior of known proteins, such as their expression conditions, localization, and the biological processes they are involved in. In this thesis, both strategies were adopted to analyze novel small proteins and adaptation of *Pseudomonas putida* in different environmental conditions.

Pseudomonas putida is a Gram-negative bacterium that is tolerant to many antibiotics and environmental stresses (1). Due to the non-pathogenicity and capability for toxic organic reagent degradation (2), *P. putida* is an appropriate bacterium for industrial application in bioremediation, biodegradation and chemical production. With this bacterium, we first focused on novel small proteins identification.

Small proteins (s-proteins) are defined as proteins with 50 amino acids or less. Their small size leads to difficulties in their identification and validation. The major challenge of s-protein identification is the lack of appropriate research methods. Both bioinformatic approaches, such as bioinformatic prediction and homology analysis, and experimental approaches, such as protein analysis and transcription analysis, have obvious weaknesses for s-protein analysis. First, based on the methods that bioinformatic tools use for ORF prediction and identification, short ORFs are easily missed from identification (3, 4). Second, the contradiction between sensitivity and accuracy of experimental approaches limits their application for s-protein analysis. This challenge can be partly solved by combining different approaches and experimental validation of protein expression.

Environmental factors have impact on protein expression of bacteria. Proteins can be differentially expressed in different conditions either in response to certain environmental

factors or are involved in resistance towards toxic compounds. The second focus of the thesis addresses the environmental adaptation mechanisms of *P. putida*. Using quantitative proteomics approach, the proteome of *P.putida* under two stress conditions and growing in two different media was analyzed, and proteins involved in the response towards the different conditions were discussed.

This thesis contains five chapters. The first chapter (Chapter 1) introduces the field that the thesis focuses on and the motivation of the research. The second chapter (Chapter 2) introduces sORFs in bacteria and approaches for novel sORF identification. The third chapter (Chapter 3) focuses on quantitative proteomics and its application to the analysis of multiple samples. Conclusions and perspectives are provided in the fourth chapter (Chapter 4). The full-length papers and manuscripts are enclosed in the last chapter (Chapter 5).

Chapter 2. Small protein identification and validation

2.1 Small proteins in bacteria

Small proteins with fifty amino acids or less have been known as ribosomal proteins, leader peptides or toxic peptides for decades (5). However, more recent characterization of additional small proteins has revealed that they have diverse functions. One of the first characterized functional small proteins was SpoVM(6, 7), a 26-amino-acid membrane-associated protein initiating spore formation by recognizing the spore formation site and recruiting ATPase(8). It was characterized in a *Bacillus subtilis* strain and conserved in endospore-forming species (9). Further analysis on spore formation in *Bacillus* led to the discovery of another functional small protein CmpA, a 37-amino-acid protein that inhibits the cortex assembly and negatively regulates the spore formation process(10, 11).

Research on *Escherichia coli*, *Salmonella* and other bacteria has revealed the ubiquity and functional diversity of small proteins in bacteria. In *Bacillus*, *Caulobacter*, and *Escherichia coli,* small proteins (MciZ, SidA, Blr) have been confirmed to participate in cell division. They regulate cell division either by associating with the key factor of the divisome, FtsZ, (MciZ in *Bacillus*)(12), or other components of divisome (SidA and Blr in *Caulobacter* and *Escherichia coli*) (13–17). Some transporters and enzymes are also regulated by small proteins. One example is AcrZ, a regulator of transporter that is involved in stress response of bacteria. The AcrZ is confirmed to interact with the AcrB-AcrA-TolC efflx pump in *E. coli*. The expression of AcrZ is induced by antibiotics, detergents, and oxidizing compounds, and the mutation of AcrZ undermines the resistance of bacteria to antibiotics. This suggests that AcrZ promotes the AcrB-AcrA-TolC efflux pump to export substrates (18, 19). Another example is CydX, a 37-amino-acid conserved protein that modulates cytochrome bd oxidase in *E. coli*. It regulates the oxidase activity by interacting with the CydA subunit of cytochrome bd oxidase (20, 21). Small proteins also interact with protein kinases and thus regulate signal transduction. One example is a cytosolic small protein, Sda, This 46 amino-acids protein inhibits the sporulation signal transduction pathway KinA/KinB-Spo0F-SpoOB-Spo0A in *Bacillus subtilis* by inhibiting

the kinase activity of KinA(22, 23). Another example is MgrB, an *E.coli* protein involved in regulation of the PhoQ-PhoP system. The PhoQ-PhoP system regulates groups of genes in response to stress conditions (24), such as low magnesium and the presence of some antimicrobial peptides (25). The *mgrB* gene, which encodes MgrB protein, is also regulated by PhoQ-PhoP system (26). However, studies indicated that MgrB interacted with PhoQ, resulting in negative feedback with the PhoQ-PhoP signal transduction system (27). In bacteria, small proteins also act as chaperones in biological processes. Two examples are the small proteins, FbpB and FbpC,that are regulated by the level of iron in the environment. The function of FbpB and FbpC is to facilitate the function of a rRNA, FsrA, in gene expression regulation (28–30). (Table 1)

Table 1: Well-characterized s-proteins in bacteria[a].

Name	Size (aa)	Functions	Organism[b]	Location	Reference
MciZ	40	Cell Division	*B.subtilis*	Membrane	(12)
SidA	29	Cell Division	*C. crescentus*	Membrane	(13, 14)
Blr	41	Cell Division	*E.coli*	Membrane	(15–17)
MntS	42	Chaperone	*E coli*	Cytoplasmic	(31, 32)
FbpB	48	Chaperone	*B.subtilis*	Cytoplasmic	(28, 29)
FbpC	29	Chaperone	*B. subtilis*	Cytoplasmic	(28, 30)
CydX	37	Membrane-Bound Enzyme Regulator	*E.coli*	Membrane	(20, 21, 33)
PmrR	29	Membrane-Bound Enzyme Regulator	*Salmonella*	Membrane	(34, 35)
MgtR	30	Membrane-Bound Enzyme Regulator	*Salmonella*	Membrane	(36, 37)
MgrB	47	Regulator of Protein Kinase and Signal Transduction	*E. coli*	Membrane	(24, 26, 27, 38)
Sda	46	Regulator of Protein Kinase and Signal Transduction	*B. subtilis*	Membrane	(22, 23, 39)
SpoVM	26	Spore Formation	*B.subtilis*	Membrane	(6–9)
Cmp	37	Spore Formation	*B. subtilis*	Membrane	(10, 11)

A					
KdpF	29	Transport regulation for kdpABC K^+ transporter	*E. coli*	Membrane	(40, 41)
AcrZ	49	Transport regulation for AcrB-AcrA-TolC efflux pump	*E. coli*	Membrane	(18, 19)
SgrT	43	Transport regulation for glucose uptake	*E.coli*	Membrane	(42, 43)

[a]Ribosomal proteins, small toxins, and leader peptides are not included in this table.
[b]The bacterium where the s-protein was first identified with experimental approaches is indicated.

Tens of functional small proteins have been characterized in bacteria. While small proteins have various functions in bacteria, most of those characterized to date are membrane-associated proteins. They are either inserted into the membrane directly or interact with large membrane proteins. Due to their small size, small proteins have limitations as far as adopting complicated three-dimensional structures. Thus it is unlikely that they contain typical functional domains that can be identified. However, through the interaction with functional proteins or domains, small proteins can block or activate biological processes. This indicates that many s-proteins would probably be involved in regulation of biological processes.

2.2 Small protein discovery and sORF identification

Most of the well-characterized small proteins were discovered serendipitously. They were identified either by mutation or deletion of intergenic regions adjacent to functional genes(6, 7, 10), or by re-identification of RNA sequences (31, 42). In the former case, the small proteins were identified only if the mutation or deletion displayed a strong phenotype. In the latter case, small proteins were identified by sequence conservation and ORF prediction of RNA sequences, in most cases, small regulatory RNAs. In addition, some small proteins were identified by their protein-protein interactions with well-known proteins involved in biological processes (12, 34, 35). However, random discovery of small proteins is largely dependent on the phenotype and functional research on specific biological processes. Approaches that are used for s-protein random discover are limited

on novel small protein identification. Thus, a genome-wide approach is required for small protein identification and validation.

The development of genomics and DNA-sequencing techniques for the first time gives access to genome-wide information of organisms. The improvement of bioinformatics provides tools for genome-wide gene annotation as well as sORF prediction. In addition, transcriptomics and proteomics approaches give the possibility to identify novel small proteins at the transcription and translation level.

2.2.1 Genome-wide gene prediction and sORF prediction in prokaryotes

As a quick, large-scale approach for data analysis, bioinformatic approaches are widely used for ORF prediction and genomic analysis. The genome annotation, which provides most of genetic information about a certain organism, is also established by bioinformatic approaches. Based on several bioinformatics methods, a list of ORF prediction tools have been developed for analyzing different types of DNA sequence data for different purposes (Table 2).

The first and also the simplest method is a sequence graphic search. With this method, the bioinformatic tools identify ORFs by checking the feature sequences of a gene, such as start codon, stop codon, and SD sequence, from sequence data. For example, the bioinformatic tool, ORF finder (http://www.ncbi.nlm.nih.gov/orffinder), is used to check the start and stop codon from long DNA sequences. An online tool, RBS finder (http://parts.igem.org/Ribosome_Binding_Sites), is used to check the possible ribosome-binding site from DNA sequences. This method is also adopted by some genetic research platforms, such as CLC Workbench, as a tool for ORF prediction.

While it gives a simple idea of possible ORFs on DNA sequences, the sequence-feature-based bioinformatic method has its problem on ORF prediction and is not credible enough for genome annotation. First, the sequences with start and stop codon in a bacterial genome do not necessarily encode any proteins or RNAs. Second, most of bioinformatics

tools with RBS analysis search for the Shine-Dalgarno sequence (SD-sequence) as the ribosome binding site. However, while the SD-sequence initiates translation (44) and has a great impact on protein expression levels (45), the typical SD-sequence does not always exist in the 5′ -flanking region of a gene (46). A recent study in *E.coli* identified an alternative cytosine-rich sequence that is unmistakably complementary to 16S rRNA and initiates translation (47), suggesting that the SD-sequence is not the only ribosome binding site on mRNAs. A ribosome profiling study in bacteria *E.coli* and *B. Subtilis* concluded that the SD-like features are associated with the pausing of translation (48), indicating that groups of SD-like sequences are located within ORFs. These studies on the SD-sequence demonstrate that bioinformatic ORF prediction based on the SD-sequence feature may lead to missed and mis-prediction of ORFs.

The method that is used for genome annotation and ORF prediction is based on a comparative genomics approach. In this method, the genome sequence or DNA sequences obtained from DNA libraries, cDNA libraries, metagenomics, are used to compare genetic databases from other bacterial strains. The sequence comparison, in general, is achieved by a bioinformatic model called the Markov model, and its variants (49). This model is a self-training system that captures the features of groups of genes and uses these features for ORF prediction. In this model, by comparing with large numbers of coding regions from multiple bacteria strains, different combinations of nucleotides will be given a value of possibility, which is used to evaluate the possibility of a coding sequence. The possible coding regions can be isolated from non-coding regions by calculating this value across the entire sequence. The regions with high possibility values will be predicted as ORFs(3). This model is widely adopted by popular ORF prediction software including Glimmer v1.0, the tool used for genome annotation in pseudomonads (3). An important update of the bioinformatic model is called the interpolated context model (ICM), and it is adopted by Glimmer v2.0. In comparison to Glimmerv1.0, Glimmer v2.0 is more efficient in resolving overlapping genes(4).

A recent improvement of ORF prediction is to predict the RBS with a comparative genomics method. This is based on the fact that many ORFs do not have typical SD

sequences in their 5´-flanking regions. By combining RBS and coding region prediction, the method increases the accuracy of results(50).

Table 2: Popular bioinformatics tools for ORF prediction in bacterial genomes

Aprroaches	Principle for ORF prediction	Oganism	Reference
FrameD	Interpolated Markov model and RBS analysis	Prokaryotes and eukaryotes	(51)
GeneMark	Homogeneous and non-homogeneous Markov models	Prokaryotes and eukaryotes	(52, 53)
GeneTack	Hidden Markov model	Prokaryotes	(54, 55)
MED	EDP model	Prokaryotes	(56)
GLIMMER 3.0	Interpolated Markov model and RBS analysis	Prokaryotes	(48–50)
EuGene	Markov model	Prokaryotes and eukaryotes	(59, 60)
AMIGene	Markov model	Prokaryotes	(61)
Prodigal	Conservation and RBS analysis	Prokaryotes	(62)
MetaGene-Annotator	Heuristically derived model and RBS analysis	Prokaryotes and Phage	(63–65)
EasyGene	Hidden Markov model	Prokaryotes	(66, 67)
ORPHEUS	Seed ORFs model and RBS analysis	Prokaryotes	(68)
EcoParse	Hidden Markov model	Prokaryotes	(69)
BactgeneSHOW	Hidden Markov model	Prokaryotes	(70)

Bioinformatic approaches have been used for sORF prediction in bacteria. In the *E. coli* K-12 MG1655 strain, more than 60 small proteins were identified by comparative genomic approaches (71, 72). Most of these small proteins were further validated by expression experiments (73, 74). A bioinformatic program, BactgeneSHOW(70), was used to predict sORFs in streptococci. This program uses a hidden Markov model (HMM) that calculates both the presence of RBSs and the nucleotide composition of the coding region. Thirty-five ORFs encoding less than 60 amino acids were predicted, including 24 sORFs that were shorter than 150bp (75).

2.2.2 Proteomic approaches for small protein identification

Proteomic analysis is used for the large-scale study of proteins and aims at obtaining global information of the whole proteome or complicated protein samples. Many techniques such as immunoassay, protein chip and electron microscopy have been used for proteome analysis (76). Another technique, mass spectrometry was also used for proteomics analysis for decades. The mass spectrometry based proteomic approach is established based on tandem mass spectrometry (MS/MS). With this approach, the protein sample first needs to be denatured and digested into small peptides. Then the peptides are ionized to create the fragments that can be screened by mass spectrometry. Two series-connected mass spectrometry instruments (MS/MS) are used to separate the peptide fragments by m/z ratio (mass to charge ratio) and obtaining final mass spectra of each fragment for data analysis. The amino acids sequences of the fragments are obtained by mass spectrometry analysis and the protein sequences are obtained by comparing the fragment sequences with a protein database. (Figure 1)

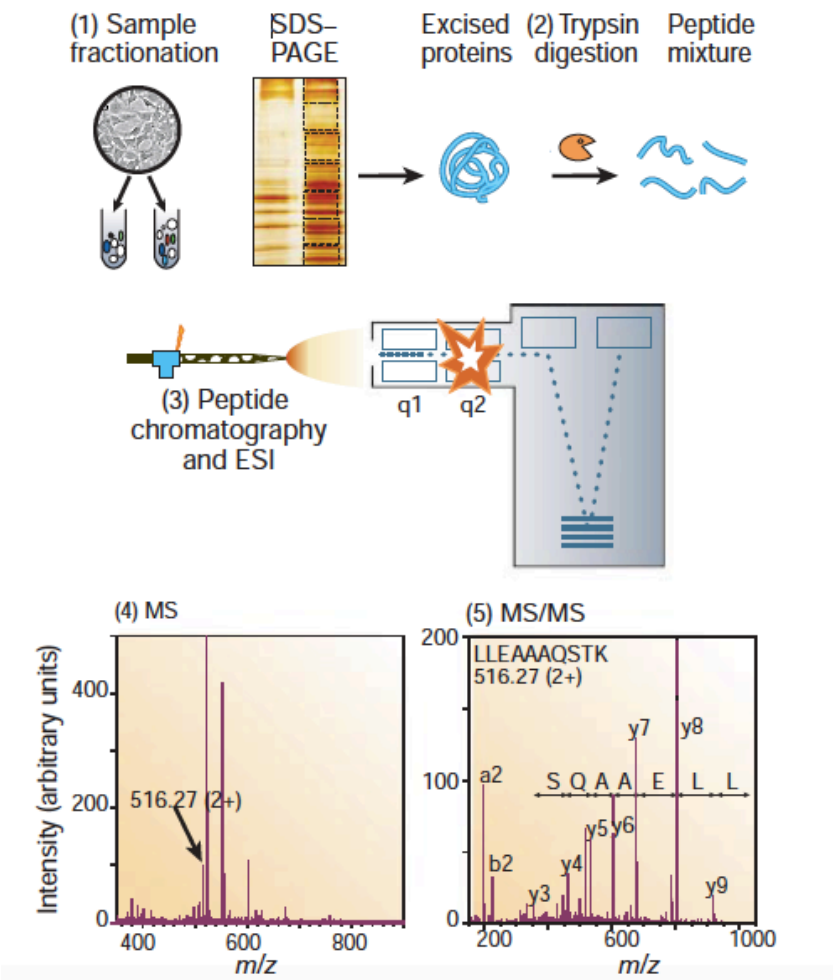


Figure 1: Workflow of the MS/MS approach (76). (The picture was obtained from reference 76). *Step 1: Protein sample preparation and fractionation. Step 2: Enzymatic digestion of protein into peptides. Trypsin is one of the most popular enzymes for protein digestion. Step 3: Ionization of peptides with Electrospray ionization (ESI)*(77) *or Matrix-assisted laser desorption/ionization (MALDI)*(78, 79)*. Step 4: Ionized peptides are analyzed with first MS and mass spectrum for peptides are recorded. Step 5: Fragmentation of peptides by energetic collision with gas and MS/MS spectrum are recorded.*

The first generation of proteomic MS approaches was established based on 2D-gel and MS/MS techniques (80). In these approaches, the total proteins are first separated by 2-dimensional polyacrylamide gel electrophoresis. The protein dots on the gel are then collected for MS/MS analysis. The major advantage is that, based on the staining color and locations of dots on the gel, the expression level and kinds of proteins can be

evaluated prior to mass spectrometry. However, the disadvantages are significant. Due to the limited number of dots that can be shown on the gel and analyzed by MS/MS, the approach only provides a partial view of the proteome. Another limitation is that 2D-gel electrophoresis is suitable for abundant proteins with appropriate masses (between 30 to 200 kDa). Proteins with low expression levels or having extremely large and small sizes are easily missed in these experiments (81–83).

The 2D-gel method was eventually replaced by gel-free approaches where a liquid chromatography column is used to separate total proteins before MS/MS analysis (LC-MS/MS). The advantage of this approach is that the whole proteome will be analyzed by MS/MS. While mass spectrometry-based proteomic analysis has been continuously updated and methods with specific goals established during the last twenty years, the principle of MS/MS analysis is similar (Figure 1).

While the MS-based proteomic approaches have been used for small protein identification in eukaryotes, few studies have been reported in bacteria. One example is from the genome-reduced bacterium *Mycoplasma pneumonia* (84). In this study, the total protein samples were fractionated by either SDS-gel or size exclusion chromatography. The samples were then analyzed by MS/MS together with the total protein sample without fractionation and total DNA binding protein samples. Thirty-five small proteins, including 14 previously unknown small proteins, were identified and many of small protein genes located within larger annotated genes. This suggests that sORFs are misannotated during genome annotation and emphasizes that sORFs are generally overlooked. Another study in *Pseudomonas fluorescens* Pf0-1(85) detected 2 novel sORFs by MS/MS. The bioinformatic analysis showed that none of these novel sORFs were predicted by the gene prediction tools GenemarkS and Glimmer, indicating that some sORFs are not detected via bioinformatic gene prediction.

2.2.3 Ribosome profiling and transcriptomic approaches for small protein identification

Ribosome profiling is a novel technique established by Nick Ingolia and Jonathan Weissman for monitoring *in vivo* gene translation(86, 87). This technique is based on the fact that gene translation relies on the binding of ribosomes to mRNAs and the movement of ribosomes along the coding regions of mRNAs. By deep sequencing of ribosome footprints on mRNAs, the translated regions of mRNAs can be identified(88, 89). With ribosome profiling technique, ribosome footprints are obtained by isolating and sequencing the mRNA fragments that are protected by ribosome (87, 90). The sequencing data is mapped with genomic databases and compared with mRNA-seq data to analyze the translated regions (Figure 2).

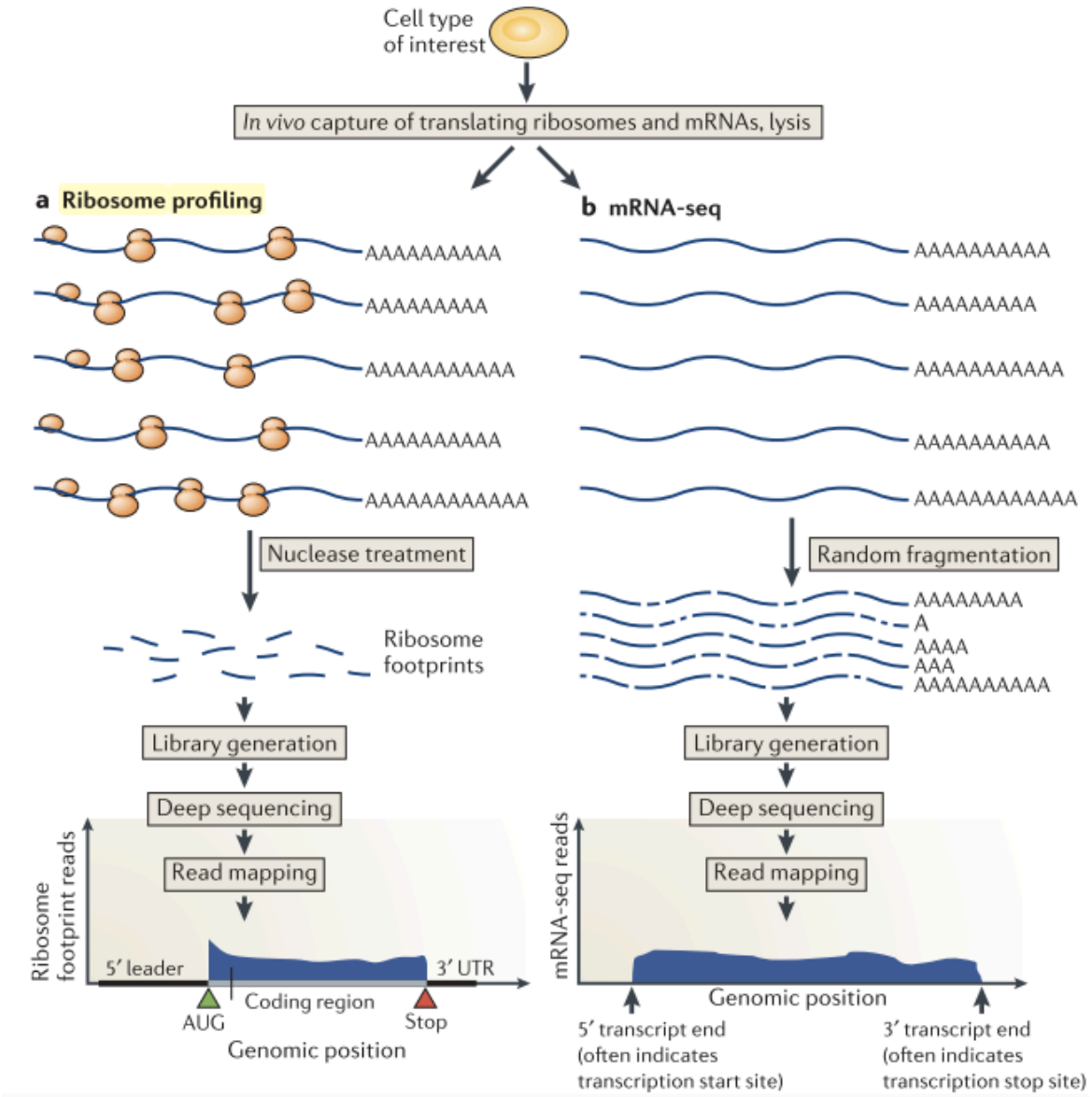


Figure 2: **An overview of ribosome profiling** (91). (Figure adapted from reference 91)

a). Ribosome-bound mRNAs are isolated from a cell lysate. A nonspecific nuclease, such as RNase I and micrococcal nuclease, is used to degrade mRNAs that are not protected by ribosomes, resulting in ribosome-protected mRNA fragments termed "footprints". The ribosome footprints are used to prepare a library for deep sequencing. With mapping the sequencing reads of ribosome footprints on the mRNA or genome sequences, the coding regions are identified. b). The mRNA sequencing data is used to compare with ribosome profiling data. The mRNAs are fragmented randomly and fragments of mRNAs are used to prepare a library for deep sequencing. The mRNA-seq data offers information of approximate transcript boundaries.

As a newly established genomic-wide analysis approach, ribosome profiling has been used for the identification of translated regions, monitoring protein translation and maturation, and measurement of cellular protein synthesis in various oraganisms, including eukaryotes, prokaryotes and viruses (92–100). Ribosome profiling analysis in vertebrates, insects and yeast has led to the discovery of large number of un-annotated ORFs(95–100). Many of them were located in the 5′-UTRs of mRNAs and identified as upstream ORFs (uORFs), similar to ORFs encoding leader peptides in bacteria (101).

The ribosome profiling technique has been used for coding region analysis in several bacterial species. One example is recent work on ORF identification in *Caulobacter crescentus* (102). In the study, the coding regions of bacterial genomes were identified by ribosome profiling, RNA-seq, LC-MS/MS and global 5′-RACE. Ninty-four novel sORFs encoding proteins less than 50 amino acids were identified. This study also proved that SD sequence is not only necessary for translation but also for pausing of ribosomes inside of coding regions in many cases in bacteria. Another example is a study in Mycobacteria (103). By adopting ribosome profiling and RNA-seq, 22 unannotated ORFs were identified. In addition, ribosome profiling has also been adopted also in research on *E.coli*. One example is the study of sRNA regulatory targets (104). Based on quantitative analysis of ribosome profiling data and RNA sequencing data, a list of genes that were regulated by an sRNA, RyhB, including novel targeted genes, were obtained. In bacteriophage lambda, ribosome profiling was used to identify novel ORFs in the genome. Based on ribosome profiling data, 55 potential novel ORFs were identified (105).

The major improvement of the ribosome profiling technique is that the protein translation can be analyzed directly from RNA sequencing data. This is not possible with microarrays or normal RNA sequencing techniques. In addition, the DNA strand that the ribosome footprint is located on can be identified, indicating that overlapping genes on different DNA strands can also be distinguished by the ribosome profiling approach.

2. 3 Genome annotation and annotated sORFs in *P. putida* KT-2440

The genome of *P.putida* KT2440 was annotated twice. The first annotation was in 2002 when the genome was sequenced (106).The GLlMMER was used for genome annotation and the ORF membership in families and superfamilies was identified by two sets of hidden Markov models (HMMs), PFAM(107, 108) and the TIGR orthologue resource (109). The functions of ORFs were determined by sequence conservation. A total of 5420 ORFs were annotated within the chromosome of 6,181,873 bp, including 164 sORFs with less than 154 bp. Within 164 annotated sORFs, 79 sequences were classified as coding sequences (CDS) encoding hypothetical proteins, one sORF (PP_0380) is identified as *pqqA* that encodes a small protein involved in coenzyme PQQ synthesis. Two genes encoded ribosomal proteins comprised of less than 51 amino acids. The remaining 82 sequences were classified as tRNA or rRNA genes.

The genome of *P. putida* KT2440 was re-annotated in 2016. Based on the genome sequencing and transcriptomic data, the genome was re-annotated with two bioinformatic tools, AMLGene and Prodigal. The AMLGene annotated the genome by a comparative genomics approach while another bioinformatic tool, Prodigal, annotated the genome by the sequence conservation and analysis of ribosome binding sites (110). The functional annotation was established based on the knowledge of metabolic pathways. A total number of 242 new ORFs, including 6 new sORFs were annotated. One of the newly annotated sORFs was identified as KdpF while others were identified as hypothetical proteins with unknown function.

The studies on genome annotation and re-annotation detected 86 sORFs, while none of them have been validated by experimental approaches. Except one ORF that has been identified as pqqA gene, none of the sORFs have identified functions, indicating that small proteins might have unique functions that remain undiscovered.

2.4 Challenges for small protein identification and detection

The bottlenecks of small protein research are protein identification and detection. Due to their small size, small proteins are difficult to detect by classical research methods. For example, the expression of small proteins is difficult to detect by classical protein electrophoresis and coomassie staining due to low expression levels and small size. While there are new methods such as improved protein gel systems and peptide tags that can facilitate small protein detection, there are still lots of challenges in protein identification. Bioinformatic, proteomic and transcriptomic approaches have been used for sORF and small protein identification. However, there are some challenges that limit the performance of these approaches with small protein analysis that are described below.

2.4.1 Challenges for bioinformatic prediction

Due to the fact that most of bioinformatic approaches and genomic databases are designed for large protein or normal ORF prediction, the main challenge of bioinformatic prediction is the compatibility and credibility of bioinformatic tools for small protein prediction.

First, due to the fact that sORFs lack sufficient sequence for domain and homology analysis, computational annotation of sORF is almost impossible (82, 111, 112). It also impedes the functional prediction of sORFs. Also, due to the fact that long ORFs have longer sequences for comparison and the homology is more significant than sORFs, bioinformatic tools tend to collect long ORFs instead of sORFs, especially when long ORFs and sORFs are overlapping (3, 4, 111, 113).

Second, the genetic sequence databases and protein databases lack information about sORFs and small proteins. It is partly because of the limited knowledge about small proteins and sORFs. In addition, the sequence size cutoff during the sequence submission excludes small proteins and sORF sequences from many databases. For example, the genetic sequence database of National Institutes of Health refuses the submission of individual sequences of less than 200 nucleotides (183). The limited sequence information of small proteins and sORFs in databases impedes the bioinformatic prediction of sORFs by sequence comparison.

2.4.2 Challenges for proteomic approaches

MS/MS analysis provides information on protein sequences and direct evidence of protein expression, which is an advantage for novel protein identification. In addition, the MS/MS analysis also contains information about conditions that are suitable for protein expression. This provides a convenient method for novel protein validation and characterization. However, there are some challenges that limit the application of MS/MS for small protein identification.

First, the concentration of protein affects the credibility and reproducibility (114) of the MS/MS result. Compared with larger proteins, small proteins generally have lower expression levels. Their signals in mass spectra are, in many cases, as low as the noise level and thus overlooked by mass spectrometry.

Second, the number of enzymatic digestion sites affects the number of ionized peptides that can be captured by MS/MS, and thus affects the quality of MS/MS results. In general, a protein will be identified as an authentic protein only if more than two peptides can be captured. Due to the size and limited number of trypsin-digestion sites of small proteins, this number is often set to one to improve the sensitivity on small protein signal capture. However, this raises the risk of false positives and undermines the reproducibility of the results.

Third, the MS/MS data analysis relies on protein databases. However, due to the limited knowledge about small proteins in bacteria, few small protein sequences are included in protein databases. The general absence of sORFs from conventional protein databases has the consequence that small proteins are likely to escape from detection even if their signals can be captured by mass spectrometry. Some studies established custom database for small protein analysis. For example, a study in *Pseudomonas fluorescens* established a custom database by separating genome sequence from stop codon to stop codon and translating the sequences into amino acid sequences (85). However, the collection of

sequences without considering the coding/non-coding regions increases the risk for false positives and complicates the data analysis.

Chapter 3 Quantitative proteomics analysis and adaptation of *Pseudomonas putida* to different environments

3.1 Adaptation of *P. putida* to different environments

Members of the *Pseudomonas* genus exist in variety of environments. A major characteristic of *Pseudomonas* species is metabolic versatility and the ability to adapt to different environments, including stress conditions induced by antibiotics and chemicals. As one of the most studied models in the genus, *P. putida* widely exists in soil, water and organic polluted environments. In addition, *P. putida* strains can also establish a beneficial relationships with plants (115, 116). (Figure 3)

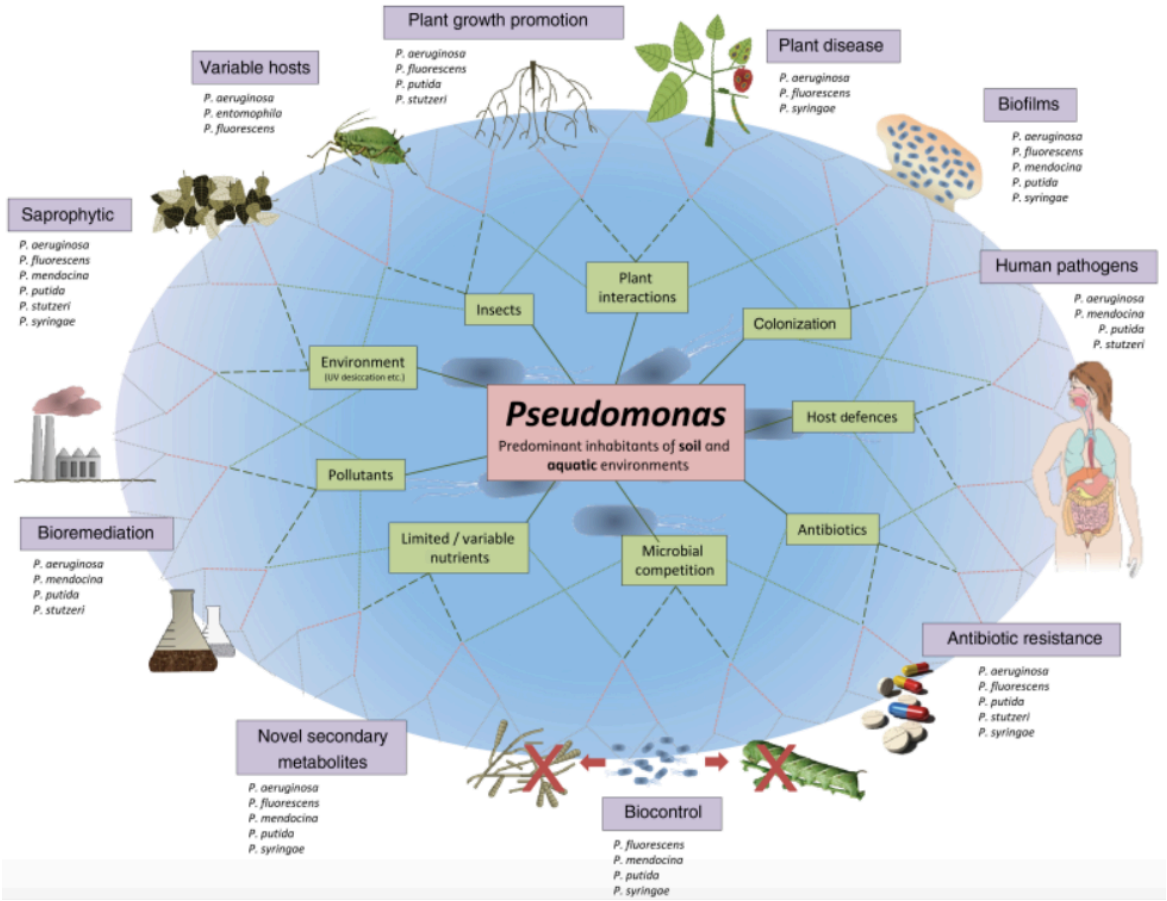


Figure 3: The functional and environmental range of *Pseudomonas* spp. (117). (Figure adapted from reference 117)

3.1.1 Stress resistance of *P. putida*

P. putida is able to tolerate stress conditions, including antibiotics (118–122), organic solvents (123–127), oxidative stress (128, 129), osmotic stress (130–132), as well as variations of pH and temperature (120). Many studies have focused on revealing the response of *P. putida* to these stress conditions. A proteomic study about the response of *P. putida* to tetracycline stress demonstrated that components of ABC transporters were up-regulated under tetracycline stress, indicating the role of ABC transporter system in tetracycline stress response (118). Oxidative stress generates reactive oxygen species (ROS) and leads to damage for all the components of cells (133). In response, a group of stress-regulators such as FinR, Fpr, as well as detoxification enzymes KatA, AhpC etc. were induced(134–142). The oxidative stress also affected metal homeostasis, indicating the involvement of iron in ROS elimination. The research on resistance of *P. putida* on organic compounds demonstrated that aromatic compounds induced a similar response in *P. putida*. Under this stress condition, efflux pumps were activated and chaperones that were involved in the ROS response were also activated (143). Study on phenol-induced stress (144) showed high expression of ABC transporters, a group of chaperones and antioxidants, indicating that many proteins were involved in the response against different stresses. Research focusing on the response of bacteria for acid environment demonstrated the differential protein expression in cell division and electron transport processes. The expression level of group of regulators was also affected in acid environment (145).

3.1.2 Adaptation of *P. putida* to different nutrient environments

With versatile metabolism, *P. putida* is able to assimilate a wide range of compounds and survive with a wide spectrum of nutrients. Research about the carbon source assimilation indicated that *P.putida* selectively used amino acids as carbon source when glucose and amino acids were both abundant in the environment(146). Another analysis demonstrated that *P.putida* consumed succinate prior to glucose when both compounds were present in the medium (147), indicating that organic acids and amino acids rather than glucose were the optimal carbon sources for *P. putida*. The Crc regulator that was first found in *P.*

aeruginosa (148), was later validated in *P. putida* (149–151). This regulator controls the carbon source selection of the bacterium.

3.2 Quantitative proteomic MS approaches

Protein expression is deeply involved in the interaction between bacteria and the environment. On the one hand, the expression of many proteins is largely affected by environment. On the other hand, the activation of certain biological pathways and protein expression reduces detrimental effects caused by toxic factors in the environment and guarantees the survival of bacteria in harmful environments. Better understanding of the differential protein expression of bacteria on the proteome level is meaningful for uncovering the principles of interaction between bacteria and the environment.

Due to the fact that the proteome provides the most comprehensive information of total protein expression in an organism, mass spectrometry-based proteomic techniques were designed for testing the differential expression of the proteome among multiple samples. With these techniques, the total proteins obtained from different samples are analyzed by MS/MS. This technique has been used for the analysis of adaptability of *P.putida* in many toxic environments, such as in the presence of antibiotics and organic reagents, as well as other environmental factors, such as nutrient sources and temperature (118, 144, 152, 153).

The quantitative proteomic analysis is established based on comparable MS/MS data from more than one sample. Depending on the proteomic approach and methods used for sample preparation, the MS/MS data can be obtained in two ways. First, the data can be obtained by isotope labeling quantitative MS/MS. With this method, different samples are mixed and used for MS/MS. The peptides from different samples are separated by differentially labeled tags. Second, the MS/MS data can also be obtained via a label-free method. In this method, each sample is independently analyzed by MS/MS and the protein expression among different samples is compared by a complex data analysis approach (154). In the former case, the signals of peptides from different samples need to be

identified and distinguished by MS/MS. Second, mass spectra and signals of peptides obtained from different samples need to be comparable. This requires data normalization during the MS/MS data analysis (Figure 3).

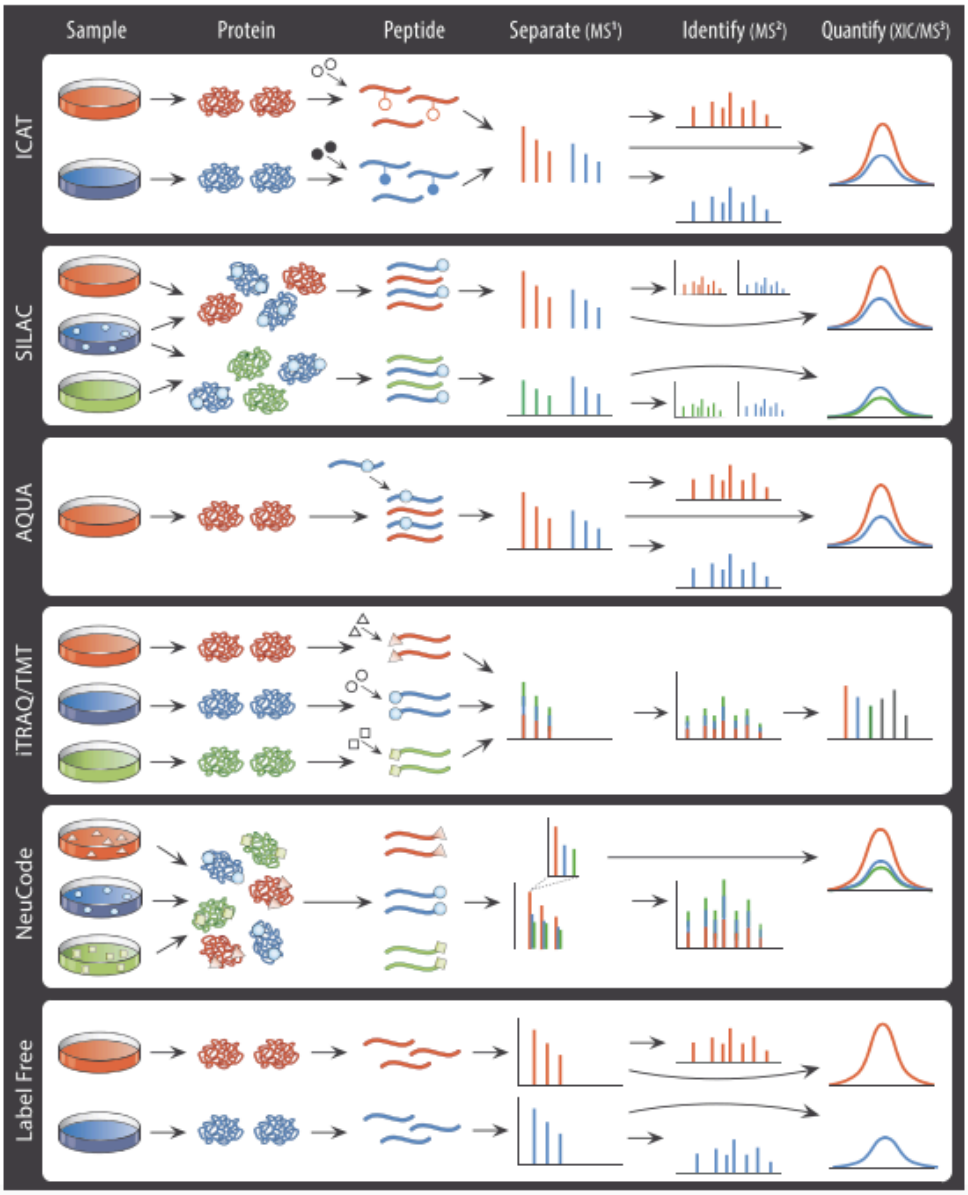


Figure 3: Quantitative analysis of labeling and label-free proteomic mass spectrometry techniques (154)**.** (Figure adapted from reference 154) *The figure consists of workflows of six techniques. For SILAC and NeuCode techniques, labels are introduced into proteins during the sample growth. With iTRAQ/TMT and ICAT techniques, samples are labeled after protein digestion. The AQUA is adopted for absolute quantification. The label free technique is used for relative quantification. MS^1 is the mass spectrometry for peptide separation. MS^2 is the mass spectrometry for peptide sequence identification. The extracted-ion chromatogram (XIC) is a figure combined with chromatography and MS^1 signals. It reflects the ion intensity at each time point during the chromatography retention time.*

3.2.1 Stable isotope labeling quantitative MS

The MS/MS data analysis is based on the mass to charge ratio (m/z) of proteins or peptides. Due to the fact that each protein or peptide has its own m/z value on mass spectra, in general, the MS signal of the same peptide originating from different samples is overlapped and cannot be distinguished by MS/MS. One solution is to introduce labels that can change the m/z value of peptides and create the identifiable changes of MS signals that can be used to distinguish the peptides from different samples. Stable isotopes, such as ^{13}C, ^{15}N, ^{18}O, and ^{2}H, can be introduced to proteins by many approaches. Most of the approaches are based on either metabolic labeling or chemical labeling strategies.

Metabolic labeling

The metabolic labeling strategy is established according to the protein metabolism of organism. The isotope-labeled amino acids are used for feeding the organism and are incorporated into every protein of the organism by protein synthesis. One of the most popular approaches of metabolic labeling is SILAC (stable isotope labeling with amino acids in culture) (155). With this method, isotope–labeled arginine and lysine are used for organism cultivation. The cells with and without labeling are mixed together for protein extraction, the mixed protein samples are loaded on the MS/MS for analysis. The intensities of labeled and unlabeled peptides in mass spectra are used for relative quantification analysis. (Figure 3) The major advantage of this strategy is reproducibility of the results.

One of the problems of the SILAC approach is that unlabeled amino acids coming from the degradation of proteins are recycled into new rounds of protein synthesis and lead to the underestimation of protein turnover rates (156, 157). The underestimated turnover rates may distort the ratio measured between heavy and light sample, and thus significantly affect the quantification result. A variant SILAC approach, pulse SILAC (158, 159), was designed for measuring the synthesis and degradation rates of proteins,

and correcting the deviation of the protein turnover rate. In brief, the organism is first cultivated in normal medium and then incubated with heavy-isotope-labeled amino acids for a certain time. The protein synthesis and degradation rates can be measured by comparing the MS signal ratio of labeled to unlabeled proteins of the same sample.

Another disadvantage of the classical SILAC approach is the limitation on the number of samples that can be analyzed within one reaction. In general, the maximum number of samples is three. This is because of the limited repertoire of labeled amino acids that can be used for isotope labeling (160). Extending SILAC to more labels requires the involvement of deuterated amino acids. However, the incompatibility of LC with deuterated amino acids separation hampers the accuracy of quantification (161). In addition, more samples increase the possibility of overlapping isotope clusters, which affect the quantification. A recent study introduced a new SILAC approach called Neutron-encode SILCA (NeuCode) for multiple-sample analysis (162–164). This approach was established based on the detection of slight mass changes resulting from the isotopic shift for ^{15}N and ^{13}C in the stable-isotope tag. The concomitant swapping of $^{1}H/^{2}H$, $^{16}O/^{18}O$, and $^{32}S/^{34}S$ can also be used for designing new tags. A mass spectrometer with high mass resolution is necessary for detecting the slight mass changes that result from isotopic shift. With this approach, 39 isotopologues were synthesized for multiplexing samples labeling. This approach enables high levels of multiplexing with high accuracy (164).

Chemical labeling

Another popular strategy is to label the peptide samples with chemical reactions. Different from metabolic labeling, the digested protein sample instead of living organism is used for isotopic labeling. In order to introduce labels to proteins, a series of isotope-labeled tags are designed. In addition to isotope-labeling regions, these tags contain reactive groups that target proteins. The peptide is labeled by chemical reaction instead of metabolism.

The major challenge of chemical labeling approaches is that the differences in sample preparation are taken into account, leading to variations of results among biological repeats (165).

The isotope-coded affinity tag (ICAT) is the first tag that was designed for quantitative MS/MS (166). The tag contains a reactive group for amino acid side chain labeling, isotope-labeled linker, and an affinity tag for labeled protein/peptide isolation. In practice, the side chains of cysteine residues are labeled with the "light" and "heavy" ICAT reagents. The quantification of cysteine-containing peptides is analyzed directly via the first MS spectra (Figure 4). The major weakness of this method is that the quantification analysis is limited to the cysteine-containing peptides. The proteins that contain few or no cysteine residues cannot be analyzed with the ICAT approach.

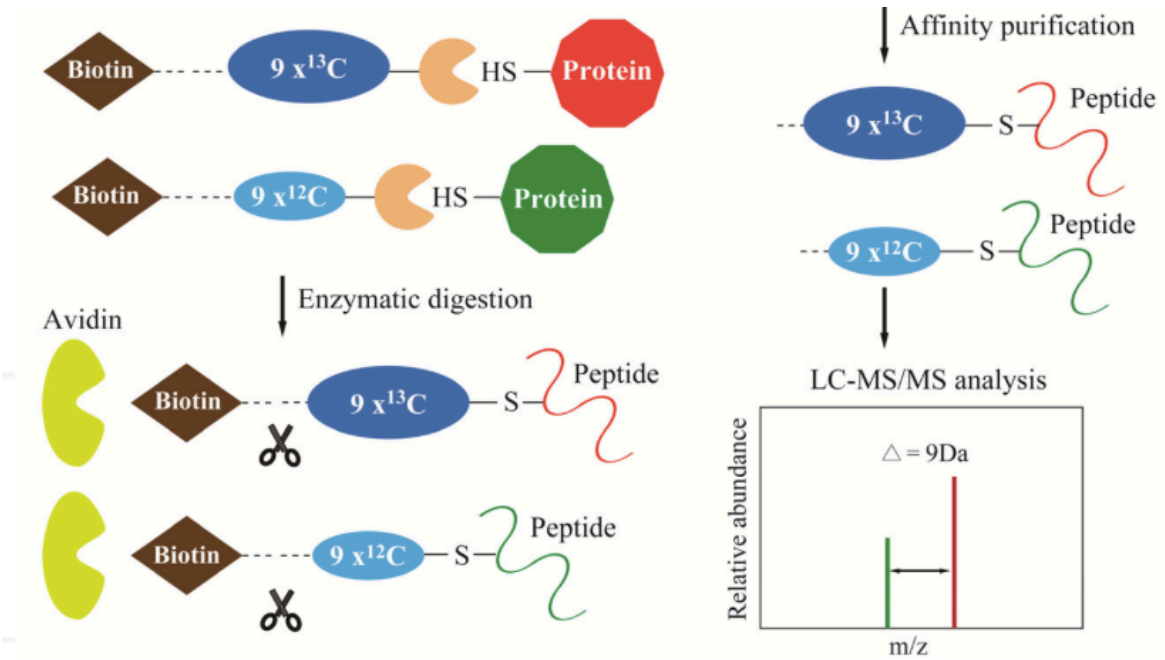


Figure 4: ICAT tags and quantitative proteomics analysis (167). (Figure adapted from reference 167) *The ICAT tags contain reactive groups that interact with side chains of amino acids (on the left), isotope labeled linker (linker in the middle) and biotin that is used for affinity purification of labeled proteins. The biotin is removed from the tags before LC-MS/MS. The signals of "light" and "heavy" labeled peptides are separated from each other on mass spectrum.*

The isobaric mass tags solved the weakness of ICAT and make it possible for multiplexing analysis. Two popular isobaric mass tags are isobaric tags for relative and absolute quantitation (iTRAQ) and tandem mass tag (TMT) (168–170). These tags consist of isotope-encoded reporter groups for quantitative analysis, balance groups for unifying the mass of different tag reagents and amine specific peptide reactive groups (NHS) for linking with peptide (Figure 5). In practice, the isobaric tags are incorporated at the N-terminus and side-chain amines of peptides. Due to the equal mass of tag reagents, differentially labeled proteins do not have differences in mass and are not distinguished by sample fractionation. This design reduces the possibility of signal overlap in LC or MS[1], which increases the credibility of analysis and capability of multiplexing analysis. The signals of peptides from different samples are distinguished during the MS/MS when fragmentation within the mass spectrometry releases the isotope-encoded reporters. The quantification is achieved based on the quantitative comparison among reporters from different samples.

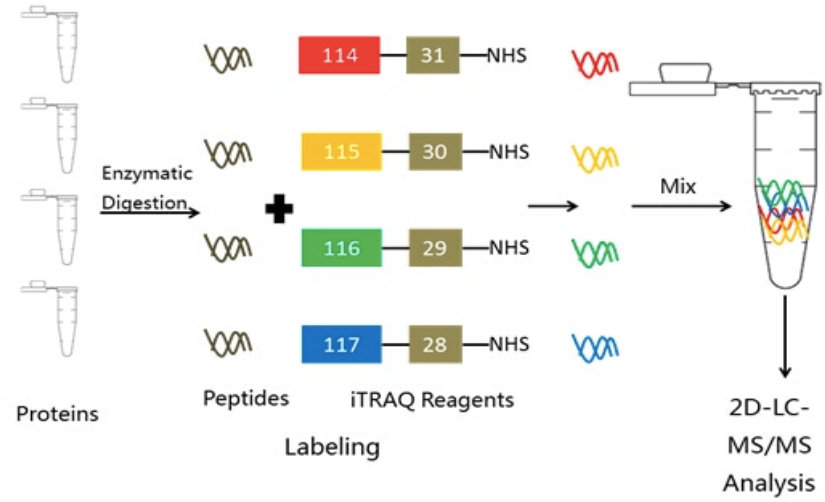


Figure 5: The iTRAQ/TMT labeling system (184). (Figure adapted from reference 184) *The TMT tags contain isotope-encoded reporter groups for quantitative analysis (the left part with numbers from 114 to 117), balance groups for unifying the mass of different tag reagents (in the middle with numbers from 31 to 28) and amine specific peptide reactive groups (NHS). The relative quantification of peptides from different samples is calculated by comparing the levels of each isotope-encoded reporters in the samples.*

3.2.2 Label free MS/MS

The label free MS/MS is established on the improvement of instrument, MS/MS data acquisition and analysis (171, 172). Due to the fact that the label free approach does not require the labeling treatment of protein samples, the experimental procedure is simpler than labeling approaches. However, it requires very high reproducibility of the results and the high replicate numbers of reads needed for good statistics analysis. The methods that are used for data analysis are also more complex than labeling proteomics (173–176). With this method, all protein samples without labeling are separately analyzed by LC-MS/MS. The quantification among samples is analyzed either by spectral counting of proteins (at MS^2 level) or measurement of peptide intensities in mass spectrometry (at MS^1 level).

In the former case (177–179), the number of peptide-to-spectrum matches (PSMs) for each protein is calculated. The relative quantification of protein abundance is established on the comparison of PSMs of each protein in different samples. The major advantage is that the analysis is very straightforward and only counts the number of PSMs. The major disadvantage is that the number of PSMs is largely dependent on the size of proteins and at same time, is easily influenced by experimental parameters. This indicates the variability of quantification, especially for short proteins.

In the later case, the ion intensities of peptides are evaluated based on signals of Mass spectra on MS^1 (180–182). The signal intensities of each peptide in each sample are used for comparison and quantification. The intensity approach requires the high accuracy mass spectrometry to eliminate the interfering signals of similar but distinct mass.

Chapter 4. Concluding remarks and perspective

P. putida is an appropriate bacterium for compound production. Increasing the knowledge of *P. putida* is meaningful for its broader application in industry. This thesis focuses on a group of understudied proteins, s-proteins, in the bacterium and for the first time validates the expression of 14 s-proteins in *P. putida*. In addition, a quantitative proteomic approach is adopted for protein analysis at the proteome level. The global analysis of protein expression under different stress and nutrient conditions increases our knowledge about the response of the bacterium to different conditions.

Three articles are included in this thesis. **Article 1** focuses on novel small protein identification and validation. By combining bioinformatic prediction, proteomics and transcriptomics approaches, 186 putative small open reading frames encoding proteins of 50 amino acids or less were identified. Based on the plasmid-based system and affinity tag specific for small protein labeling, expression of 14 novel small proteins was validated by western blot. Based on the orthology and synteny analysis, the relationship between novel sORF and their neigbour gene were conserved in most cases. This gave the clue about the possible function of validated sORFs.

Article 2 focuses on the proteomic analysis of *P. putida* under different conditions. In this article, a quantitative proteomic MS approach was used to analysis the protein expression level under H_2O_2, imipenem stress, as well as LB and M9+glucose medium. By comparing the proteome of stress-treated sample with M9 medium, lists of genes that were response to stress condition were analyzed. The proteomic data showed that some proteins, such as ABC transporters, detoxification proteins, and DNA repair proteins are involved in the response for both H_2O_2 and imipenem stresses. The components of flagella were exclusively up-regulated under imipenem stress and some detoxification protein, KatA and ahpF were specifically up-reguated under H_2O_2 stresses. The different nutrient environments great impacted protein and amino acid synthesis, energy metabolism and fatty acid metabolism.

Article 3 focuses on a new technique, ribosome profiling, and its application to novel ORF identification. Ribosome profiling analyzes the ribosome binding regions on mRNA sequences, which can be used for ORF identification. The ribosome footprint cDNA libraries were established from 1 stress condition, imipenem stress, and two growth media, LB and M9+glucose. The cDNA sequencing technique was used for ribosome footprint sequencing. The ribosome footprints were grouped by the overlap of RNA-seq reads. A custom-made ORF prediction approach was used to predict the ORFs based on grouped footprint sequences within intergenic regions, and more than 1200 putative ORFs with larger than 47 bp were collected.

The technical improvements in genome and protein analysis have had a great impact on biological research. However, these improvements on bioinformatics, proteomics, transcriptomics methods still have their own disadvantage on sORF and s-protein analysis. The combination of multiple approaches would increase the capability for s-protein analysis. In this thesis, tens of s-proteins were successfully validated, however, the challenge is the functional analysis of this group of proteins. The major obstacle is the small size of the proteins that makes homology analysis almost impossible. Based on our knowledge with this group of novel small proteins, the next step of the research will be the characterization of some of novel small proteins that may be involved in certain functional operons. Another goal is a more in-depth analysis of the putative ORFs obtained from ribosome profiling.

Both **Article 2** and **Article 3** are in preparation. Especially for the **Article 3**, the result has only been preliminary discussed. More work is necessary to fulfill the articles into publication. For the **Article 2**, our future work will focus on the improvement of language and organization of text. More details about differential expressed proteins will be added to make the text easier to understand. In addition, the quantitative transcriptomics data will be introduced and compared with MS data. Issues, such as Post-transcriptional regulation, will be discussed in the future.

For the **Article 3**, the first work we will do is to check the credibility of the putative ORFs located in intergenic regions. The work will include the homology analysis and genome context analysis. Second, the ribosome footprints that are mapped with annotated gene will be analyzed. The result will be used for the analysis on three issues: 1) leaderless transcription; 2) Pause of ribosome on mRNA during translation; 3). Alternative ORFs within annotated genes.

References

1. Nikel PI, Martínez-García E, de Lorenzo V (2014) Biotechnological domestication of pseudomonads using synthetic biology. *Nat Rev Microbiol* 12(5):368–79.
2. Godoy P, Ramos-González MI, Ramos JL (2001) Involvement of the TonB system in tolerance to solvents and drugs in Pseudomonas putida DOT-T1E. *J Bacteriol* 183(18):5285–5292.
3. Salzberg SL, Delcher AL, Kasif S, White O (1998) Microbial gene identification using interpolated Markov models. 26(2):544–548.
4. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* 27(23):4636–4641.
5. Storz G, Wolf YI, Ramamurthi KS (2014) Small proteins can no longer be ignored. *Annu Rev Biochem* 83:753–77.
6. Sandman K, Losick R, Youngman P (1987) Genetic analysis of Bacillus subtilis spo mutations generated by Tn917-mediated insertional mutagenesis. *Genetics* 117(4):603–617.
7. Levin PA, et al. (1993) An unusually small gene required for sporulation by Bacillus subtilis. *Mol Microbiol* 9(4):761–771.
8. Van Ooij C, Losick R (2003) Subcellular localization of a small sporulation protein in Bacillus subtilis. *J Bacteriol* 185(4):1391–1398.
9. Galperin MY, et al. (2012) Genomic determinants of sporulation in Bacilli and Clostridia: Towards the minimal set of sporulation-specific genes. *Environ Microbiol* 14(11):2870–2890.
10. Ebmeier SE, Tan IS, Clapham KR, Ramamurthi KS (2012) Small proteins link coat and cortex assembly during sporulation in Bacillus subtilis. *Mol Microbiol* 84(4):682–696.
11. Schmalisch M, et al. (2010) Small genes under sporulation control in the Bacillus subtilis genome. *J Bacteriol* 192(20):5402–5412.
12. Handler AA, Lim JE, Losick R (2008) Peptide inhibitor of cytokinesis during sporulation in Bacillus subtilis. *Mol Microbiol* 68(3):588–599.
13. Modell JW, Hopkins AC, Laub MT (2011) A DNA damage checkpoint in Caulobacter crescentus inhibits cell division through a direct interaction with FtsW. *Genes Dev* 25(12):1328–43.
14. Mohammadi T, et al. (2011) Identification of FtsW as a transporter of lipid-linked cell wall precursors across the membrane. *EMBO J* 30(8):1425–1432.
15. Wong RSY, McMurry LM, Levy SB (2000) "Intergenic" blr gene in Escherichia coli encodes a 41-residue membrane protein affecting intrinsic susceptibility to certain inhibitors of peptidoglycan synthesis. *Mol Microbiol* 37(2):364–370.
16. Karimova G, Davi M, Ladant D (2012) The Beta-lactam resistance protein Blr, a small membrane polypeptide, is a component of the Escherichia coli cell division machinery. *J Bacteriol* 194(20):5576–5588.
17. Hobbs EC, Astarita JL, Storz G (2010) Small RNAs and small proteins involved in resistance to cell envelope stress and acid shock in Escherichia coli: analysis of a bar-coded mutant collection. *J Bacteriol* 192(1):59–67.
18. Hobbs EC, Yin X, Paul BJ, Astarita JL, Storz G (2012) Conserved small protein

associates with the multidrug efflux pump AcrB and differentially affects antibiotic resistance. *Proc Natl Acad Sci U S A* 109(41):16696–701.
19. Du D, et al. (2014) Structure of the AcrAB-TolC multidrug efflux pump. *Nature* 509(7501):512–5.
20. VanOrsdel CE, et al. (2013) The Escherichia coli CydX protein is a member of the CydAB cytochrome bd oxidase complex and is required for cytochrome bd oxidase activity. *J Bacteriol* 195(16):3640–50.
21. Sun Y-H, et al. (2012) The small protein CydX is required for function of cytochrome bd oxidase in Brucella abortus. *Front Cell Infect Microbiol* 2:47.
22. Burkholder WF, Kurtser I, Grossman AD (2001) Replication initiation proteins regulate a developmental checkpoint in Bacillus subtilis. *Cell* 104(2):269–279.
23. Cunningham KA, Burkholder WF (2009) The histidine kinase inhibitor Sda binds near the site of autophosphorylation and may sterically hinder autophosphorylation and phosphotransfer to Spo0F. *Mol Microbiol* 71(3):659–677.
24. Lippa AM, Goulian M (2012) Perturbation of the oxidizing environment of the periplasm stimulates the PhoQ/PhoP system in Escherichia coli. *J Bacteriol* 194(6):1457–1463.
25. Moon K, Gottesman S (2009) A PhoQ/P-regulated small RNA regulates sensitivity of Escherichia coli to antimicrobial peptides. *Mol Microbiol* 74(6):1314–1330.
26. Miyashiro T, Goulian M (2007) Stimulus-dependent differential regulation in the Escherichia coli PhoQ PhoP system. *Proc Natl Acad Sci U S A* 104(41):16305–16310.
27. Lippa AM, Goulian M (2009) Feedback inhibition in the PhoQ/PhoP signaling system by a membrane peptide. *PLoS Genet* 5(12). doi:10.1371/journal.pgen.1000788.
28. Gaballa A, et al. (2008) The Bacillus subtilis iron-sparing response is mediated by a Fur-regulated small RNA and three small, basic proteins. *Proc Natl Acad Sci U S A* 105(33):11927–11932.
29. Baichoo N, Wang T, Ye R, Helmann JD (2002) Global analysis of the Bacillus subtilis fur regulon and the iron starvation stimulon. *Mol Microbiol* 45(6):1613–1629.
30. Chen L, James LP, Helmann JD (1993) Metalloregulation in Bacillus subtilis: Isolation and characterization of two genes differentially repressed by metal ions. *J Bacteriol* 175(17):5428–5437.
31. Waters LS, Sandoval M, Storz G (2011) The Escherichia coli MntR miniregulon includes genes encoding a small protein and an efflux pump required for manganese homeostasis. *J Bacteriol* 193(21):5887–5897.
32. Martin JE, Waters LS, Storz G, Imlay JA (2015) The Escherichia coli Small Protein MntS and Exporter MntP Optimize the Intracellular Concentration of Manganese. *PLoS Genet* 11(3). doi:10.1371/journal.pgen.1004977.
33. Allen RJ, et al. (2014) Conservation analysis of the CydX protein yields insights into small protein identification and evolution. *BMC Genomics* 15(1):946.
34. Kato A, Chen HD, Latifi T, Groisman EA (2012) Reciprocal Control between a Bacterium's Regulatory System and the Modification Status of Its Lipopolysaccharide. *Mol Cell* 47(6):897–908.
35. Alix E, Blanc-Potard A-B (2009) Hydrophobic peptides: novel regulators within

bacterial membrane. *Mol Microbiol* 72(1):5–11.
36. Alix E, Blanc-Potard A-B (2008) Peptide-assisted degradation of the Salmonella MgtC virulence factor. *EMBO J* 27(3):546–557.
37. Moncrief MBC, Maguire ME (1998) Magnesium and the role of mgtC in growth of Salmonella typhimurium. *Infect Immun* 66(8):3802–3809.
38. Cardenal-Muñoz E, Ramos-Morales F (2013) DsbA and MgrB regulate steA expression through the two- component system PhoQ/PhoP in salmonella enterica. *J Bacteriol* 195(10):2368–2378.
39. Whitten AE, et al. (2007) The Structure of the KinA-Sda Complex Suggests an Allosteric Mechanism of Histidine Kinase Inhibition. *J Mol Biol* 368(2):407–420.
40. Gaßel M, Möllenkamp T, Puppe W, Altendorf K (1999) The KdpF subunit is part of the K+-translocating Kdp complex of Escherichia coli and is responsible for stabilization of the complex in vitro. *J Biol Chem* 274(53):37901–37907.
41. Gannoun-Zaki L, Alibaud L, Carrère-Kremer S, Kremer L, Blanc-Potard AB (2013) Overexpression of the KdpF Membrane Peptide in Mycobacterium bovis BCG Results in Reduced Intramacrophage Growth and Altered Cording Morphology. *PLoS One* 8(4). doi:10.1371/journal.pone.0060379.
42. Vanderpool CK, Gottesman S (2004) Involvement of a novel transcriptional activator and small RNA in post-transcriptional regulation of the glucose phosphoenolpyruvate phosphotransferase system. *Mol Microbiol* 54(4):1076–1089.
43. Wadler CS, Vanderpool CK (2007) A dual function for a bacterial small RNA: SgrS performs base pairing-dependent regulation and encodes a functional polypeptide. *Proc Natl Acad Sci* 104(51):20454–20459.
44. Shine J, Dalgarno L (1974) The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci U S A* 71(4):1342–6.
45. Ma J, Campbell A, Karlin S (2002) Correlations between Shine-Dalgarno Sequences and Gene Features and Operon Structures. 184(20):5733–5745.
46. Barendt PA, Shah NA, Barendt GA, Kothari PA, Sarkar CA (2013) Evidence for context-dependent complementarity of non-shine-dalgarno ribosome binding sites to *Escherichia coli* rRNA. *ACS Chem Biol* 8(5):958–966.
47. Barendt PA, Shah NA, Barendt GA, Sarkar CA (2012) Broad-specificity mrna-rrna complementarity in efficient protein translation. *PLoS Genet* 8(3). doi:10.1371/journal.pgen.1002598.
48. Li G-W, Oh E, Weissman JS (2012) The anti-Shine–Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* 484(7395):538–541.
49. Vogl C, Futschik A (2010) Hidden Markov models in biology. *Methods Mol Biol* 609:241–253.
50. Delcher AL, Bratke KA, Powers EC, Salzberg SL (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. 23(6):673–679.
51. Schiex T, Gouzy J, Moisan A, de Oliveira Y (2003) FrameD: A flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences. *Nucleic Acids Res* 31(13):3738–3741.
52. Besemer J, Lomsadze A, Borodovsky M (2001) GeneMarkS : a self-training method for prediction of gene starts in microbial genomes . Implications for finding sequence motifs in regulatory regions. 29(12):2607–2618.

53. Zhu W, Lomsadze A, Borodovsky M (2010) Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res* 38(12). doi:10.1093/nar/gkq275.
54. Antonov I, Borodovsky M (2010) Genetack: frameshift identification in protein-coding sequences by the Viterbi algorithm. *J Bioinform Comput Biol* 8(3):535–551.
55. Antonov I, Baranov P, Borodovsky M (2013) GeneTack database: Genes with frameshifts in prokaryotic genomes and eukaryotic mRNA sequences. *Nucleic Acids Res* 41(D1). doi:10.1093/nar/gks1062.
56. Zhu H, Hu G-Q, Yang Y-F, Wang J, She Z-S (2007) MED: a new non-supervised gene prediction algorithm for bacterial and archaeal genomes. *BMC Bioinformatics* 8:97.
57. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL (1999) Improved microbial gene identification with G LIMMER. 27(23):4636–4641.
58. Delcher AL, Bratke KA, Powers EC, Salzberg SL (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 23(6):673–679.
59. Sallet E, Gouzy J, Schiex T (2014) EuGene-PP: A next-generation automated annotation pipeline for prokaryotic genomes. *Bioinformatics* 30(18):2659–2661.
60. Sallet E, et al. (2013) Next-generation annotation of prokaryotic genomes with EuGene-P: Application to sinorhizobium meliloti 2011. *DNA Res* 20(4):339–353.
61. Bocs S, Cruveiller S, Vallenet D, Nuel G, Médigue C (2003) AMIGene: Annotation of MIcrobial Genes. *Nucleic Acids Res* 31(13):3723–3726.
62. Hyatt D, et al. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119.
63. Noguchi H, Taniguchi T, Itoh T (2008) Meta gene annotator: Detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res* 15(6):387–396.
64. Noguchi H, Park J, Takagi T (2006) MetaGene: Prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res* 34(19):5623–5630.
65. Besemer J, Borodovsky M (1999) Heuristic approach to deriving models for gene finding. *Nucleic Acids Res* 27(19):3911–3920.
66. Ã PN, Krogh A (2005) Genome analysis Large-scale prokaryotic gene prediction and comparison to genome annotation. 21(24):4322–4329.
67. Larsen TS, Krogh A (2003) EasyGene--a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics* 4(1):21.
68. Frishman D, Mironov A, Mewes HW, Gelfand M (1998) Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res* 26(12):2941–2947.
69. Krogh A, Mian IS, Haussler D (1994) A hidden Markov model that finds genes in E. coli DNA. *Nucleic Acids Res* 22(22):4768–78.
70. Nicolas P, et al. (2002) Mining Bacillus subtilis chromosome heterogeneities using hidden Markov models. *Nucleic Acids Res* 30(6):1418–1426.
71. Blattner FR, et al. (1997) The Complete Genome Sequence of Escherichia coli K-12. *Sci Mag* 277(September):1453–1462.
72. Rudd KE (1998) Linkage map of Escherichia coli K-12, edition 10: the physical map. *Microbiol Mol Biol Rev* 62(3):985–1019.
73. Hemm MR, Paul BJ, Schneider TD, Storz G, Rudd KE (2008) Small membrane proteins found by comparative genomics and ribosome binding site models. *Mol*

Microbiol 70(6):1487–1501.
74. Hemm MR, et al. (2010) Small stress response proteins in Escherichia coli: proteins missed by classical proteomic studies. *J Bacteriol* 192(1):46–58.
75. Ibrahim M, et al. (2007) A genome-wide survey of short coding sequences in streptococci. *Microbiology* 153(Pt 11):3631–44.
76. Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. *Nature* 422(6928):198–207.
77. Ho CS, et al. (2003) Electrospray ionisation mass spectrometry: principles and clinical applications. *Clin Biochem* 24(1):3–12.
78. Karas M, Kr??ger R (2003) Ion formation in MALDI: The cluster ionization mechanism. *Chem Rev* 103(2):427–439.
79. Knochenmuss R (2006) Ion formation mechanisms in UV-MALDI. *Analyst* 131(9):966–986.
80. Wilkins MR, et al. (1996) From proteins to proteomes: large scale protein identification by two- dimensional electrophoresis and amino acid analysis. *Biotechnol (N Y)* 14(1):61–65.
81. Han M-J, Lee SY (2006) *The Escherichia coli proteome: past, present, and future prospects.* doi:10.1128/MMBR.00036-05.
82. Rudd KE, Humphery-Smith I, Wasinger VC, Bairoch A (1998) Low molecular weight proteins: A challenge for post-genomic research. *Electrophoresis* 19(4):536–544.
83. Fountoulakis M, Takács MF, Berndt P, Langen H, Takács B (1999) Enrichment of low abundance proteins of Escherichia coli by hydroxyapatite chromatography. *Electrophoresis*, pp 2181–2195.
84. Lluch-senar M, et al. (2015) Defining a minimal cell : essentiality of small ORFs and ncRNAs in a genome-reduced bacterium. 1–7.
85. Kim W, et al. (2009) Proteomic detection of non-annotated protein-coding genes in Pseudomonas fluorescens Pf0-1. *PLoS One* 4(12):e8455.
86. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324(5924):218–23.
87. Ingolia NT (2010) *Genome-wide translational profiling by Ribosome footprinting* (Elsevier Inc.). 2nd Ed. doi:10.1016/S0076-6879(10)70006-9.
88. Latif H, et al. (2015) A streamlined ribosome profiling protocol for the characterization of microorganisms. *Biotechniques* 58(6):329–32.
89. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS (2012) The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc* 7(8):1534–50.
90. Freeberg L, Kuersten S, Syed F (2013) Isolate and sequence ribosome-protected mRNA fragments using size-exclusion chromatography. *Nat Publ Gr* 10(5):i–ii.
91. Brar G a, Weissman JS (2015) Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat Rev Mol Cell Biol* 16(11):651–664.
92. Woolstenhulme CJ, Guydosh NR, Green R, Buskirk AR (2015) High-Precision analysis of translational pausing by ribosome profiling in bacteria lacking EFP. *Cell Rep* 11(1):13–21.
93. Kannan K, et al. (2014) The general mode of translation inhibition by macrolide

antibiotics. *Proc Natl Acad Sci U S A* 111(45):1–6.
94. Li GW, Burkhardt D, Gross C, Weissman JS (2014) Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell* 157(3):624–635.
95. Ingolia NT, Lareau LF, Weissman JS (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147(4):789–802.
96. Stern-Ginossar N, et al. (2012) Decoding human cytomegalovirus. *Science (80-)* 338(6110):1088–1093.
97. Brar G a., et al. (2012) High-Resolution View of the Yeast Meiotic Program Revealed by Ribosome Profiling. *Science (80-)* 335(6068):552–557.
98. Bazzini A a, et al. (2014) Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J* 33(9):981–93.
99. Aspden JL, et al. (2014) Extensive translation of small ORFs revealed by Poly-Ribo-Seq. *Elife*:e03528.
100. Smith JE, et al. (2014) Translation of Small Open Reading Frames within Unannotated RNA Transcripts in Saccharomyces cerevisiae. *Cell Rep* 7(6):1858–1866.
101. Vilela C, McCarthy JEG (2003) Regulation of fungal gene expression via short open reading frames in the mRNA 5???untranslated region. *Mol Microbiol* 49(4):859–867.
102. Schrader JM, et al. (2014) The Coding and Noncoding Architecture of the Caulobacter crescentus Genome. *PLoS Genet* 10(7). doi:10.1371/journal.pgen.1004463.
103. Shell SS, et al. (2015) Leaderless Transcripts and Small Proteins Are Common Features of the Mycobacterial Translational Landscape. 1–31.
104. Wang J, et al. (2015) Identification of bacterial sRNA regulatory targets using ribosome profiling. *Nucleic Acids Res*:gkv1158–.
105. Liu X, Jiang H, Gu Z, Roberts JW (2013) High-resolution view of bacteriophage lambda gene expression by ribosome profiling. *Proc Natl Acad Sci U S A* 110(29):11928–33.
106. Nelson KE, et al. (2002) Complete genome sequence and comparative analysis of the metabolically versatile Pseudomonas putida KT2440. 4:799–808.
107. Sonnhammer ELL, Eddy SR, Birney E, Bateman A, Durbin R (1998) Pfam: Multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* 26(1):320–322.
108. Sonnhammer ELL, Eddy SR, Durbin R (1997) Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins Struct Funct Genet* 28(3):405–420.
109. Quackenbush J, Liang F, Holt I, Pertea G, Upton J (2000) The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res* 28(1):141–5.
110. Belda E, Heck RGA Van, Lopez-sanchez MJ, Cruveiller S The revisited genome of Pseudomonas putida KT2440 enlightens its value as a robust metabolic chassis. 1–49.
111. Basrai MA, Hieter P, Boeke JD (1997) Small open reading frames: Beautiful

needles in the haystack. *Genome Res* 7(8):768–771.
112. Cliften PF, et al. (2001) Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res* 11(7):1175–1186.
113. Kastenmayer JP, et al. (2006) Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae. *Genome Res* 16(3):365–73.
114. Ma J, et al. (2014) Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. *J Proteome Res* 13(3):1757–65.
115. Espinosa-Urgel M, Salido A, Ramos JL (2000) Genetic analysis of functions involved in adhesion of Pseudomonas putida to seeds. *J Bacteriol* 182(9):2363–2369.
116. Degrassi G, et al. (2002) Plant growth-promoting Pseudomonas putida WCS358 produces and secretes four cyclic dipeptides: Cross-talk with quorum sensing bacterial sensors. *Curr Microbiol* 45(4):250–254.
117. Silby MW, Winstanley C, Godfrey SAC, Levy SB, Jackson RW (2011) Pseudomonas genomes: Diverse and adaptable. *FEMS Microbiol Rev* 35(4):652–680.
118. Yun S, et al. (2006) Proteome Analysis of Cellular Response of Pseudomonas putida KT2440 to Tetracycline Stress. 53:95–101.
119. Fernandez M, et al. (2012) Mechanisms of Resistance to Chloramphenicol in Pseudomonas putida KT2440. *Antimicrob Agents Chemother* 56(2):1001–1009.
120. Martins dos Santos VAP, Heim S, Moore ERB, Strätz M, Timmis KN (2004) Insights into the genomic basis of niche specificity of Pseudomonas putida KT2440. *Environ Microbiol* 6(12):1264–1286.
121. Kim SE, et al. (2012) Bacteremia: High Rates of Carbapenem Resistance and Mortality. *Chonnam Med J* 48(2):91.
122. Santos PM, Benndorf D, Sá-correia I (2004) Insights into Pseudomonas putida KT2440 response to phenol-induced stress by quantitative proteomics. 2640–2652.
123. Nakazawa T, Yokota T (1973) Benzoate Metabolism in Pseudomonas putida(arvilla) mt 2: Demonstration of Two Benzoate Pathways. *J Bacteriol* 115(1):262–267.
124. Dunn NW, Gunsalus IC (1973) Transmissible Plasmid Coding Early Enzymes of Naphthalene Oxidation in Pseudomonas putida. *J Bacteriol* 114(3):974–979.
125. Zhen D, et al. (2006) Plasmid-mediated degradation of 4-chloronitrobenzene by newly isolated Pseudomonas putida strain ZWL73. *Appl Microbiol Biotechnol* 72(4):797–803.
126. Dean HF, Cheevadhanarak S, Skurray RA, Bayly RC (1989) Characterisation of a degradative plasmid in Pseudomonas putida that controls the expression of 2, 4-xylenol degradative genes. *FEMS Microbiol* 61:153–157.
127. Herrmann H, Janke D, Krejsa S, Kunze I (1987) Involvement of the plasmid pPGH1 in the phenol degradation of Pseudomonas putida strain H. *FEMS Microbiol Lett* 43:133–137.
128. Geszvain K, McCarthy JK, Tebo BM (2013) Elimination of manganese(II,III) oxidation in Pseudomonas putida GB-1 by a double knockout of two putative multicopper oxidase genes. *Appl Environ Microbiol* 79(1):357–66.
129. Kim J, Park W (2014) Oxidative stress response in Pseudomonas putida. *Appl Microbiol Biotechnol* 98(16):6933–6946.

130. Kumar Deshwal V, Kumar P (2013) Effect of Salinity on Growth and PGPR Activity of Pseudomonads. *J Acad Ind Res* 2(6):353–356.
131. Castro AG De, Bredholt H, Strøm AR, Tunnacliffe A (2000) Anhydrobiotic Engineering of Gram-Negative Bacteria. *Appl Environ Microbiol* 66(9):4142–4144.
132. Manzanera M, et al. (2002) Hydroxyectoine Is Superior to Trehalose for Anhydrobiotic Engineering of *Pseudomonas putida* KT2440. *Appl Environ Microbiol* 68(9):4328–4333.
133. Nyska A, Kohen R (2002) Oxidation of biological systems: oxidative stress phenomena, antioxidants, redox reactions, and methods for their quantification. *Toxicol Pathol* 30(6):620–650.
134. Chavarría M, Nikel PI, Pérez-Pantoja D, De Lorenzo V (2013) The Entner-Doudoroff pathway empowers Pseudomonas putidaKT2440 with a high tolerance to oxidative stress. *Environ Microbiol* 15(6):1772–1785.
135. Hishinuma S, Yuki M, Fujimura M, Fukumori F (2006) OxyR regulated the expression of two major catalases, KatA and KatB, along with peroxiredoxin, AhpC in Pseudomonas putida. *Environ Microbiol* 8(12):2115–2124.
136. Kim J, Jeon CO, Park W (2008) Dual regulation of zwf-1 by both 2-keto-3-deoxy-6-phosphogluconate and oxidative stress in Pseudomonas putida. *Microbiology* 154(12):3905–3916.
137. Park W, Peña-Llopis S, Lee Y, Demple B (2006) Regulation of superoxide stress in Pseudomonas putida KT2440 is different from the SoxR paradigm in Escherichia coli. *Biochem Biophys Res Commun* 341(1):51–56.
138. Yeom J, Jeon CO, Madsen EL, Park W (2009) Ferredoxin-NADP+ Reductase from Pseudomonas putida Functions as a Ferric Reductase. *J Bacteriol* 191(5):1472–1479.
139. Yeom J, Jeon CO, Madsen EL, Park W (2009) In vitro and In vivo Interactions of Ferredoxin-NADP+ Reductases in Pseudomonas putida. *J Biochem* 145(4):481–491.
140. Yeom S, Yeom J, Park W (2010) NtrC-Sensed Nitrogen Availability Is Important for Oxidative Stress Defense in Pseudomonas putida KT2440. *J Microbiol* 48(2):153–159.
141. Yeom S, Yeom J, Park W (2010) Molecular characterization of FinR, a novel redox-sensing transcriptional regulator in Pseudomonas putida KT2440. *Microbiology* 156(5):1487–1496.
142. Yeom J, Lee Y, Park W (2012) ATP-dependent RecG Helicase Is Required for the Transcriptional Regulator OxyR Function in Pseudomonas species. *J Biol Chem* 287(29):24492–24504.
143. Ramos JL, et al. (2015) Mechanisms of solvent resistance mediated by interplay of cellular factors in Pseudomonas putida. *FEMS Microbiol Rev* 39(4):555–566.
144. Martins Dos Santos VAP, Heim S, Moore ERB, Str??tz M, Timmis KN (2004) Insights into the genomic basis of niche specificity of Pseudomonas putida KT2440. *Environ Microbiol* 6(12):1264–1286.
145. Functional Genomics of Stress Response in Pseudomonas putida KT2440.pdf
146. La Rosa R, Nogales J, Rojo F (2015) The Crc/CrcZ-CrcY global regulatory system helps the integration of gluconeogenic and glycolytic metabolism in Pseudomonas putida. *Environ Microbiol* 17(9):3362–3378.

147. Rojo F (2010) Carbon catabolite repression in Pseudomonas: optimizing metabolic versatility and interactions with the environment. *FEMS Microbiol Rev* 34(5):658–684.
148. Wolff JA, MacGregor CH, Eisenberg RC, Phibbs P V. (1991) Isolation and Characterization of Catabolite Repression Control Mutants of Pseudomonas aeruginosa PAO. *J Bacteriol* 173(15):4700–4706.
149. Moreno R, Martínez-Gomariz M, Yuste L, Gil C, Rojo F (2009) The Pseudomonas putida Crc global regulator controls the hierarchical assimilation of amino acids in a complete medium: Evidence from proteomic and genomic analyses. *Proteomics* 9(11):2910–2928.
150. Moreno R, Rojo F (2008) The Target for the Pseudomonas putida Crc Global Regulator in the Benzoate Degradation Pathway Is the BenR Transcriptional Regulator. *J Bacteriol* 190(5):1539–1545.
151. Morales G, Linares JF, Beloso A, Albar JP, Rojo F (2004) The Pseudomonas putida Crc Global Regulator Controls the Expression of Genes from Several Chromosomal Catabolic Pathways for Aromatic Compounds. *J Bacteriol* 186(5):1337–1344.
152. Kurbatov L, Albrecht D, Herrmann H, Petruschka L (2006) Analysis of the proteome of Pseudomonas putida KT2440 grown on different sources of carbon and energy. 8:466–478.
153. Krayl M, Benndorf D, Loffhagen N, Babel W (2003) Use of proteomics and physiological characteristics to elucidate ecotoxic effects of methyl tert -butyl ether in Pseudomonas putida. 1544–1552.
154. Bakalarski C, Kirkpatrick D (2016) A Biologist's Field Guide to Multiplexed Quantitative Proteomics. *Mcp* 13975(615):1–31.
155. Ong S-E (2002) Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Mol Cell Proteomics* 1(5):376–386.
156. Doherty MK, Whitehead C, McCormack H, Gaskell SJ, Beynon RJ (2005) Proteome dynamics in complex organisms: Using stable isotopes to monitor individual protein turnover rates. *Proteomics* 5(2):522–533.
157. Cambridge SB, et al. (2011) Systems-wide proteomic analysis in mammalian cells reveals conserved, functional protein turnover. *J Proteome Res* 10(12):5275–5284.
158. Lam YW, Lamond AI, Mann M, Andersen JS (2007) Analysis of Nucleolar Protein Dynamics Reveals the Nuclear Degradation of Ribosomal Proteins. *Curr Biol* 17(9):749–760.
159. Milner E, Barnea E, Beer I, Admon A (2006) The turnover kinetics of major histocompatibility complex peptides of human cancer cells. *Mol Cell Proteomics* 5(2):357–365.
160. Bantscheff M, Lemeer S, Savitski MM, Kuster B (2012) Quantitative mass spectrometry in proteomics: Critical review update from 2007 to the present. *Anal Bioanal Chem* 404(4):939–965.
161. Zhang R, Sioma CS, Wang S, Regnier FE (2001) Fractionation of isotopically labeled peptides in quantitative proteomics. *Anal Chem* 73(21):5142–5149.
162. Werner T, et al. (2012) High-resolution enabled TMT 8-plexing. *Anal Chem* 84(16):7188–7194.
163. McAlister GC, et al. (2012) Increasing the multiplexing capacity of TMTs using

reporter ion isotopologues with isobaric masses. *Anal Chem* 84(17):7469–7478.
164. Hebert AS, et al. (2013) Neutron-encoded mass signatures for multi-plexed proteome quantification. Nature methods.10(4):332–334.
165. Oda Y, Huang K, Cross FR, Cowburn D, Chait BT (1999) Accurate quantitation of protein expression and site-specific phosphorylation. *Proc Natl Acad Sci* 96(12):6591–6596.
166. Gygi SP, et al. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* 17(10):994–9.
167. Kozuka-hata H, Goto Y, Oyama M (2013) Phosphoproteomics-Based Characterization of Cancer Cell Signaling Networks. *Oncogenomics Cancer Proteomics – Nov Approaches Biomarkers Discov Ther Targets Cancer*:185–206.
168. Philip L. Ross‡ YNH (2004) Multiplexed Protein Quantitation in\nSaccharomyces cerevisiae Using\nAmine-reactive Isobaric Tagging Reagents*. *Mol Cell Proteomics* 3.12:1154–1169.
169. Wiese S, Reidegeld KA, Meyer HE, Warscheid B (2007) Protein labeling by iTRAQ: A new tool for quantitative mass spectrometry in proteome research. *Proteomics* 7(3):340–350.
170. Thompson A, et al. (2003) Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* 75(8):1895–1904.
171. Zhu W, Smith JW, Huang CM (2010) Mass spectrometry-based label-free quantitative proteomics. *J Biomed Biotechnol* 2010. doi:10.1155/2010/840518.
172. Chromatography NL, Spectrometry M (2006) Label-free Quantitative Proteomics Using Large Peptide Data Sets Generated by. *Mol Cell Proteomics*:1338–1347.
173. Bakalarski CE, et al. (2008) The impact of peptide abundance and dynamic range on stable-isotope-based quantitative proteomic analyses. *J Proteome Res* 7(11):4756–65.
174. Griffin NM, et al. (2010) Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nat Biotechnol* 28(1):83–89.
175. Callister SJ, et al. (2006) Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J Proteome Res* 5(2):277–286.
176. Luber CA, et al. (2010) Quantitative Proteomics Reveals Subset-Specific Viral Recognition in Dendritic Cells. *Immunity* 32(2):279–289.
177. Washburn MP, Wolters D, Yates JR (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* 19(3):242–7.
178. Zybailov B, Coleman MK, Florens L, Washburn MP (2005) Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. *Anal Chem* 77(19):6218–6224.
179. Dong M-Q, et al. (2007) Quantitative mass spectrometry identifies insulin signaling targets in C. elegans. *Science* 317(2007):660–663.
180. Chelius D, Bondarenko P V. (2002) Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. *J Proteome Res* 1(4):317–323.

181. Wiener MC, Sachs JR, Deyanova EG, Yates NA (2004) Differential mass spectrometry: A label-free LC-MS method for finding significant differences in complex peptide and protein mixtures. *Anal Chem* 76(20):6085–6096.
182. Higgs RE, Knierman MD, Gelfanova V, Butler JP, Hale JE (2005) Comprehensive label-free method for the relative quantification of proteins from biological samples. *J Proteome Res* 4(4):1442–1450.
183. Natl. Cent. Biotechnol. 2011. What kind of data can be submitted to GenBank? In GenBank Submissions Handbook,Natl. Cent. Biotechnol., Bethesda, Md. http://www.ncbi.nlm.nih.gov/books/NBK53707/
184. Creative proteomics: http://www.creative-proteomics.com/services/itraq-based-proteomics-analysis.htm

Chapter 5 Papers and manuscripts

Manuscript 1

Identification and validation of novel small proteins in *Pseudomonas putida*

Xiaochen Yang, Sheila I. Jensen, Tune Wulff, Scott J. Harrison,

Katherine S. Long

Identification and validation of novel small proteins in *Pseudomonas putida*

Xiaochen Yang, Sheila I. Jensen, Tune Wulff, Scott J. Harrison[#], Katherine S. Long*

Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Hørsholm, Denmark

[#] Current Address: School of BioSciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT

* Corresponding author
Katherine S. Long
Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kogle Allé 6, DK-2970 Hørsholm, Denmark
Tel: (+45) 45258024, Fax: (+45) 45258001
E-mail: kalon@biosustain.dtu.dk

Running title: Novel small proteins in *P. putida*

Keywords: Small proteins, *Pseudomonas putida*, KT2440, sORFs, SPA tag, proteomics, bioinformatics, gene expression/regulation

Summary

Small proteins of fifty amino acids or less have been understudied due to difficulties that impede their annotation and detection. In this study, bioinformatic and proteomic approaches have been used to identify putative small open reading frames (sORFs) in the well-characterized *Pseudomonas putida* strain KT2440. These combined with hypothetical sORFs annotated in the genome were assessed based on criteria of evidence of transcription, sequence conservation and genome context to identify candidates for validation. A plasmid-based system was established for sORF validation, enabling expression of C-terminal sequential peptide affinity (SPA) tagged variants and their detection via protein immunoblotting. Out of 22 tested putative sORFs, the expression of fourteen sORFs was confirmed, where all except one are novel. All of the validated sORFs except one are located adjacent to annotated genes on the same strand and three are in close proximity to genes with known functions. These include an ABC transporter operon and the two transcriptional regulators Fis and CysB involved in biofilm formation and cysteine biosynthesis, respectively. These are interesting small proteins for future studies involving identification of binding partners and further functional characterization.

Introduction

Although the functions of numerous large proteins have been investigated in great detail, knowledge on small proteins, especially those composed of fifty amino acids or less is comparatively sparse (Storz et al., 2014). Small proteins (s-proteins) are polypeptides encoded by small open reading frames (sORFs) that are translated by the ribosome. They are often excluded from genome annotations by minimum size cutoffs to limit the number of hypothetical ORFs. In addition, they readily escape detection with standard gel electrophoresis systems used for larger proteins.

In bacteria, most characterized s-proteins can be classified as ribosomal proteins, leader peptides or toxic peptides. A number of s-proteins with unknown functions have been identified, particularly in the model organism *Escherichia coli* (Hemm et al., 2008; Hemm et al., 2010). Progress in characterizing these shows that they exhibit great functional diversity, including acting as chaperones (Gaballa et al., 2008; Waters et al., 2011), playing roles in spore formation (Levin et al., 1993; Ebmeier et al., 2012) and cell division (Handler et al., 2008; Modell et al., 2011; Karimova et al., 2012), as well as in the regulation of transport (Gassel et al., 1999; Wadler and Vanderpool, 2007; Hobbs et al., 2012), membrane-bound enzymes (Alix and Blanc-Potard, 2008; Kato et al., 2012; VanOrsdel et al., 2013), protein

kinases and signal transduction (Burkholder et al., 2001; Lippa and Goulian, 2009). Most of these s-proteins are localized to and function at the membrane (Hemm et al., 2008; Storz et al., 2014). Although there are a few examples of s-proteins (KdpF, CydX) encoded within operons containing larger proteins that are found in more than one phylogenetic class (Gassel et al., 1999; VanOrsdel et al., 2013), they typically show limited conservation (Storz et al., 2014).

The ubiquitous Gram-negative bacterium *Pseudomonas putida* is found in a variety of terrestrial and aqueous environments, including strains that are part of the rhizosphere or isolated from soil contaminated with chemical waste. Its characteristics include a versatile metabolism, robustness towards stressors and tolerance towards organic solvents, combined with the availability of genetic tools make it an attractive host for applications in industrial biotechnology and synthetic biology (Nikel et al., 2014). In this work, bioinformatic and proteomic approaches have been used to identify candidate s-proteins in the well-characterized *P. putida* strain KT2440. Based on RNA sequencing data and BLAST analysis, 22 sORF candidates were selected and the expression of 14 was validated by western blot. These include three sORFs in close proximity to annotated genes, where one is likely part of an ABC transporter operon and two are adjacent to genes encoding transcriptional regulators. This study represents the first step in understanding the collection of s-proteins encoded in the *P. putida* genome and pinpoints a set of novel sORFs that are interesting targets for further functional characterization.

Results and Discussion

Identification of putative sORFs via genome analysis and bioinformatic prediction

Computational prediction is a rapid method to identify putative ORFs in bacterial genomes using sequence homology, but its utility for sORFs is limited by their small size and poor annotation. The genome of *P. putida* KT-2440 was first examined to investigate the annotated genes that encode putative s-proteins. The genome annotation is mainly based on computational prediction with interpolated Markov models (Salzberg et al., 1998; Nelson et al., 2002), a method focused on genome comparison that is more accurate for long ORF prediction. A total of 164 annotated sequences of less than 154 base pairs were present in the *Pseudomonas* genome database (Winsor et al., 2011), of which 82 sequences corresponded to tRNA or rRNA genes. Two genes (*rmpJ* and *rmpH*) encoded ribosomal proteins comprised of less than 51 amino acids. Of the remaining 80 sequences classified as coding sequences (CDS), only one sORF (PP_0380) is identified as *pqqA* that encodes a small

96 protein that functions as a substrate for coenzyme PQQ synthesis (Puehringer et al., 2008).
97 The remaining sequences were all annotated as hypothetical proteins. As there is no
98 experimental validation of these 80 annotated sequences, they were collected as
99 hypothetical sORFs for further analysis (Figure 1).
100 In order to identify additional sORFs not included in the genome annotation, a bioinformatic
101 approach was used to search for sORFs not predicted by homology analysis. To this end, a
102 custom bioinformatic tool was developed to detect putative sORFs with an upstream Shine-
103 Dalgarno (SD) sequence, which is a ribosome-binding site upstream of start codons on
104 bacterial mRNAs that can base pair with a complementary sequence at the 3′-end of 16S
105 rRNA, recruiting the 30S subunit to mRNA during initiation of protein synthesis (Shine and
106 Dalgarno, 1974; Steitz and Jakes, 1975). The SD sequence and the spacing between SD
107 sequence and start codon affect the efficiency of protein expression, but the sequence can
108 vary and is not necessary in all cases for protein expression. Here, the oligonucleotide
109 sequence GGAG was used to identify sORFs. The sequences corresponding to the intergenic
110 regions of the genome were selected and examined with the custom Python program
111 designed to identify the SD sequence feature (GGAG), start codons (ATG, CTG, TTG, GTG),
112 and stop codons (TAA, TGA, TAG). The distance between the SD sequence feature and the
113 putative start codon was limited to 20 bp or less. The sequences with lengths between 45
114 and 153 bp that fulfilled these requirements were collected as putative sORFs. A total of 273
115 putative sORFs were obtained from the intergenic regions using this custom bioinformatics
116 tool (Figure 1).
117 Two other small protein genes have been identified recently in *P. putida*. The small protein
118 gene *cydZ* was identified in a conservation study to find homologs of CydX, an essential small
119 protein in the cytochrome *bd* oxidase complex (Allen et al., 2014). In *P. putida*, the *cydZ* gene
120 (PP_4649) is located in the *cioAB* operon encoding cyanide-insensitive cytochrome oxidases
121 that exhibit decreased sensitivity to hydrogen cyanide. The CydZ protein was not included in
122 the group of hypothetical sORFs collected from the *Pseudomonas* genome database because
123 it is comprised of 51 amino acids and larger than the maximum length of 50 amino acids.
124 Another recent study reanalyzed the *P. putida* KT2440 genome and improved the
125 annotation, including the identification of 242 new ORFs and functional reannotation of
126 1548 genes (Belda et al., 2016). The gene encoding the small protein KdpF (PP_5660) was
127 identified and the presence of a complete Kdp transport system involved in mediating
128 postassium ion uptake was confirmed. The *kdpF* gene was also identified in this study by the
129 bioinformatic tool described above (BIO-10) (Table 1).

Identification of putative sORFs by proteomics

Mass spectrometry was chosen as an orthogonal approach to identify sORFs as this method detects peptides that are evidence of sORF translation. There are two significant challenges, however, that impede the application of mass spectrometry. The first issue relates to the manner in which MS hits are identified via comparison of MS/MS data with a protein sequence database. Due to the limited knowledge of small proteins in *P. putida* and other bacteria, the protein sequence database for MS/MS data analysis is inadequate with the consequence that small proteins, although detected by mass spectrometry would not be identified because their sequences are largely absent from the database. This was addressed by collecting all possible small protein sequences from the *P. putida* genome. A customized database was constructed by determining all the DNA sequences with less than 200 bp from start codon (ATG, CTG, TTG, GTG) to stop codon (TAA, TGA, TAG) and translating them into amino acid sequences. More than 80,000 putative small protein sequences were included in the database.

The second issue with s-proteins is that their size limits the number of trypsin digestion sites they contain, the number of peptide fragments produced and ultimately the number of ionized fragments detected by mass spectrometry. In standard proteomics data analysis of larger proteins, a minimum of two peptides must be identified for positive identification of a protein sequence in order to limit the number of false positive hits. In this study, detection of one peptide was sufficient for the positive identification of an s-protein hit. Thus, small protein sequences were not discarded in the event of a single identifiable tryptic peptide.

The protein samples used for proteomics were total proteins extracted from *P. putida* KT2440 grown in LB broth with cells harvested at OD_{600} values of 1.5 and 4.5. The experimental details are provided in the supplementary information (Supplementary File 1). Briefly, the total protein samples were separated on polyacrylamide gels and the gel lanes containing the samples were each cut into 10 to 12 pieces. The proteins from each gel piece were extracted, digested with trypsin, and analyzed on a Synapt G2 Q-TOF instrument using electrospray ionization (ESI) with a NanoLock-spray source. The protein sequences were identified by the Protein Lynx global server using the MSe search function against the custom database.

The mass spectrometry data from all total protein samples yielded 93 small protein sequences with lengths less than 66 amino acids. Ten sequences were ribosomal proteins or fragments of annotated proteins. The remaining 83 sequences were mapped to the genome

and around 75% were located within annotated genes. Of these, 44 sequences were antisense sORFs located on the opposite strand (Figure 2A), while 20 sequences were sORFs overlapping annotated genes on the same strand but using a different reading frame (Figure 2B). Ten sORFs were partly overlapping annotated genes, either on the same or opposite strand (Figure 2C and 2D), whereas only nine sORFs were located in intergenic regions of the genome (Figure 2E). The latter were not detected with the custom bioinformatics tool because their upstream sequences did not contain the SD sequence feature (GGAG). Upon checking which gel slices the s-proteins originated from, it was found that most of s-proteins (71/83) were detected from gel slices of significantly higher molecular weight compared to their expected size. This indicates that many s-proteins did not migrate according to size despite the denaturing conditions used for electrophoresis and suggests that they could be associated with larger proteins. Several characterized bacterial small proteins are known to associate with larger proteins, including small protein regulators of small molecule transport (AcrZ, SgrT) (Hobbs et al., 2012; Kosfeld and Jahreis, 2012) and membrane-bound enzymes (CydX, MgtR) (Lee et al., 2013; VanOrsdel et al., 2013).

While mass spectrometry has been widely used for sORF identification in eukaryotes (Yang et al., 2011; Slavoff et al., 2013; Ericson et al., 2014; Ma et al., 2014), there are only a few cases where it has been used in bacteria (Kim et al., 2009; Lluch-Senar et al., 2015). One study in *Pseudomonas fluorescens* Pf0-1 used proteomics to detect 16 novel ORFs, including 2 sORFs, which were further validated with RT-PCR (Kim et al., 2009). Among the novel ORFs, five were located in intergenic regions, but nine were located on the opposite strand and one overlapping in a different reading frame relative to annotated genes. Although five of the novel ORFs were detected with the gene prediction tools GenemarkS (Besemer et al., 2001) and Glimmer (Delcher et al., 1999; Delcher et al., 2007) these did not include the sORFs, supporting the notion that these tools are not suitable for sORF prediction. Another study combined mass spectrometry with size exclusion chromatography to detect small proteins in the genome-reduced bacterium *Mycoplasma pneumoniae* (Lluch-Senar et al., 2015). It was found that the majority of small proteins eluted in fractions of significantly higher molecular weight compared to the expected size of the individual small proteins, supporting similar observations made here with a different method and experimental conditions.

Selection of candidate sORFs for validation

Several criteria were used to evaluate the putative sORFs and select candidates for further validation. The first approach was to analyze each sORF for evidence of transcription using strand-specific RNA-sequencing data. The transcriptome data was collected from *P. putida* cells cultivated in LB or M9 minimal media at exponential and stationary phase (D'Arrigo et al., 2016) (Bojanovic et al., unpublished data). The number of reads mapping to the genome positions corresponding to each putative sORF was checked. Out of a total of 436 putative sORFs, 24 obtained from mass spectrometry that overlap or partly overlap annotated genes on the same strand (Figure 2B and 2D) were excluded from the analysis as it was not possible to unambiguously check for transcription with RNA-seq data. Out of the remaining 412 putative sORFs, transcripts corresponding to 186 sORFs were detected in the RNA sequencing data. Evidence supporting transcription of half of the putative sORFs (30/59) obtained by mass spectrometry was detected in the RNA sequencing data. However, for the MS sORFs that were located in intergenic regions, 8 out of 9 sORFs had evidence of transcription. The proportion of putative sORFs obtained from the *Pseudomonas* genome database and bioinformatic prediction with transcription evidence was lower than 50% (39% for annotated sORFs and 46% for bioinformatics prediction) (Figure 1B).

To further evaluate the putative sORFs for expression validation, sequence conservation and genome context analysis were used. The tBlastN program was used to analyze the conservation of DNA sequences and it showed that 81 out of 186 sORFs were conserved within pseudomonads. The genome contexts of putative sORFs located in intergenic regions were analyzed using the *Pseudomonas* genome and KEGG databases (Kanehisa and Goto, 2000; Winsor et al., 2011; Kanehisa et al., 2016) with the aim of using the annotation of neighboring genes to gain information on the possible biological processes they could be involved in. Based on the conservation and genome context analyses, twenty-two sORFs candidates were selected for further validation and are summarized in Table 1. All of these sORFs were located in intergenic regions and comparatively close to annotated genes (less than 300bp), although the annotated genes were in some cases hypothetical proteins.

Validation of sORF candidates

A plasmid-based expression system was developed in order to further validate expression of the putative sORFs. The pSEVA-631 plasmid (Silva-Rocha et al., 2013) was chosen as the backbone and modified for sORF validation. A pSEVA631-GFP plasmid (Figure 3A) was first generated (Supplementary Files 1 and 2), where GFP expression is controlled by the *rrnB* promoter with activity positively correlated to the growth rate of the cell (Bartlett and

231 Gourse, 1994). In order to facilitate detection of the sORFs on polyacrylamide gels, a
232 sequence encoding a sequential peptide affinity (SPA) (Zeghouf et al., 2004) tag was
233 introduced into the plasmid to enable expression of C-terminal SPA-tagged sORFs. This
234 affinity tag has a size (8 kDa) similar to the sORF candidates and contains three tandem FLAG
235 sequences that can be detected with a commercial antibody. The SPA-tag sequence was first
236 amplified from plasmid pJL-148 (Zeghouf et al., 2004) and then inserted at the C-terminus of
237 the GFP ORF to yield plasmid pSEVA-631-GFP-SPA (Figure 3B). This was followed by
238 replacement of the GFP and its Shine-Dalgarno sequence with sORF sequences (Figure 3C) to
239 yield the pSEVA-631-sORF-SPA plasmids (Figure 3D). The sORF sequences were amplified
240 and inserted in between the *rrnB* promoter and SPA-tag. The sORF sequences inserted into
241 the plasmid also contain 15-50bp of upstream sequence which may include necessary
242 elements for translation (Figure 3C). As transcription of the sORF candidates has already
243 been verified with RNA sequencing data, the *rrnB* promoter will not affect the validity of the
244 results, but rather provide a predictable transcription of the insertion. The insertions of 22
245 sORF recombinant plasmids were confirmed by DNA sequencing and transferred into
246 competent *P. putida* cells for validation of protein expression. The *P. putida* KT2440 strain
247 without a plasmid was used as a control. The total protein samples were extracted from cells
248 harvested in exponential phase, separated on polyacrylamide gels and tested for the
249 expression of sORF-SPA-tag fusion proteins via western blotting. The western blotting
250 experiment was repeated three to five times for each sample to ensure reproducibility of
251 the data.
252 The western blot results with positive evidence for the expression of fourteen sORF
253 candidates are shown in Figure 4E. For most of the sORF candidates, the bands
254 corresponding to the sORF-SPA fusion proteins have sizes between 10-15 kDa (Figure 4),
255 which is consistent with their expected sizes. For some candidates the tagged sORFs have
256 bands that are larger than expected (15-20 kDa) and in some cases more than one band is
257 present. As the lengths of upstream sORF sequence introduced into the pSEVA-631-sORF-
258 SPA plasmids were very short, it is unlikely that larger or multiple bands are due to multiple
259 open reading frames.
260 For most of the validated sORFs, potential SD sequences can be identified (Table 1).
261 However, for the validated sORFs obtained from mass spectrometry MS-2 lacks an
262 identifiable SD sequence and for MS-4 the most likely SD sequence is located 40 bp
263 upstream of the start codon (Table 1). In the case of PGD-4, there is an upstream AGG
264 sequence, but this may be too short to function as an effective SD sequence. This suggests

265 that a strong SD sequence is not required for the expression of all sORFs. Eight of the sORF
266 candidates, including BIO-10 (*kdpF*), were not validated by western blot. This does not
267 necessarily mean that these sORFs do not encode small proteins but rather that their
268 expression could not be confirmed under the tested conditions in this study. One possible
269 scenario is the posttranscriptional regulation of s-protein expression by RNA secondary
270 structures including RNA thermometers that occlude the ribosome binding site and prevent
271 translation. There are many examples of small proteins in *E. coli* that accumulate under
272 specific growth conditions or are stress induced, where a subset are likely subject to
273 posttranscriptional regulation (Hemm et al., 2008; Hemm et al., 2010).
274
275 **Genomic context of validated sORFs**
276 As a first step towards understanding the possible functions of the validated sORFs, their
277 genomic context and proximity to neighboring annotated genes were examined and the
278 results are summarized in Table 2. All of the validated sORFs are located within 200 bp of
279 annotated genes on the same strand with the exception of BIO-11. In this case, there are no
280 annotated genes within 5 kbp on the same strand. Eleven of the validated sORFs are in close
281 proximity to annotated genes or operons with specific functions, while for two sORFs the
282 neighboring genes were hypothetical ORFs. The two sORFs PGD-4 and PGD-8 are annotated
283 as part of operons (Mao et al., 2009). The three sORFs MS-2, BIO-1 and BIO-6 are located in
284 positions adjacent to annotated operons. While there are no annotated operons close to the
285 sORFs MS-4, BIO-2, BIO-4, BIO-8, and BIO-9, they are located close to annotated genes of
286 known function and this information may lead to the identification of new operons. In the
287 new *P. putida* KT2440 genome annotation, the newly identified gene PP_5559 (Belda et al.,
288 2016) matched the validated sORF BIO-9.
289 In order to determine whether the relationships between the sORFs and neighboring genes
290 are conserved in other pseudomonads, the sORFs together with their adjacent genes were
291 analyzed with BlastN. For the eight sORFs discovered with bioinformatic prediction and
292 proteomics, all (MS-2, MS-4, BIO-2, BIO-4, BIO-6, BIO-8, BIO-9) except BIO-1 were conserved
293 in pseudomonads. The presence of synteny between these sORFs as well as those obtained
294 from the *Pseudomonas* genome database and their respective neighboring genes supports
295 possible functional relationships and may provide clues to the biological processes the sORFs
296 are involved in.
297 Three of the validated sORFs are located adjacent to genes encoding transporters. In
298 particular, MS-2 is located four bp upstream of the first gene of an ABC transporter operon

299 containing four genes (Eichhorn et al., 1997). In the new *P. putida* KT2440 genome
300 annotation, another new gene (PP_5431) was annotated 73 bp upstream of MS-2 (Belda et
301 al., 2016), indicating that the ABC transporter operon may be larger than expected. Several
302 small proteins are known to associate with small molecule transporters, including AcrZ and
303 SgrT in *E. coli*. The AcrZ protein interacts with the AcrB, the inner membrane component of
304 the AcrAB-TolC efflux pump, and may serve to enhance export of specific substrates (Hobbs
305 et al., 2012). The SgrT protein interacts with the glucose transporter of the
306 phosphoenolpyruvate –dependent glucose phosphotransferase system, $EIICB^{Glc}$, and inhibits
307 glucose uptake (Wadler and Vanderpool, 2007).
308 Two validated sORFs are located close to genes encoding transcriptional regulators. The BIO-
309 2 sORF is located 60 bp upstream of the Fis transcriptional regulator that is known to
310 activate rRNA transcription (Ross et al., 1990). It has also been associated with biofilm
311 formation in *P. putida* (Jakovleva et al., 2012). The BIO-4 sORF is located around 50 bp
312 downstream of the CysB transcriptional regulator gene. CysB is a member of the LysR family
313 of transcriptional regulatory proteins and regulates the expression of genes associated with
314 cysteine biosynthesis in bacteria (Delic-Attree et al., 1997; Lochowska et al., 2001).
315
316 **Concluding remarks**
317 In this first investigation of small proteins in *P. putida*, the expression of fourteen small
318 proteins identified via bioinformatic prediction and proteomic approaches is documented.
319 Although it was possible to validate expression of a subset of sORFs identified with each
320 approach, there was essentially no overlap between the putative sORFs obtained from the
321 different approaches, suggesting that there are more small proteins to be identified. This
322 study represents an initial step towards the identification and characterization of the
323 collection of s-proteins encoded in the *P. putida* genome. An approach such as ribosome
324 profiling has promising potential to provide unbiased information on the hitherto largely
325 ignored small proteome, without requirements such as broad sequence conservation, prior
326 knowledge regarding upstream sequences or compatibility issues with mass spectrometry.
327
328 **Acknowledgments**
329 We thank Isotta D´Arrigo and Klara Bojanovič for sharing their RNA-seq data prior to
330 publication, Stefan Kol and Helle Munck Petersen for their expertise and help with protein
331 electrophoresis and western blotting, Holger Døssing and Marta Matos for helpful advice on

the Python program and Christine Finnie for helpful discussions. Gisela Storz and Aixia Zhang are acknowledged for providing plasmid pJL-148 for SPA-tag amplification.

This work is supported by the Novo Nordisk Foundation and a PhD grant from the People Programme (Marie Curie Actions) of the European Union Seventh Framework Programme [FP7-People-2012-ITN], under grant agreement no. 317058,"BACTORY".

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

References

Alix, E., and Blanc-Potard, A.B. (2008) Peptide-assisted degradation of the Salmonella MgtC virulence factor. *EMBO J* **27**: 546-557.

Allen, R.J., Brenner, E.P., VanOrsdel, C.E., Hobson, J.J., Hearn, D.J., and Hemm, M.R. (2014) Conservation analysis of the CydX protein yields insights into small protein identification and evolution. *BMC Genomics* **15**: 946.

Bartlett, M.S., and Gourse, R.L. (1994) Growth rate-dependent control of the rrnB P1 core promoter in Escherichia coli. *J Bacteriol* **176**: 5560-5564.

Belda, E., van Heck, R.G., Lopez-Sanchez, M.J., Cruveiller, S., Barbe, V., Fraser, C. et al. (2016) The revisited genome of Pseudomonas putida KT2440 enlightens its value as a robust metabolic chassis. *Environ Microbiol.*

Besemer, J., Lomsadze, A., and Borodovsky, M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* **29**: 2607-2618.

Burkholder, W.F., Kurtser, I., and Grossman, A.D. (2001) Replication initiation proteins regulate a developmental checkpoint in Bacillus subtilis. *Cell* **104**: 269-279.

D'Arrigo, I., Bojanovič, K., Yang, X., Rau, M.H., and Long, K.S. (2016) Genome-wide mapping of transcription start sites yields novel insights into the primary transcriptome of *Pseudomonas putida. Environmental Microbiology, accepted.*

Delcher, A.L., Bratke, K.A., Powers, E.C., and Salzberg, S.L. (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**: 673-679.

Delcher, A.L., Harmon, D., Kasif, S., White, O., and Salzberg, S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27**: 4636-4641.

364 Delic-Attree, I., Toussaint, B., Garin, J., and Vignais, P.M. (1997) Cloning, sequence and
365 mutagenesis of the structural gene of Pseudomonas aeruginosa CysB, which can activate
366 algD transcription. *Mol Microbiol* **24**: 1275-1284.
367 Ebmeier, S.E., Tan, I.S., Clapham, K.R., and Ramamurthi, K.S. (2012) Small proteins link coat
368 and cortex assembly during sporulation in Bacillus subtilis. *Mol Microbiol* **84**: 682-696.
369 Eichhorn, E., van der Ploeg, J.R., Kertesz, M.A., and Leisinger, T. (1997) Characterization of
370 alpha-ketoglutarate-dependent taurine dioxygenase from Escherichia coli. *J Biol Chem* **272**:
371 23031-23036.
372 Ericson, M., Janes, M.A., Butter, F., Mann, M., Ullu, E., and Tschudi, C. (2014) On the extent
373 and role of the small proteome in the parasitic eukaryote Trypanosoma brucei. *BMC Biol* **12**:
374 14.
375 Gaballa, A., Antelmann, H., Aguilar, C., Khakh, S.K., Song, K.B., Smaldone, G.T., and Helmann,
376 J.D. (2008) The Bacillus subtilis iron-sparing response is mediated by a Fur-regulated small
377 RNA and three small, basic proteins. *Proc Natl Acad Sci U S A* **105**: 11927-11932.
378 Gassel, M., Mollenkamp, T., Puppe, W., and Altendorf, K. (1999) The KdpF subunit is part of
379 the K(+)-translocating Kdp complex of Escherichia coli and is responsible for stabilization of
380 the complex in vitro. *J Biol Chem* **274**: 37901-37907.
381 Handler, A.A., Lim, J.E., and Losick, R. (2008) Peptide inhibitor of cytokinesis during
382 sporulation in Bacillus subtilis. *Mol Microbiol* **68**: 588-599.
383 Hemm, M.R., Paul, B.J., Schneider, T.D., Storz, G., and Rudd, K.E. (2008) Small membrane
384 proteins found by comparative genomics and ribosome binding site models. *Mol Microbiol*
385 **70**: 1487-1501.
386 Hemm, M.R., Paul, B.J., Miranda-Rios, J., Zhang, A., Soltanzad, N., and Storz, G. (2010) Small
387 stress response proteins in Escherichia coli: proteins missed by classical proteomic studies. *J*
388 *Bacteriol* **192**: 46-58.
389 Hobbs, E.C., Yin, X., Paul, B.J., Astarita, J.L., and Storz, G. (2012) Conserved small protein
390 associates with the multidrug efflux pump AcrB and differentially affects antibiotic
391 resistance. *Proc Natl Acad Sci U S A* **109**: 16696-16701.
392 Jakovleva, J., Teppo, A., Velts, A., Saumaa, S., Moor, H., Kivisaar, M., and Teras, R. (2012) Fis
393 regulates the competitiveness of Pseudomonas putida on barley roots by inducing biofilm
394 formation. *Microbiology* **158**: 708-720.
395 Kanehisa, M., and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic*
396 *Acids Res* **28**: 27-30.

397 Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016) KEGG as a
398 reference resource for gene and protein annotation. *Nucleic Acids Res* **44**: D457-462.
399 Karimova, G., Davi, M., and Ladant, D. (2012) The beta-lactam resistance protein Blr, a small
400 membrane polypeptide, is a component of the Escherichia coli cell division machinery. *J*
401 *Bacteriol* **194**: 5576-5588.
402 Kato, A., Chen, H.D., Latifi, T., and Groisman, E.A. (2012) Reciprocal control between a
403 bacterium's regulatory system and the modification status of its lipopolysaccharide. *Mol Cell*
404 **47**: 897-908.
405 Kim, W., Silby, M.W., Purvine, S.O., Nicoll, J.S., Hixson, K.K., Monroe, M. et al. (2009)
406 Proteomic detection of non-annotated protein-coding genes in Pseudomonas fluorescens
407 Pf0-1. *PLoS One* **4**: e8455.
408 Kosfeld, A., and Jahreis, K. (2012) Characterization of the Interaction Between the Small
409 Regulatory Peptide SgrT and the EIICBGlc of the Glucose-Phosphotransferase System of E.
410 coli K-12. *Metabolites* **2**: 756-774.
411 Lee, E.J., Pontes, M.H., and Groisman, E.A. (2013) A bacterial virulence protein promotes
412 pathogenicity by inhibiting the bacterium's own F1Fo ATP synthase. *Cell* **154**: 146-156.
413 Levin, P.A., Fan, N., Ricca, E., Driks, A., Losick, R., and Cutting, S. (1993) An unusually small
414 gene required for sporulation by Bacillus subtilis. *Mol Microbiol* **9**: 761-771.
415 Lippa, A.M., and Goulian, M. (2009) Feedback inhibition in the PhoQ/PhoP signaling system
416 by a membrane peptide. *PLoS Genet* **5**: e1000788.
417 Lluch-Senar, M., Delgado, J., Chen, W.H., Llorens-Rico, V., O'Reilly, F.J., Wodke, J.A. et al.
418 (2015) Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced
419 bacterium. *Mol Syst Biol* **11**: 780.
420 Lochowska, A., Iwanicka-Nowicka, R., Plochocka, D., and Hryniewicz, M.M. (2001) Functional
421 dissection of the LysR-type CysB transcriptional regulator. Regions important for DNA
422 binding, inducer response, oligomerization, and positive control. *J Biol Chem* **276**: 2098-
423 2107.
424 Ma, J., Ward, C.C., Jungreis, I., Slavoff, S.A., Schwaid, A.G., Neveu, J. et al. (2014) Discovery of
425 human sORF-encoded polypeptides (SEPs) in cell lines and tissue. *J Proteome Res* **13**: 1757-
426 1765.
427 Mao, F., Dam, P., Chou, J., Olman, V., and Xu, Y. (2009) DOOR: a database for prokaryotic
428 operons. *Nucleic Acids Res* **37**: D459-463.

Modell, J.W., Hopkins, A.C., and Laub, M.T. (2011) A DNA damage checkpoint in Caulobacter crescentus inhibits cell division through a direct interaction with FtsW. *Genes Dev* **25**: 1328-1343.

Nelson, K.E., Weinel, C., Paulsen, I.T., Dodson, R.J., Hilbert, H., Martins dos Santos, V.A. et al. (2002) Complete genome sequence and comparative analysis of the metabolically versatile Pseudomonas putida KT2440. *Environ Microbiol* **4**: 799-808.

Nikel, P.I., Martinez-Garcia, E., and de Lorenzo, V. (2014) Biotechnological domestication of pseudomonads using synthetic biology. *Nat Rev Microbiol* **12**: 368-379.

Puehringer, S., Metlitzky, M., and Schwarzenbacher, R. (2008) The pyrroloquinoline quinone biosynthesis pathway revisited: a structural approach. *BMC Biochem* **9**: 8.

Ross, W., Thompson, J.F., Newlands, J.T., and Gourse, R.L. (1990) E.coli Fis protein activates ribosomal RNA transcription in vitro and in vivo. *EMBO J* **9**: 3733-3742.

Salzberg, S.L., Delcher, A.L., Kasif, S., and White, O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res* **26**: 544-548.

Shine, J., and Dalgarno, L. (1974) The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci U S A* **71**: 1342-1346.

Silva-Rocha, R., Martinez-Garcia, E., Calles, B., Chavarria, M., Arce-Rodriguez, A., de Las Heras, A. et al. (2013) The Standard European Vector Architecture (SEVA): a coherent platform for the analysis and deployment of complex prokaryotic phenotypes. *Nucleic Acids Res* **41**: D666-675.

Slavoff, S.A., Mitchell, A.J., Schwaid, A.G., Cabili, M.N., Ma, J., Levin, J.Z. et al. (2013) Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat Chem Biol* **9**: 59-64.

Steitz, J.A., and Jakes, K. (1975) How ribosomes select initiator regions in mRNA: base pair formation between the 3' terminus of 16S rRNA and the mRNA during initiation of protein synthesis in Escherichia coli. *Proc Natl Acad Sci U S A* **72**: 4734-4738.

Storz, G., Wolf, Y.I., and Ramamurthi, K.S. (2014) Small proteins can no longer be ignored. *Annu Rev Biochem* **83**: 753-777.

VanOrsdel, C.E., Bhatt, S., Allen, R.J., Brenner, E.P., Hobson, J.J., Jamil, A. et al. (2013) The Escherichia coli CydX protein is a member of the CydAB cytochrome bd oxidase complex and is required for cytochrome bd oxidase activity. *J Bacteriol* **195**: 3640-3650.

461 Wadler, C.S., and Vanderpool, C.K. (2007) A dual function for a bacterial small RNA: SgrS
462 performs base pairing-dependent regulation and encodes a functional polypeptide. *Proc*
463 *Natl Acad Sci U S A* **104**: 20454-20459.
464 Waters, L.S., Sandoval, M., and Storz, G. (2011) The Escherichia coli MntR miniregulon
465 includes genes encoding a small protein and an efflux pump required for manganese
466 homeostasis. *J Bacteriol* **193**: 5887-5897.
467 Winsor, G.L., Lam, D.K., Fleming, L., Lo, R., Whiteside, M.D., Yu, N.Y. et al. (2011)
468 Pseudomonas Genome Database: improved comparative analysis and population genomics
469 capability for Pseudomonas genomes. *Nucleic Acids Res* **39**: D596-600.
470 Yang, X., Tschaplinski, T.J., Hurst, G.B., Jawdy, S., Abraham, P.E., Lankford, P.K. et al. (2011)
471 Discovery and annotation of small proteins using genomics, proteomics, and computational
472 approaches. *Genome Res* **21**: 634-641.
473 Zeghouf, M., Li, J., Butland, G., Borkowska, A., Canadien, V., Richards, D. et al. (2004)
474 Sequential Peptide Affinity (SPA) system for the identification of mammalian and bacterial
475 protein complexes. *J Proteome Res* **3**: 463-468.
476
477

Table 1. Candidate sORFs selected for validation

sORF candidate	Coordinates[a]	Strand	Length (bp)	Potential SD sequence[b]	Start codon	Distance between SD sequence and start codon (bp)[c]	Validation
MS-1	1236387-1236452	+	66	AGAA	CTG	4	No
MS-2	221114-221224	+	111	none	CTG	NA	Yes
MS-3	1607694-1607584	-	111	AAAG	ATG	4	No
MS-4	6148076-6147897	-	180	AGGAAG	CTG	40	Yes
MS-11	5466651-5466788	+	72	GAAAA	ATG	16	No
BIO-1	4022806-4022699	-	108	GGGAGG	ATG	8	Yes
BIO-2	60766-60671	-	96	AGGAG	ATG	10	Yes
BIO-4	2655328-2655447	+	120	GGGAG	ATG	15	Yes
BIO-5	2139686-2139829	+	144	GGAGAA	ATG	10	No
BIO-6	2528995-2529066	+	72	GGAGGG	ATG	11	Yes
BIO-7	5338386-5338454	+	180	GGAGA	ATG	2	No
BIO-8	2412871-2412963	+	72	AGGAGG	TTG	12	Yes
BIO-9	3466245-3466126	-	99	GGAGAG	ATG	4	Yes
BIO-10 (*kdpF*)	4705261-4705172	-	117	GGGAGA	ATG	7	No
BIO-11	1332685-1332542	-	123	AGGAGA	ATG	5	Yes
PGD-1 (*pqqA*)	PP_0380	-	72	GAAGGA	ATG	8	Yes
PGD-2	PP_0419	+	99	AGG	GTG	5	No
PGD-3	PP_1149	+	117	GGGAGG	ATG	5	Yes
PGD-4	PP_1685	+	123	AGG	ATG	9	Yes
PGD-5	PP_2644	+	147	AAG	ATG	15	No
PGD-6	PP_0023	+	126	AGGAAG	ATG	6	Yes
PGD-8	PP_3434	-	111	GAAGGA	ATG	24	Yes

[a] The genomic coordinates or PP gene designations are indicated.

[b] The possible Shine-Dalgarno (SD) sequence is given. In cases where there is more than one possible SD sequence, that closest to the start codon is provided.

[c] The number of base pairs between the possible SD sequence and start codon is indicated.

Table 2. Small open reading frames with validated expression

Name	Protein -SPA MW (kDa)[a]	5′/3′ Flanking gene[b]	Direction[c]	tBlastN conser-vation[d]	Operon/ Closest adjacent gene[b]	Operon/ gene annotation
MS-2	12.2	0168/0169	>>>	P	0169-0172	ABC transporter
MS-4	14.3	5394/5393	<<<	P	5393	Heavy metal transport/ detoxification protein
BIO-1	11.7	3548/3547	<<<	P	3550-3548	EmrB/QacA family drug resistance transporter
BIO-2	11.6	0052/0051	>><	P	0051	Transcript-ional regulator Fis
BIO-4	11.5	2327/2328	>><	P	2327	Transcript-ional regulator CysB
BIO-6	10.7	2217/2218	>>>	P	2212-2217	fad operon
BIO-8	11.3	2113/2114	>>>	P	2113	RNA 2′-O-ribose methyl-transferase
BIO-9	12.5	3081/3080	<<<	P	3081	PEBP family protein
BIO-11[e]	13.0	1158/1157	>><	pt	-	-
PGD-3	12.3	1148/1150	>>>	B	1147-1148	hypothetical
PGD-4	12.6	1684/1686	>>>	P	1684-1686	glutathione peroxidase[f]
PGD-6	13.0	0024/0022	<<<	P	0024	hypothetical
PGD-8	12.1	3435/3433	>><	P	3434-3433	hydroxyphenyl pyruvate dioxygenase[f]
PGD-1	10.9	0381/0379	<<<	B	0379 (*pqqB*)	PqqB

[a] Molecular weights of the SPA-tagged small proteins are given, where the SPA tag is 8.1 kDa.

[b] In cases where sORFs are in close proximity to operons, the genes included in the operon are indicated. Otherwise the closest gene is given. The gene names are indicated with their PP number designations.

[c] The symbols denote the orientations of the 5′ flanking gene, sORF gene, and 3′flanking gene.

[d] The degree of nucleotide conservation is denoted by: pt, limited to *P. putida*, P, within the *Pseudomonas* genus, or B, conserved in additional bacterial genera.

[e] There are no adjacent genes within 5 kbp on the same strand.

[f] These annotations refer to the single genes that are positioned closest to the sORF. The other genes in the operon have unknown functions.

502 **Figure legends**

503

504 **Figure 1.** Strategy of sORF identification and validation. (A) The approaches used for
505 identification of putative sORFs and the workflow for selection of candidate sORFs for
506 validation is diagrammed. (B) A summary of the numbers of sORFs obtained with each
507 approach and at each step of the workflow.

508

509 **Figure 2.** The location of mass spectrometry hits relative to annotated genes. The sORF hits
510 are overlapping annotated genes on the opposite strand (A), overlapping annotated genes
511 on the same strand but in a different reading frame (B), display partial overlap with genes on
512 the same strand but in a different reading frame (C), have partial antisense location relative
513 to annotated genes (D), or are located in intergenic regions (E) relative to the adjacent
514 annotated genes. Annotated genes and sORFs are depicted with black and gray arrows/bars,
515 respectively. The numbers of sORF hits are shown in parentheses. The sORF hits in groups B
516 and D were excluded from the transcription analysis, as it was not possible to distinguish
517 transcription of sORFs from the annotated genes in these cases. The sORF hits in group E
518 were selected for genome context analysis.

519

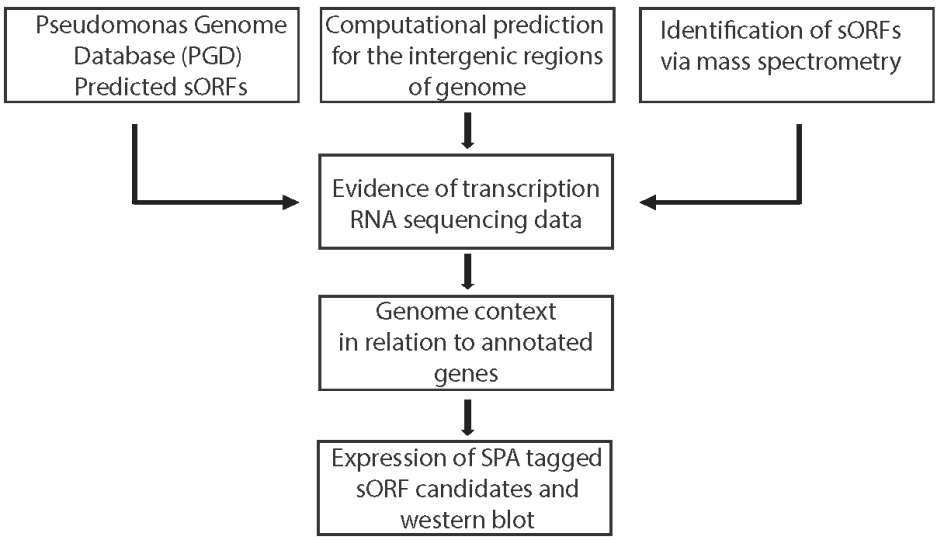
520 **Figure 3.** Validation of sORF expression. The construction of the plasmid system for
521 validation of sORF expression is depicted in panels A-D. The recombinant plasmid was
522 constructed from the pSEVA631-GFP plasmid containing a gentamicin resistance gene (Gm^R)
523 (A). The sequential peptide affinity (SPA) tag sequence was amplified from the pJL-148
524 plasmid and inserted into the pSEVA631-GFP backbone to generate pSEVA-631-GFP-SPA (B).
525 The sequences encoding the sORFs together with short upstream sequences were amplified
526 from the *P. putida* chromosome (C) and USER cloning was used to replace the Shine-
527 Dalgarno (SD) sequence and green fluorescent protein (GFP) of (B) to generate the
528 pSEVA631-sORF-SPA plasmids (D). The pSEVA631-sORF-SPA plasmids were transferred into
529 *P.putida* KT-2440, followed by cell growth, harvest at exponential phase, total protein
530 extraction, polyacrylamide gel separation, and detection of sORF-SPA fusions by western
531 blotting (E). The results of the sORF candidates with positive validation are shown. The lanes
532 are labeled with – (negative control), M (size markers with specific bands labeled with their
533 sizes in kDa) and abbreviated sORF candidate names (where P, B and M stand for hits
534 obtained from the *Pseudomonas* genome database, bioinformatic prediction or mass
535 spectrometry, respectively) corresponding to those listed in Tables 1 and 2.

Supplementary files

Supplementary file 1. Detailed experimental procedures

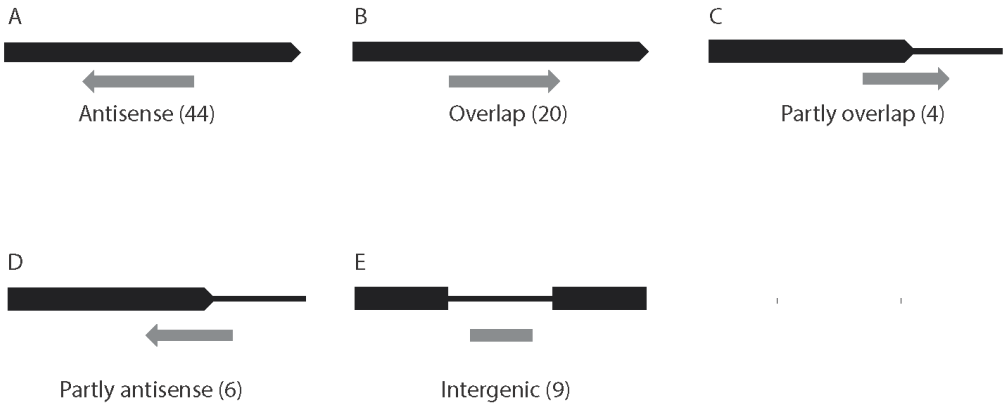
Supplementary file 2. Oligonucleotide primers used in this study

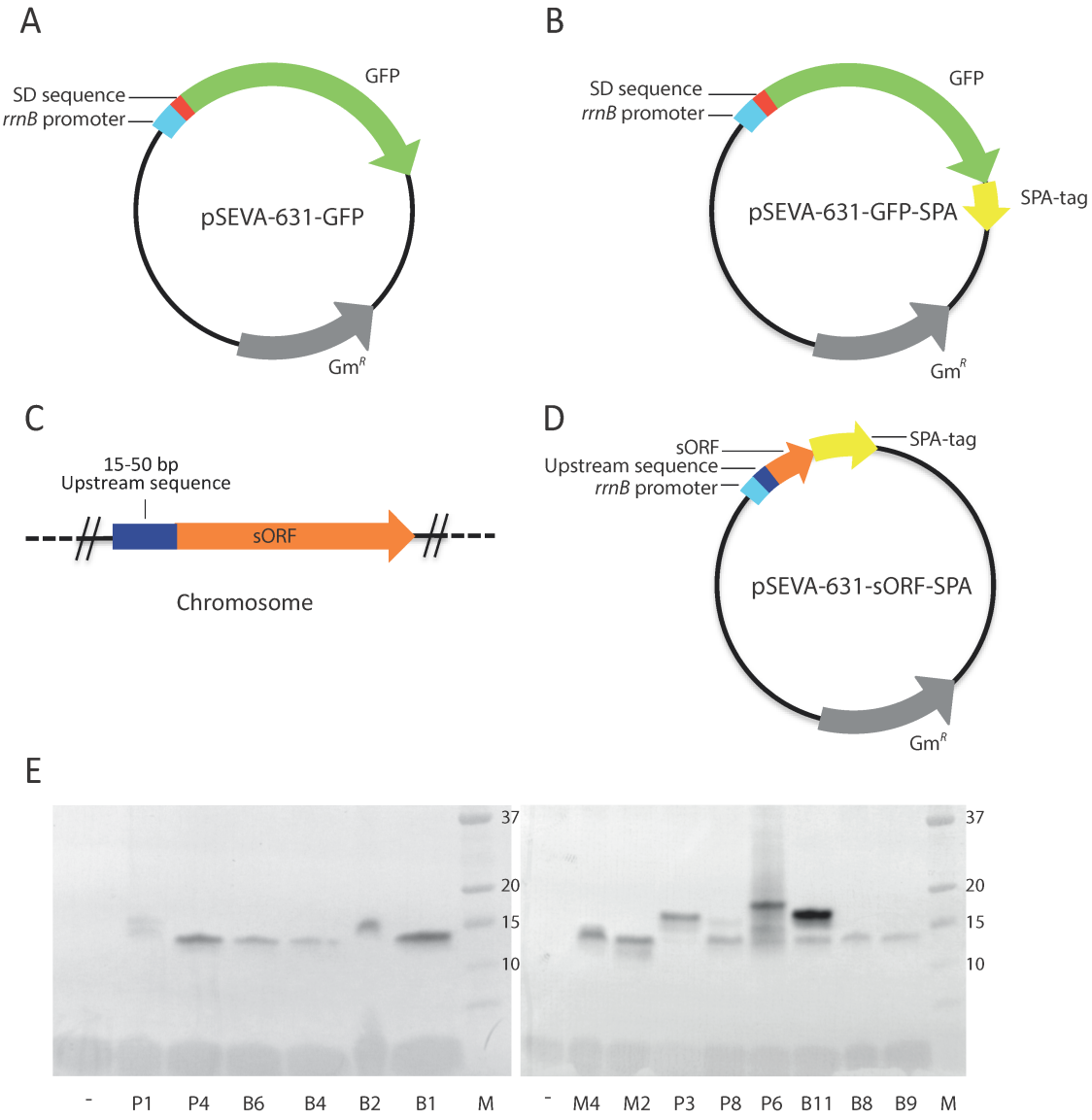
A



B

Approach	Identification	Evidence of transcription	Candidates for validation	Validated sORFs
PGD	80	31	7	5
Bioinformatics	273	125	10	7
Mass spectrometry	83	30	5	2
Total	436	186	22	14


A
Antisense (44)
B
Overlap (20)
C
Partly overlap (4)
D
Partly antisense (6)
E
Intergenic (9)


A
SD sequence
rrnB promoter
GFP
pSEVA-631-GFP
Gm^R
B
SD sequence
rrnB promoter
GFP
SPA-tag
pSEVA-631-GFP-SPA
Gm^R
C
15-50 bp
Upstream sequence
sORF
Chromosome
D
sORF
Upstream sequence
rrnB promoter
SPA-tag
pSEVA-631-sORF-SPA
Gm^R
E
37
20
15
10
- P1 P4 B6 B4 B2 B1 M
- M4 M2 P3 P8 P6 B11 B8 B9 M

S1: Detailed experimental procedures

Strains, plasmid and Growth conditions

Pseudomonas putida KT-2440 was grown in LB medium (10 g l^{-1} of tryptone, 5 g l^{-1} of yeast extract and 5 g l^{-1} of NaCl) with chloramphenicol (Cm) 30 ug ml^{-1} at 30 °C, and *Escherichia coli* XL1-blue was grown in LB medium at 37 °C. Recombinant *P. putida* KT-2440 and *E. coli* XL1-blue strains with modified pSEVA-631 plasmid were grown in LB broth with gentamicin (Gm) 10 ug ml^{-1}. The *E. coli* strain that contains pJL-148 plasmid (Zeghouf et al., 2004) was grown in LB medium with (Gm) 10 ug ml^{-1}.

Bioinformatics prediction and annotated sORFs collection

The DNA sequences of intergenic regions were collected from *P. putida* KT-2440 genome sequence. A Python program was designed to check the Shine-Dalgarno sequence feature (GGAG), start codon (ATG, CTG, TTG, GTG), and stop codon (TAA, TGA, TAG). The distance between putative SD-sequences and the start codon was limited to 20bp. The sequences that met these requirements were collected from start codon to stop codon. Hits with less than 154bp were collected as putative sORFs.

The annotated sORFs were collected from Pseudomonas Genome database (www.pseudomonas.com) with advanced annotation search function. The hypothetical annotated sORFs with equal or less than 153bp were collected.

Total protein sample preparation and mass spectrometry

P. putida cells were cultivated in 50 ml LB in 250 ml shake flasks at 30 °C with 250 rpm shaker. Cells were collected at an OD_{600} of 1.5 and 4.5. Total protein extraction was done with the ReadyPrep™ Extraction Kit according to manufacturer's instruction (Total Protein) (BIO-RAD). The total proteins were fragmented with a protein gel system (NuPAGE Novex Bis-Tris 4-10%) at 180V, 35min. Each lane was cut into 10-12 pieces from the top of the gel to bottom. The proteins from each gel piece were collected for trypsin digestion.

For Nanoscale LC separation, the supernatant from each gel lane was loaded on a nanoACQUITY TM System (Waters Corp., USA) equipped with a Symmetry C18 (5µm, 180µm x 20mm) pre-column and a nanoaAcquity BEH130 C18 (1.7 µm, 75 µm x 250 mm) analytical reversed-phase column (Waters, USA). Prior to separation samples were trapped on a pre-column. Peptides were separated using a reversed-phase gradient over 90 min going from 5–40 % acetonitrile in water with a flow rate of 250 nL.min-1 and a column temperature of 35°C. Mobile phase A consisting of 0.1 formic acid in water and mobile phase B consist of 0.1% formic acid in acetonitrile.

Data was acquired on a Synapt G2 (Waters, Manchester UK) Q-TOF instrument using ESI with a NanoLock-spray source. The mass spectrometer was operated in positive and resolution mode with continuum spectra being acquired. Data were continuously calibrated using Leucine encephalin as lock mass. Data was acquired using MSE during which the mass spectrometer alternated between low and high energy mode using a scan time of 0.8 s for each mode over a 50-2000 Da interval. In the low-energy MS mode, data was collected at constant collision energy of 4 eV. In the elevated-energy MS mode, the collision energy was increased from 15 to 40 eV.

Custom Database for MS analysis

The custom database was established by collecting all the possible sORFs sequences from genome sequence. The genome sequence of *P. putida* KT-2440 was analyzed by collecting all the sequences with start and stop codon. The sequences with less than 200bp were collected and translated into amino acids sequences. The CLC Lab-bench 7 platform was used for sequence collection and the online platform SMS (Sequence Manipulation Suite) was used for sequence translation.

Transcription analysis

RNA sequencing data was obtained from unpublished work done in our laboratory using methods described previously (Gómez-Lozano et. al, 2012). The data was analyzed with Rockhopper. The transcription of putative sORFs and the transcription levels were analyzed by checking the number of the RNA sequencing reads in the genome where the putative sORFs are located.

pSEVA-631-sORF-SPA plasmid construction and recombinant strain construction

The SPA expression vectors were created by sequential cloning of PCR products into the pSEVA-631-GFP vector. The recombinant vector was constructed with USER cloning technique. In the first step, the SPA tag fragment was introduced into the pSEVA-631-GFP plasmid to construct the pSEVA-631-GFP-SPA plasmid. The plasmid backbone was amplified with the following primers (Primers: SEVA-F: AGCGCAUACCTGCAG GCATGCAAGCT; GFP-SPA-R: ATGGAUCCTTTG TAGAGCTCATCCATGCCA) and the PCR-product was treated with DpnI for template digestion. The SPA-tag sequence was amplified from pJL-148 plasmid using (Primers: SPA-F: ATCCAUGGAAAAGAGAAGATGGAAAAAGAATT TCA. SPA-R: ATGCGCUACTTGTCATCGTCATCCTTGT) and inserted downstream of the GFP by USER cloning.

In the second step, the pSEVA-631-GFP-SPA plasmid backbone was PCR- amplified with (Primers: SPA-F:ATCCAUGGAAAAGAGAAGATGGAAAAAGAATTTCA; SEVA-R: AATTCUAGAGTGGTGTCGCATTATAGG). The linearized backbone contains the whole plasmid except GFP and its upstream SD sequence. The sORF fragment was amplified from *P. putida* cells and combined with the linearized backbone by USER cloning. . After plasmid extraction and verification by sequencing, the plasmids were transferred into competent *P. putida* cells by electroporation using protocols described previously (Choi, Kumar, & Schweizer, 2006).

Protein expression and Western blot

P. putida cells containing the different pSEVA-631-GFP-SPA vectors were cultivated in 50 ml LB in 250 ml shake flasks at 30°C with 250rpm shaker. The cells were harvested at the OD_{600}= 0.6, and cell pellets were collected by centrifugation (4000 x g, 10 min, 4 °C). The cell pellet was washed with 1 ml of ice-cold lysis buffer (20mM Tris pH 8, 150mM KCl, 1mM MgCl2, 1mM DTT) and centrifuged at 4000 x g, 10 min, 4 °C. The pellet was re-suspended in 1 ml of ice cold lysis buffer and transferred into a 2 ml tube for bead beating. 0.6-0.8 ml of glass beads and protease inhibitor (cOmplete™ ULTRA Tablets, Mini, EASYpack, Roche; 1 pinch of the tablet per tube) was added to the tube, and the cells were broken with the bead-beater (20s burst followed by chilling on ice for 1 min, for three times). The lysed cells was centrifuged at 16000 x g for 30 min at 4 °C. The cleared lysate (0.6-0.8 ml) was collected as total protein sample. The concentration of total protein samples was measured with BCA kit(Novagen).

10 ug of total protein from each samples were loaded on a Tris-Bis gel (NuPAGE Novex Bis-Tris 4-10%) under 180V for 35 min. The proteins were transferred into membrane (NC membrane from sigma) with 20V for 7 min and treated with the M2 antibody (1:4000) and secondary antibody (1:5000). The ECL reagent (Bio-Rad) was used to stain the membrane. The immunoblot bands were captured using a bio-imager (Syngene).

Reference:
Choi, K.-H., Kumar, A., & Schweizer, H. P. (2006). A 10-min method for preparation of highly electrocompetent Pseudomonas aeruginosa cells: application for DNA fragment transfer between chromosomes and plasmid transformation. *Journal of Microbiological Methods*, *64*(3), 391–7.
Gómez-Lozano, M., Marvig, R. L., Molin, S., & Long, K. S. (2012). Genome-wide identification of novel small RNAs in Pseudomonas aeruginosa. *Environmental Microbiology*, *14*(8), 2006–16.
Zeghouf, M., Li, J., Butland, G., Borkowska, A., Canadien, V., Richards, D., … Greenblatt, J. F. (2004). Sequential Peptide Affinity (SPA) System for the Identification of Mammalian and Bacterial Protein Complexes research articles, 463–468.

S2. Oligonucleotide primers used in this study

Primer	Sequence (5´to 3´)	Application
MS1-F	AGAATUTGAACGGTCAGTCACGATATG	Amplification of MS-1
MS1-R	ATGGAUCCGGACAATTCCCACATCCCAG	Amplification of MS-1
MS2-F	AGAATUCTGCAGGACGCTCCCGCGAC	Amplification of MS-2
MS2-R	ATGGAUCCCTCCCACGCAATGCGCGAC	Amplification of MS-2
MS3-F	AGAATUTGACACTTTTTGCCGGCC	Amplification of MS-3
MS3-R	ATGGAUCCTGACAGGTAGGCATTTGCCTG	Amplification of MS-3
MS4-F	AGAATUCCTATGAGAGGAAGTTCCGCCA	Amplification of MS-4
MS4-R	ATGGAUCCTCTTCGCCCTAAAATCTCGTCC	Amplification of MS-4
MS11-F	AGAATUCCGTCTCTCGAAAATCGCTAATGG	Amplification of MS-11
MS11-R	ATGGAUCCTGAATTGCCGGTCGCGGCCATGTAC	Amplification of MS-11
BIO1-F	AGAATUCTTGCGAAGGGAGGTGACG	Amplification of BIO-1
BIO1-R	ATGGAUCCTCCGGCCCCTGCGCTACG	Amplification of BIO-1
BIO2-F	AGAATUACAGGAGCTTTATCAAGATGCATG	Amplification of BIO-2
BIO2-R	ATGGAUCCGTCGACTGCCACCTGGG	Amplification of BIO-2
BIO4-F	AGAATUTGTGGGAGTAACGGCCCTT	Amplification of BIO-4
BIO4-R	ATGGAUCCGGGCGGCCTGTAGGAGC	Amplification of BIO-4
BIO5-F	AGAATUTGGTGGAGAAGTAGGCAAGGATG	Amplification of BIO-5
BIO5-R	ATGGAUCCCATGACCGGTGCCTGGC	Amplification of BIO-5
BIO6-F	AGAATUACTTCAGGAGGCCGAGCGTAG	Amplification of BIO-6
BIO6-R	ATGGAUCCCGGTCTTGCTCCCGGGA	Amplification of BIO-6
BIO7-F	AGAATUTGG CGG AGG CGC AGC TAA	Amplification of BIO-7
BIO7-R	ATGGAUCCGGCGCTGCTTTTAGTGCAGTTTTTG	Amplification of BIO-7
BIO8-F	AGAATUACTTCAGGAGGCCGAGCGTA	Amplification of BIO-8
BIO8-R	ATGGAUCCCGCTCCGGGGTTCCCTCGCT	Amplification of BIO-8
BIO9-F	AGAATUTCACTGGAGAGCAACATGCC	Amplification of BIO-9
BIO9-R	ATGGAUCCGTGGCTAAACCAATACCCAAAGTAAT AGCTG	Amplification of BIO-9
BIO10-F	AGAATUAACACGGGGAGATCCACACATGAACAT	Amplification of BIO-10
BIO10-R	ATGGAUCCGCTGCGATCGGCGCGCAGCA	Amplification of BIO-10
BIO11-F	AGAATUTCAGGGAGCGACAGTCAGG	Amplification of BIO-11

BIO11-R	ATGGAUCCTGGAGAACACGAGATCATGACTATTGGT	Amplification of BIO-11
PGD1-F	AGAATUATTCGGAAGGAGTAATTCCATGTG	Amplification of PP_0380
PGD1-R	ATGGAUCCGCGGTTAGCGAAGTACATGGT	Amplification of PP_0380
PGD2-F	AGAATUTTTACAGGCATATTGCGCCCGT	Amplification of PP_0419
PGD2-R	ATGGAUCCGGCGCGCCAACGCCAAC	Amplification of PP_0419
PGD3-F	AGAATUACGAATGGGGAGGCCAGC	Amplification of PP_1149
PGD3-R	ATGGAUCCGCGCGGCTTGCGCTTGTTCG	Amplification of PP_1149
PGD4-F	AGAATUAACGTCCGGTTTCACGCC	Amplification of PP_1685
PGD4-R	ATGGAUCCACACACACTCATGCCCAGTTG	Amplification of PP_1685
PGD5-F	AGAATUAACACCGGTCCCGACACC	Amplification of PP_2644
PGD5-R	ATGGAUCCAGGTAGCCGTGCAGACGC	Amplification of PP_2644
PGD6-F	AGAATUTTGTCGGAGGCTCTTATTCTGGCCTC	Amplification of PP_0023
PGD6-R	ATGGAUCCCTCCCCCTTTTTGTGCTTGTGGGGTTT	Amplification of PP_0023
PGD8-F	AGAATUCGTCCGAAGGATCAAGCCC	Amplification of PP_3434
PGD8-R	ATGGAUCCTTGATTGCGTGGGATGTGGAACAGGAAC	Amplification of PP_3434
SEVA-F	AGCGCAUACCTGCAGGCATGCAAGCT	Amplification of pSEVA631-GFP backbone
GFP-SPA-R	ATGGAUCCTTTGTAGAGCTCATCCATGCCA	Amplification of pSEVA631-GFP backbone
SEVA-R	AATTCUAGAGTGGTGTCGCATTATAGG	Amplificaiton of pSEVA631-GFP-SPA backbone
SPA-F	ATCCAUGGAAAAGAGAAGATGGAAAAAGAATTTCA	Amplification of SPA tag and pSEVA631-GFP-SPA backbone
SPA-R	ATGCGCUACTTGTCATCGTCATCCTTGT	Amplification of SPA tag
SEVA_seq_F	TCTAGGGCGGCGGATTTG	Insertion check and DNA sequencing primer
SEVA_seq_R	CCGAGCGTTCTGAACAAATC	Insertion check and DNA sequencing primer

Manuscript 2

Quantitative analysis of the cytosolic proteome of *Pseudomonas putida* under stress conditions and in different nutrient environments.

Xiaochen Yang, Daniela Zühlke , Katharina Maria Riedel , Katherine S. Long

Quantitative analysis of the cytosolic proteome of *Pseudomonas putida* under stress conditions and in different nutrient environments.

Xiaochen Yang [1], Daniela Zühlke [2], Katharina Maria Riedel [2], Katherine S. Long [1]

1. Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Hørsholm, Denmark

2. Institute of Microbiology, Ernst-Moritz-Arndt-University, Greifswald, Germany

Abstract:

Pseudomonas putida is a Gram-negative bacterium that is tolerant to many antibiotics and environmental stresses. To investigate the mechanisms underlying stress responses and adaptation to different nutrient conditions, a label-free quantitative proteomic approach was designed to analyze the variations of cytosolic protein levels in *P.putida* KT-2440 under two stress conditions, H_2O_2 stress and imipenem stress, and two media, M9 minimal medium and complex LB medium. A total number of 1,302 proteins were identified via mass spectrometry. The results suggested that the DNA replication, recombination, and repair systems were highly activated under both H_2O_2 and imipenem stresses. Regulators involved in the general stress response (Hfq, RpoS) were also upregulated under both stress conditions. In addition, Different ABC transporters were involved in the response of both stresses. The components of flagella were up-regulated under imipenem stress. Proteins involved in detoxification of ROS were exclusively identified under H_2O_2 stress, including AhpF and KatA. Growth in LB medium led to higher amounts of proteins involved in protein synthesis, e.g. ribosomal proteins and tRNA synthetases while the pathways for amino acid biosynthesis were downregulated in LB compared to M9 minimal medium. Proteins that were involved in stress response were downregulated as well, while enzymes responsible for cell division DNA metabolism, fatty acid metabolism and phospholipid metabolism were accumulating in LB medium, indicating that more energy and material resources of cells were used for cell growth and division in LB medium.

1. Introduction:

The energy and material exchange between cells and environment is indispensable for the survival of living organism. Cells uptake nutrients and other necessary compounds from the

environment, and at the same time, suffer the toxic stress coming from environment. One of the common stresses is oxidative stress. The oxidative stress comes from reactive oxygen species (ROS) that widely exist in the environment, such as superoxide radicals, hydrogen peroxide and hydroxyl radicals. The peroxides and free radicals generated from ROS lead to the damage of all cell components, including DNA, proteins and lipids (Nyska, 2002).
For bacteria, another common stress is antibiotics. Antibiotics are initially produced by microbes and are used to kill or inhibit the growth of other microbes. Different from oxidative stress, antibiotics specifically block certain metabolic pathways. Depending on the type of antibiotic, they can be block synthesis of cell wall, nucleic acids, or proteins.

In order to survive in the stress environment, bacteria have evolved a series of systems for stress resistance. For example, the signal transduction system is used to identify the stress and activate the response of cells to the stress (Altuvia, 1997; McCullen, 2010). The repair system, such as DNA repair system, is used to identify and repair the damaged molecules in the cell (Webb, 1997; Wigley, 2007). In many multidrug resistance bacteria, the transporter system is used to extrude the toxic compounds out of cells (Marquez, 2005). In addition, many genes are involved in response to different nutrient the cells uptake (Molina-Henares, 2010).

Pseudomonas putida is a Gram-negative bacteria demonstrating a versatile metabolism and multidrug resistance (Nikel, 2014). In this study, a quantitative proteomic approach was used to analyze the protein expression of *P.putida* under H_2O_2 and imipenem stresses, as well as 2 media with different nutrient compositions, M9 and LB. The results yielded insights into the response of *P.putida* to environmental stresses and different nutrient sources.

2. Materials and methods:

2.1 Bacteria and cultivation conditions

Pseudomonas putida KT-2440 single colony was grown in LB medium (10 g l^{-1} of tryptone, 5 g l^{-1} of yeast extract and 5 g l^{-1} of NaCl) or in M9 minimal medium with 0.5% glucose and chloramphenicol (Cm) 30 ug ml^{-1} at 30 °C overnight. The next day 50ml of LB or M9 medium

without antibiotics were inoculated with an aliquot of the overnight culture to get an OD_{600} xxxnm of 0.05. The cells were cultured until they reached an OD_{600} and then the stress was applied. (either treating with 0.5 mM of H_2O_2 for 7 minutes or 0.1 ug/ml of imipenem for 1 hour). The concentration and time of treatment of stress compound was determined by CFU counting described in another study (Klara`s data).

2.2 Cytosolic protein extraction

50ml of the culture were centrifuged at 4°C 8.500 rpm for 10 min. The pellet was washed twice with 1ml of TE buffer (10 mM Tris, 1 mM EDTA, pH 7.5 – 8.0) at 4°C and re-suspended with 600ul of TE buffer. 400ul of glass beads (sigma, 150-212um) were added to the cell suspension and cells were disrupted using a Precellys®24 homogenizer with 6500rpm for 20s for three times. The cell lysate was centrifuged at 15.000 rpm for 30 minutes and the supernatants containing cytosolic proteins were transferred into new tubes.

2.3 Sample preparation for MS analysis

The concentrations of the cytosolic protein extracts were measured by Bradford assay using Roti Nanoquant. In this method, 1ul protein sample was mixed with 199ul A.bidest reagent and 800 ul Roti Nanoquant reagent. The solution was incubated at room temperature for 5 minutes and the ratio A_{590nm}/A_{450nm}, was measured. The concentration of protein was calculated based on the formula:

$$Protein(ug) = \frac{\frac{A590nm}{A450nm} - \frac{1}{2,6}}{0,054945}$$

The amount of 100 µg of cytosolic protein extract were combined with 5 µl 1M triethylammonium bicarbonate (TEAB; Sigma Aldrich) and 20 µl RapiGest (final concentration 0.5%; Waters) and filled to a final volume of 100 µl with distilled water in MS quality. 1 µl 500 mM tris(2-carboxylethyl) phosphine hydrochlorid (TCEP; Invitrogen) was added to each sample and incubated for 45 min at 60 °C to reduce disulphide bonds in proteins. Afterwards 2 µl 500

mM iodoacetamid (IAA; Sigma Aldrich) were added and incubated for 15 min at room temperature in the dark to block reduced cysteine residues. In the meantime Trypsin was activated by incubation with trypsin-activation buffer (1 mg/ 100 µl; Promega) for 15 min at 37 °C. 2.5 µl of activated Trypsin were added per 100 µg protein extract and incubated for 5 h at 37 °C and 900 rpm in a thermoshaker. 1-2 µl triflouracetic acid (TFA; AppliChem) was added to precipitate the RapiGest; the samples were incubated for 45 min at 37 °C, centrifuged three times for 10 min at 4 °C and 15,000 rpm and the supernatant transferred into a new reaction tube. Peptides were desalted and purified using StageTips (Thermo Scientific). StageTips were filled with 5 µl of C18 material, washed three times with 100 µl buffer A (0.1 % acetic acid; centrifuge for 1 minute at 9,000 rpm at room temperature) and twice with buffer B (0.1 % acetic acid in acetonitrile). The StageTip was washed again with 100 µl of buffer A and the digested protein extract added to the StageTip. After centrifugation the bound peptides were washed twice with buffer A. Peptides were eluted into glass vials using 30 µl of buffer B, 10 µl of high-pure distilled water were added and the samples concentrated to 5 µl in a vacuum centrifuge. 5 µl of tryptic digest of yeast alcohol dehydrogenase (Waters) were added to give a final concentration of 50 fmol/µl and filled up to 100 µl.

2.3 LC-IMS^{E} setup and data analysis

The nanoACQUITY™ UPLC™ system (Waters, Milford, MA, USA) was used to separate the peptide mixture and to introduce the samples into the mass spectrometer. The peptide mixture was directly loaded on an analytical column (nanoACQUITY™ UPLC™ column, BEH300 C18, 1.7 µm, 75 µm_200 mm, Waters) within 31 min with 99% buffer A (1% acetic acid) and a flow rate of 300 nl/min. Separation of peptides for ion mobility MS was done with a 90 min gradient from 5% buffer B to 40% buffer B. The reference peptide [Glu1]-Fibrinopeptide B (50 fmol/µl; Sigma Aldrich) was constantly infused at a flow rate of 500 nl/min and scans were acquired at an interval of 30 sec.

All MS analyses were performed on a Waters Synapt^{TM} G2 HDMS instrument equipped with an ESI source (NanoLockSpray) in positive ion mode with the analyzer mode set to resolution. Mobility-TOF mode was employed for analysis of peptide ions with the following parameters: Spectra were acquired in a mass range of 50 to 2,000 Da and a scan time of 1 sec. Low transfer

collision energy (CE) for analysis of precursor ions was set to 4 V, high transfer CE was ramped from 25 – 45 V to generate fragment ions. Wave velocity was ramped from 1,000 – 400 m/s, wave height was set to 40 V.

Data collection was controlled by MassLynx 4.1 software. LC-IMS^E data were processed using the PLGS software 3.0.1 and processing parameter were set as follows: Chromatographic peak width and MS TOF resolution were set to automatic, lock mass charge 2 set to 785.8426 Da/e with a lock mass window of 0.25 Da, low energy threshold 200.0 counts, elevated energy threshold 20.0 counts, intensity threshold 750 counts. The processed data were searched against a randomized database containing the protein sequences of *Pseudomonas putida* KT2440, translated sequences of *P. putida* KT2440 that are flanked by start- and stopcodons and were not present so far in the database. In addition, common laboratory contaminants and the yeast ADH1 sequence were included (in total 92,913 entries). Search parameters were set as follows: Trypsin was used as digestion enzyme, two missed cleavages allowed, fixed modification carbamidomethylation (Cys), variable modification deamidation (Asn, Gln), oxidation (Met), N-terminal pyrrolidone carboxylic acid. For positive protein identification the following criteria had to be met: 1 fragment ion matched per peptide, 5 fragment ions matched per protein, 1 peptide matched per protein; 2 missed cleavages allowed. The protein false discovery rate (FDR) was set to 5 %. Three biological replicates were analysed for each growth condition, each sample was measured in triplicates. For further analysis only proteins that were present in 2 out of 3 technical replicates which led to a FDR below 1% were considered.

Statistical analysis was performed using MeV v4.8.1 (Saeed, 2003). T-test was performed with the following settings: unequal group variances were assumed (Welch approximation), P-values based on all permutations with P=0.01, significance determined by adjusted Bonferroni correction. Functional prediction and classification of proteins was performed by the analysis pipeline *Prophane 2.0* (Schneider, 2011;) (http://www.prophane.de). Voronoi treemaps were generated using Paver (Decodon, Greifswald, Germany; http://www.decodon.com/).

3. Results

In this study, proteome of *P. putida* KT2440 under H_2O_2 stress, imipenem stress, LB medium and M9 medium were analyzed with a quantitative proteomic MS approach. This approach was an absolute quantification approach and the quantitative analysis was based on the comparison of absolute quantity of proteins among different samples. In data analysis, the proteome of M9 sample was used as control. The protein ratios between each of other three sample to M9 sample was calculated for quantitative analysis.

3.1 Overview of gene expression in different media and stress conditions

In this study, three biological replicates for each condition were prepared for testing the reproducibility of results. Only proteins that were detected in at least two replicates for each sample were included in the results. A total number of 1,302 proteins was identified and quantified under 4 different conditions, including 874 from LB samples, 911 from M9 sample, 1,028 from oxidative stress sample and 1,152 from imipenem stress sample.

A total of 666 proteins, corresponding to around 50% of total quantified proteins, were identified in all conditions. The proteome pattern of the LB sample shows significant differences with that of the M9 sample. Although 699 proteins were identified in both culture conditions, 212 (23 %) and 171 (20 %) proteins were detected only in the M9 and LB samples, respectively (Figure 1).

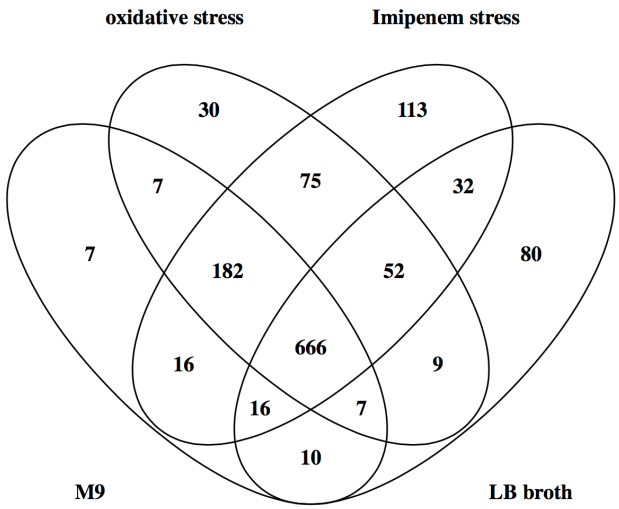


Figure 1: Comparison of cytosolic proteome composition of *P.putida* KT-2440 under four cultivation conditions. M9: Cultivation in M9+glucose medium; LB broth: Cultivation in LB medium; oxidative stress:

Cultivation of *P.putida* in M9+glucose medium with H_2O_2 treatment; Imipenem stress: Cultivation of *P.putida* in M9+glucose medium with imipenem treatment.

Compared stress conditions with M9 control, more than 96% and 94% of proteins identified in the M9 control sample were detected in oxidative-stress and imipenem samples (880 proteins for oxidative-stress samples and 862 proteins for Imipenem samples). However, more proteins were detected in stress-treated samples, indicating that expression of groups of genes was induced by stressor compounds.

.

In total 975 proteins were identified under both oxidative stress and imipenem stress conditions. This amounted to 85% of identified proteins in imipenem-treated samples and 95% of expressed genes in oxidative-stress samples. Seventy-five proteins were expressed only in both stress conditions, indicating response pathways for multiple stress conditions. The numbers of proteins identified exclusively in oxidative stress and imipenem-treated samples were 30 and 113, respectively, indicating specific responses towards each stress.

The proteome of M9 sample was used to compare with that from other three samples as a control. The differentially expressed genes in stress conditions and LB medium were determined by checking the up or down regulation of each protein in stress-treated sample and LB sample against that in the M9 sample. The proteins that had greater than 2 times fold changes were selected for further analysis (Figure 2). There were 53 and 71 differentially expressed proteins from H_2O_2 treated samples and imipenem treated samples, respectively. A total number of 155 differential expressed proteins were identified from LB samples and more than one third of them (58 proteins) were involved in energy metabolism, protein synthesis and amino acid synthesis. This suggests that the nutrient sources largely affected the material and energy metabolism of the bacterium.

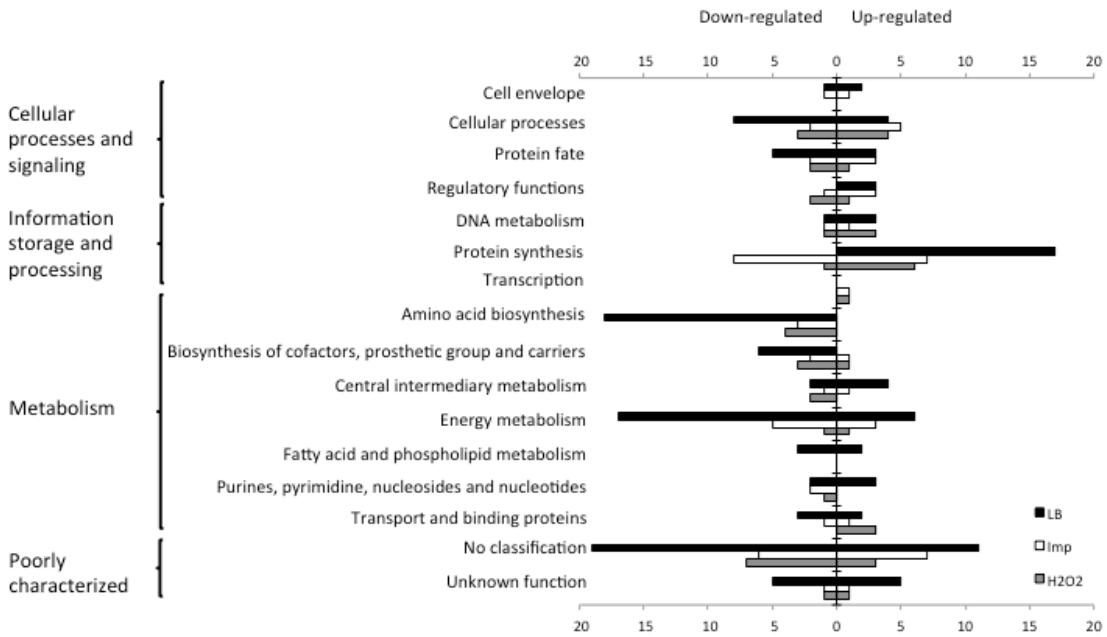


Figure 2: Impact of stress conditions and nutrient availability on the cytosolic proteome composition of *P.putida* KT-2440. The numbers of proteins with at least twofold changes under stress conditions and LB medium compared to the unstressed control in M9 medium are listed by functional classification. Functional classification is based on Tigrfams.LB: Cultivation of *P.putida* in LB medium; Imp: Cultivation of *P.putida* in M9 medium with imipenem stress; H_2O_2: Cultivation of *P.putida* in M9 with H_2O_2 stress.

3.2 Proteome analysis of H_2O_2-stressed *P.putida* KT-2440

From the 53 proteins with significantly different amounts, 25 were up-regulated and 28 were down-regulated by a factor of at least 2 under H_2O_2-stress compared to the M9 control. In addition, 166 genes were exclusively identified under imipenem stress condition. (Figure 3A, Supplementary S-1). The expression of several proteins that play key roles in stress resistance and DNA repair were up-regulated in H_2O_2-stress. One up-regulated protein was RecQ, which is involved in the RecF recombinational DNA repair pathway. RecQ helicase is involved in the repair of both double-stranded DNA breaks (DSBs) and single-stranded DNA gaps. It initiates the DNA repair reaction by unwinding the DNA strands at damage ends (Bernstein, 2010; Morimatsu, 2014; Persky, 2008). . In addition, several ABC transporters, including iron ABC transporter

PP_5196 (NP_747297.1_5127), were upregulated in the H_2O_2 stressed cells condition. The peroxidase LsfA, and bacterioferritin Bfr were also up-regulated.

The H_2O_2 stress affected the DNA transcription and protein synthesis of bacterium. A number of transcription factors, including Fis family transcriptional facoters such as PP_0888(NP_743049.1_879), PP_2790 (NP_744934.1_2764), and TetR family transcriptional facoters PP_1968 (NP_744119.1_1949), PP_2806 (NP_744950.1_2780), and ribosome proteins (e.g. RpmB, Rpsl, RpmF) were up-regulated. The biosynthesis of cofactors, prosthetic groups, and carriers was also affected by H_2O_2 stress as well. Under H_2O_2 stress, enzymes involved in the biosynthesis of thiamine, menaquinone and ubiquinone were upregulated while enzymes of the biosynthesis of biotin were downregulated.
Two proteins that were known to be involved in detoxification of ROS were identified under H_2O_2 stress. They were alkyl hydroperoxide reductase AhpF and catalase KatA that were involved in the removal of peroxide from the cell.

Two proteins that were known to be response for wide range of stresses were identified under two stress conditions. RpoS (NP_743780.1_1610), a regulator of transcription, were almost 8 fold up-regulated under both stress conditions. Hfq (NP_746997.1_4827), an sRNA binding protein was also highly up regulated under both stress conditions as well as LB.

3.3 Proteome analysis of imipenem-treated *P.putida* KT-2440

The comparison of the cytosolic proteins of imipenem-treated cells with the untreated control reveals a total number of 71 proteins exhibiting changes in abundance of at least 2-fold. This includes 35 up-regulated and 36 down-regulated proteins. In addition, 272 genes were exclusively identified under imipenem stress condition. (Figure 3:B, supplementary S-2)

Similar with the proteome under H_2O_2 stress, the expression of ABC transporter components was up-regulated. Genes that were involved in DNA replication, recombination and repair were highly expressed. The expression of many transcriptional regulators and the biosynthesis of some cofactors and prosthetic groups, such as thiamine, menaquinone and ubiquinone, were also activated under H_2O_2 stress condition.

A group of genes encoding components of flagellar were highly expressed, including, flgL and fliD, encoding hook associated protein, fliG and fliN, encoding motor switch protein, flhF and fliS, encoding flagella biosynthesis protein, cheA, involving in signal transduction (Fedi, 2016; Macnab, 2003) and three genes annotated as methyl-accepting chemotaxis sensory transducer (Nelson, 2002). Except cheA gene, which had 1.5-2 fold up-regulated, other genes in imipenem treated sample had more than two fold changes in expression.indicating the high mobility of bacterium under imipenem stress.

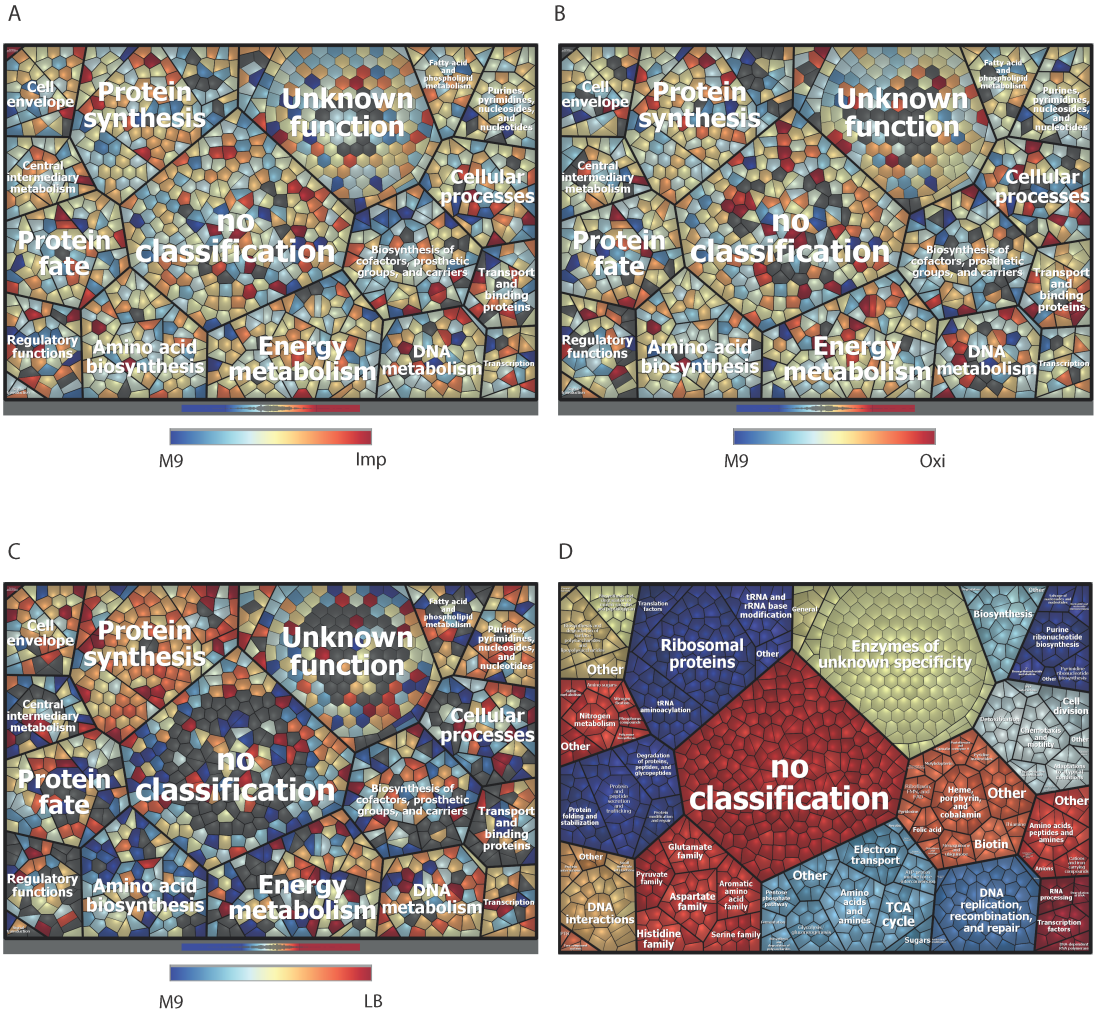


Figure 3: Functional classification of proteins identified by MS based on TIGRFAMs. In treemaps A, B and C, down-regulated genes are labeled with blue color and up-regulated genes are labeled with red color. The darkness of color is related with the fold change of expression. Genes with grey color were not quantified in the respective conditions. Treemap A: Comparison of gene expression under M9 and imipenem stress; Treemap B: Comparison of gene expression under M9 and H_2O_2 stress; C: Comparison of gene expression under M9 and LB medium. D: Sub-classification on functions of proteins.

3.4 Proteome analysis of *P.putida* KT-2440 in LB medium

From the proteins that were identified in both LB and M9 medium 155 exhibited differential expression. There were 65 up-regulated and 90 down-regulated proteins identified in the LB culture compared to the M9 control. In addition, 173 proteins were exclusively identified in LB medium while 212 proteins were exclusively identified in M9 medium (Figure 3:C, Supplementary

S-3). The major differences of protein expression were observed in metabolism, including energy metabolism and central intermediary metabolism, as well as molecule synthesis, such as amino acids biosynthesis and protein synthesis.

Compared with the proteome in the M9 condition, amino acid synthesis was down-regulated while the protein synthesis was up-regulated. Different from the stress-treated samples, not only ribosome proteins, but also genes related with tRNA aminoacylation, and tRNA and rRNA base modification were up-regulated in LB medium. This suggested that the protein synthesis rate of the cell in LB was faster than that in the other conditions.

The two proteins MinC and MinD were present in lower amounts under LB medium. The MinC and MinD are cell division inhibitors and block formation of the polar Z ring septums during cell division (de Boer, 1991; Li, 2012). In addition, stress resistance proteins, such as SodB, LsfA, KatA, were down-regulated. The components of ABC transporter system were down-regulated in LB medium. These indicated that more material and energy resources might be used for cell growth than stress resistance and compound secretion in LB condition.

4. Discussion

Protein expression of bacteria was largely affected by environment. In the present study the protein expression pattern of stressed cells (H_2O_2 and the cell-wall targeting antibiotic imipenem) and cells grown under different nutrient conditions were analyzed. The proteome data showed that the stress factors significantly affected the expression level of many key proteins involved in biological processes of *P. putida* KT2240. However, the use of different media led to even more dramatic changes in the proteome pattern compared to oxidative or antibiotic stress conditions.

4.1 Material transportation under stress conditions

ATP-binding cassette transporter systems (ABC transporter system) widely exist in prokaryotes and eukaryotes(Jones, 2004). The function of ABC transporter system is to exchange material, such as ions, amino acids, sugars, lipids, etc. between cells and the environment. In bacteria, ABC transporter effluxers and exporters are responsible for pumping certain substances, such as surface components of the bacterial cell, drugs, toxins, etc. out of cells (Davidson, 2004). The ABC

transporter contributes to the resistance of bacteria to antibiotics and environmental toxins by extruding the toxic compounds from the cell (Davidson, 2008). In this study, the genes encoding components of ABC transporters were highly expressed under two stress conditions, indicating that more ABC transporter molecules were synthesized under stress conditions and the ABC tranporter system were involved in response of both stresses.

A gene that encoded iron ABC transporter substrate-binding protein, PP_5196, was highly expressed in H_2O_2 stress but significantly down regulated in the imipenem stress condition. In addition, genes that encoded Bacterioferritin (bfr and PP_0482) were highly expressed only in the H_2O_2 stress condition. Studies in *E.coli* and *Pseudomonas aeruginosa* suggested that bacterioferritin was used to oxidize toxic Fe^{2+} iron to insoluble Fe^{3+} iron using O_2 and H_2O_2, and store the Fe^{3+} in the molecules (Orino, 2001; Rui, 2012; Weeratunga, 2009). The high expression level of bacterioferritin and iron ABC transporter substrate-binding protein indicated that the iron was involved in the response of H_2O_2 stress.

4.2 Cell mobility

In this study, the synthesis of flagella was highly activated in imipenem stress condition. A recent study on *P.putida* concluded that compared with wild-type strain, the knock-out of flagella related genes saved 30% of energy on ATP levels (Martínez-García, 2014). Due to the fact that the assembling and movement of flagella is energy consuming, the synthesis of flagella might be a prerequisite for responding to and coping with imipenem stress. In the proteome of the imipenem treated sample, groups of components of flagella were up regulated. The fleN gene, which regulated flagella synthesis negatively (Dasgupta, 2000), had more than two fold up-regulated expression under H_2O_2 stress. However, the fleN gene did not show significant changes in expression level under the imipenem stress condition. This indicated that flagella synthesis was inactivated under H_2O_2 stress but highly induced under imipenem condition. Considering the stress pressure of imipenem on cell wall synthesis of the bacterium, the highly activity of flagellar synthesis might be used for structural repair rather than promoting motility.

4.3 RpoS /Hfq system General stress response

Sigma S (RpoS) is a sigma factor that regulates the transcription of genes in bacteria. The *rpoS* gene is expressed in the late exponential phase and primarily regulates the expression of genes in stationary phase. The RpoS protein also regulates the expression of stress resistance genes and its expression can be activated by external stress conditions, such as oxidative stress, UV-radiation, pH, and temperature (Kim, 2014). In *E.coli*, the translation of RpoS under stress conditions is controlled by several small regulatory RNAs (sRNAs), including *rprA*, *rcsC* and *oxyS* (Altuvia, 1997; McCullen, 2010; Repoila, 2003; Sledjeski, 1996). A RNA chaperone, Hfq was reported to be involved in the *rpoS* regulation under stress conditions by activating sRNAs *rprA* and *rcsC*. The Hfq protein binds with sRNAs as well as tRNAs (Lee, 2008). The mutation of *hfq* in *P.putida* largely impacts the stress resistance and cell patterns of the bacterium (Arce-Rodríguez, 2015). In this study, the Hfq protein was significantly higher expressed under H_2O_2, imipenem stress as well as LB medium. The significantly higher amount of RpoS can only be found under the stress conditions. This indicated that rpoS was regulated by hfq under stress condition, and hfq might not be involved in the rpoS regulation under the condition without stress factors.

Acknowledgements

We thank Katharina Riedel`s group from Institute of Microbiology, Ernst-Moritz-Arndt-University, Greifswald, Germany for MS analysis. This work is supported by the Novo Nordisk Foundation and a PhD grant from the People Programme (Marie Curie Actions) of the European Union Seventh Framework Programme [FP7-People-2012-ITN], under grant agreement no. 317058,"BACTORY".

Reference

Altuvia, S., Weinstein-Fischer, D., Zhang, A., Postow, L., & Storz, G. (1997). A small, stable RNA induced by oxidative stress: Role as a pleiotropic regulator and antimutator. *Cell*, *90*(1), 43–53.

Arce-Rodríguez, A., Calles, B., Nikel, P. I., & de Lorenzo, V. (2015). The RNA chaperone Hfq enables the environmental stress tolerance super-phenotype of *Pseudomonas putida*. *Environmental Microbiology*, n/a–n/a. doi:10.1111/1462-2920.13052

Bernstein, K. a, Gangloff, S., & Rothstein, R. (2010). The RecQ DNA helicases in DNA repair.

Annual Review of Genetics, *44*, 393–417.

Dasgupta, N., Arora, S. K., & Ramphal, R. (2000). fleN, a gene that regulates flagellar number in Pseudomonas aeruginosa. *Journal of Bacteriology*, *182*(2), 357–364.

Davidson, A. L., & Chen, J. (2004). ATP-binding cassette transporters in bacteria. *Annual Review of Biochemistry*, *73*(1), 241–68.

Davidson, A. L., Dassa, E., Orelle, C., & Chen, J. (2008). Structure, function, and evolution of bacterial ATP-binding cassette systems. *Microbiology and Molecular Biology Reviews : MMBR*, *72*(2), 317–64.

de Boer, P. a, Crossley, R. E., Hand, a R., & Rothfield, L. I. (1991). The MinD protein is a membrane ATPase required for the correct placement of the Escherichia coli division site. *The EMBO Journal*, *10*(13), 4371–4380.

Fedi, S., Triscari Barberi, T., Rosaria Nappi, M., Sandri, F., Booth, S., J. Turner, R., … Zannoni, D. (2016). The Role of <i>cheA</i> Genes in Swarming and Swimming Motility of <i>Pseudomonas pseudoalcaligenes</i> KF707. *Microbes and Environments*, *00*(0), 0–3.

Jones, P. M., & George, A. M. (2004). The ABC transporter structure and mechanism: Perspectives on recent research. *Cellular and Molecular Life Sciences*.

Kim, J., & Park, W. (2014). Oxidative stress response in Pseudomonas putida. *Applied Microbiology and Biotechnology*, *98*(16), 6933–6946.

Lee, T., & Feig, A. L. (2008). The RNA binding protein Hfq interacts specifically with tRNAs. *RNA (New York, N.Y.)*, *14*(3), 514–23.

Li, G., & Young, K. D. (2012). Isolation and identification of new inner membrane-associated proteins that localize to cell poles in Escherichia coli. *Molecular Microbiology*, *84*(2), 276–295.

Macnab, R. M. (2003). How bacteria assemble flagella. *Annual Review of Microbiology*, *57*(1), 77–100.

Marquez, B. (2005). Bacterial efflux systems and efflux pumps inhibitors. *Biochimie*. doi:10.1016/j.biochi.2005.04.012

Martínez-García, E., Nikel, P. I., Chavarría, M., & de Lorenzo, V. (2014). The metabolic cost of flagellar motion in Pseudomonas putidaKT2440. *Environmental Microbiology*, *16*(1), 291–303.

McCullen, C. A., Benhammou, J. N., Majdalani, N., & Gottesman, S. (2010). Mechanism of positive

regulation by DsrA and RprA small noncoding RNAs: Pairing increases translation and protects rpoS mRNA from degradation. *Journal of Bacteriology*, *192*(21), 5559–5571.

Molina-Henares, M. A., de la Torre, J., Garc??a-Salamanca, A., Molina-Henares, A. J., Herrera, M. C., Ramos, J. L., & Duque, E. (2010). Identification of conditionally essential genes for growth of Pseudomonas putida KT2440 on minimal medium through the screening of a genome-wide mutant library. *Environmental Microbiology*, *12*(6), 1468–1485.

Morimatsu, K., & Kowalczykowski, S. C. (2014). RecQ helicase and RecJ nuclease provide complementary functions to resect DNA for homologous recombination. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(48), E5133–42.

Nelson, K. E., Weinel, C., Paulsen, I. T., Dodson, R. J., Hilbert, H., Santos, V. A. P. M., … Fraser, C. M. (2002). Complete genome sequence and comparative analysis of the metabolically versatile Pseudomonas putida KT2440, *4*, 799–808.

Nikel, P. I., Martínez-García, E., & de Lorenzo, V. (2014). Biotechnological domestication of pseudomonads using synthetic biology. *Nature Reviews. Microbiology*, *12*(5), 368–79.

Nyska, A., & Kohen, R. (2002). Oxidation of biological systems: oxidative stress phenomena, antioxidants, redox reactions, and methods for their quantification. *Toxicologic Pathology*, *30*(6), 620–650.

Orino, K., Lehman, L., Tsuji, Y., Ayaki, H., Torti, S. V, & Torti, F. M. (2001). Ferritin and the response to oxidative stress. *The Biochemical Journal*, *357*(Pt 1), 241–7.

Persky, N. S., & Lovett, S. T. (2008). Mechanisms of recombination: lessons from E. coli. *Critical Reviews in Biochemistry and Molecular Biology*, *43*(January), 347–370.

Repoila, F., Majdalani, N., & Gottesman, S. (2003). Small non-coding RNAs, co-ordinators of adaptation processes in Escherichia coli: The RpoS paradigm. *Molecular Microbiology*. doi:10.1046/j.1365-2958.2003.03454.x

Rui, H., Rivera, M., & Im, W. (2012). Protein dynamics and ion traffic in bacterioferritin. *Biochemistry*, *51*(49), 9900–9910.

Saeed, A. I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., … Quackenbush, J. (2003). TM4: A free, open-source system for microarray data management and analysis. *BioTechniques*, *34*(2), 374–378.

Schneider, T., Schmid, E., de Castro, J. V., Cardinale, M., Eberl, L., Grube, M., … Riedel, K. (2011). Structure and function of the symbiosis partners of the lung lichen (Lobaria pulmonaria L. Hoffm.) analyzed by metaproteomics. *Proteomics*, *11*(13), 2752–2756.

Sledjeski, D. D., Gupta, a, & Gottesman, S. (1996). The small RNA, DsrA, is essential for the low temperature expression of RpoS during exponential growth in Escherichia coli. *The EMBO Journal*, *15*(15), 3993–4000.

Webb, B. L., Cox, M. M., & Inman, R. B. (1997). Recombinational DNA repair: The RecF and RecR proteins limit the extension of RecA filaments beyond single-strand DNA gaps. *Cell*, *91*(3), 347–356.

Weeratunga, S. K., Gee, C. E., Lovell, S., Zeng, Y., Woodin, C. L., & Rivera, M. (2009). Binding of Pseudomonas aeruginosa apobacterioferritin-associated ferredoxin to bacterioferritin B promotes heme mediation of electron delivery and mobilization of core mineral iron. *Biochemistry*, *48*(31), 7420–7431.

Wigley, D. B. (2007). RecBCD: The Supercar of DNA Repair. *Cell*. doi:10.1016/j.cell.2007.11.004

Supplementary Documents

S-1: Differential expressed proteins under H_2O_2 stress condition

S-2: Differential expressed proteins under imipenem stress condition

S-3: Differential expressed proteins in LB medium

S-1: Differential expressed proteins under H_2O_2 stress condition

accession	gene	TIGRFAM main role	TIGRFAM subrole	protein description	Ox-M9 [a]
NP_742244.1_74	aroE	Amino acid biosynthesis	Aromatic amino acid family	shikimate 5-dehydrogenase	0
NP_743926.1_1756	PP_1770	Amino acid biosynthesis	Aromatic amino acid family	bifunctional cyclohexadienyl dehydrogenase/ 3-phosphoshikimate 1-carboxyvinyltransferase	-1
NP_743682.1_1512	dapE	Amino acid biosynthesis	Aspartate family	succinyl-diaminopimelate desuccinylase	0
NP_744673.1_2503	metY	Amino acid biosynthesis	Aspartate family	O-acetylhomoserine aminocarboxypropyltransferase	0
NP_745920.1_3750	dapF	Amino acid biosynthesis	Aspartate family	diaminopimelate epimerase	-1
NP_743161.1_991	argI	Amino acid biosynthesis	Glutamate family	ornithine carbamoyltransferase	0
NP_747286.1_5116	argA	Amino acid biosynthesis	Glutamate family	N-acetylglutamate synthase	0
NP_747115.1_4945	hisI	Amino acid biosynthesis	Histidine family	phosphoribosyl-AMP cyclohydrolase	0
NP_743628.1_1458	hom	Amino acid biosynthesis	Pyruvate family	homoserine dehydrogenase	-1
NP_744266.1_2096	pdxB	Amino acid biosynthesis	Serine family	erythronate-4-phosphate dehydrogenase	-1
NP_742533.1_363	bioD	Biosynthesis of cofactors, prosthetic groups, and carriers	Biotin	dithiobiotin synthetase	0
NP_742529.1_359	bioB	Biosynthesis of cofactors, prosthetic groups, and carriers	Biotin	biotin synthase	-1
NP_746320.1_4150	PP_4203	Biosynthesis of cofactors, prosthetic groups, and carriers	Chlorophyll and bacteriochlorphyll	electron-transferring-flavoprotein dehydrogenase	0
NP_743978.1_1808	folE	Biosynthesis of cofactors, prosthetic groups, and carriers	Folic acid	GTP cyclohydrolase I	-1
NP_746768.1_4598	ggt-2	Biosynthesis of cofactors, prosthetic groups, and carriers	Glutathione and analogs	gamma-glutamyltransferase	1
NP_742893.1_723	hemA	Biosynthesis of cofactors, prosthetic groups, and carriers	Heme, porphyrin, and cobalamin	glutamyl-tRNA reductase	0
NP_747298.1_5128	ubiF	Biosynthesis of cofactors, prosthetic groups, and carriers	Menaquinone and ubiquinone	2-octaprenyl-3-methyl-6-methoxy-1,4-benzoquinol hydroxylase	0
NP_743921.1_1751	ubiG	Biosynthesis of cofactors, prosthetic groups, and carriers	Menaquinone and ubiquinone	3-demethylubiquinone-9 3-methyltransferase	0
NP_747113.1_4943	ubiE	Biosynthesis of cofactors, prosthetic groups, and carriers	Menaquinone and ubiquinone	ubiquinone/menaquinone biosynthesis methyltransferase	0
NP_744271.1_2101	moaB-1	Biosynthesis of cofactors, prosthetic groups, and carriers	Molybdopterin	molybdenum cofactor biosynthesis protein B	0
NP_742546.1_376	pqqB	Biosynthesis of cofactors, prosthetic groups, and carriers	Other	pyrroloquinoline quinone biosynthesis protein PqqB	0
NP_743003.1_833	iscS	Biosynthesis of cofactors, prosthetic groups, and carriers	Other	cysteine desulfurase	-1
NP_747386.1_5216	coaBC	Biosynthesis of cofactors, prosthetic groups, and carriers	Pantothenate and coenzyme A	bifunctional phosphopantothenoylcysteine decarboxylase/phosphopantothenate synthase	0
NP_742682.1_512	thiL	Biosynthesis of cofactors, prosthetic groups, and carriers	Thiamine	thiamine monophosphate kinase	0
NP_742774.1_604	PP_0612	Biosynthesis of cofactors, prosthetic groups, and carriers	Thiamine	FAD dependent oxidoreductase	0

NP_746910.1_4740	mltB	Cell envelope	Biosynthesis and degradation of murein sacculus and peptidoglycan	lytic murein transglycosylase B	0
NP_743752.1_1582	uppS	Cell envelope	Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides	UDP diphosphate synthase	0
NP_742510.1_340	waaG	Cell envelope	Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides	lipopolysaccharide core biosynthesis protein WaaG	0
NP_744057.1_1887	kdsB	Cell envelope	Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides	3-deoxy-manno-octulosonate cytidylyltransferase	0
NP_747266.1_5096	PP_5165	Cell envelope	Other	NLPA lipoprotein	0
NP_742282.1_112	PP_0112	Cell envelope	Other	metal ABC transporter substrate-binding protein	0
NP_743603.1_1433	oprB-2	Cell envelope	Other	porin	0
NP_747403.1_5233	spoT	Cellular processes	Adaptations to atypical conditions	(p)ppGpp synthetase I SpoT/RelA	0
NP_744635.1_2465	PP_2487	Cellular processes	Adaptations to atypical conditions	aldehyde dehydrogenase	0
NP_743780.1_1610	rpoS	Cellular processes	Adaptations to atypical conditions	RNA polymerase sigma factor RpoS	1
NP_746767.1_4597	PP_4658	Cellular processes	Cell division	methyl-accepting chemotaxis sensory transducer	0
NP_747158.1_4988	PP_5057	Cellular processes	Cell division	M24/M37 family peptidase	0
NP_746457.1_4287	fleN	Cellular processes	Chemotaxis and motility	flagellar number regulator FleN	1
NP_743774.1_1604	PP_1617	Cellular processes	Detoxification	S-formylglutathione hydrolase	0
NP_744588.1_2418	ahpF	Cellular processes	Detoxification	alkyl hydroperoxide reductase	0
NP_742647.1_477	katA	Cellular processes	Detoxification	catalase (specific for oxi)	0
NP_742938.1_768	PP_0777	Cellular processes	Detoxification	glutathione peroxidase	0
NP_747481.1_5311	copA	Cellular processes	Detoxification	copper resistance protein A	0
NP_746257.1_4087	PP_4138	Cellular processes	Detoxification	NADPH-dependent FMN reductase	0
NP_742403.1_233	lsfA	Cellular processes	Detoxification	peroxidase	1
NP_743395.1_1225	PP_1235	Cellular processes	Detoxification	alkyl hydroperoxide reductase	-1
NP_745565.1_3395	oprN	Cellular processes	Detoxification	NodT family RND efflux system outer membrane lipoprotein	-1
NP_743182.2_1012	hexR	Cellular processes	DNA interactions	DNA-binding transcriptional regulator HexR	0
NP_747404.1_5234	PP_5303	Cellular processes	Other	endoribonuclease	0
NP_742563.1_393	PP_0396	Cellular processes	Sporulation and germination	hypothetical protein	0
NP_742354.1_184	pprA	Cellular processes	Sporulation and germination	LytTR family two component transcriptional regulator	1
NP_746431.1_4261	PP_4315	Cellular processes	Toxin production and resistance	PhzF family phenazine biosynthesis protein	-1
NP_743049.1_879	PP_0888	Central intermediary metabolism	Nitrogen metabolism	Fis family transcriptional regulator	0
NP_743792.1_1622	PP_1635	Central intermediary metabolism	Nitrogen metabolism	LuxR family transcriptional regulator	0
NP_742507.1_337	glnE	Central intermediary metabolism	Nitrogen metabolism	bifunctional glutamine-synthetase adenylyltransferase/deadenyltransferase	-1
NP_742935.1_765	pta	Central intermediary metabolism	Other	phosphate acetyltransferase	0
NP_745501.1_3331	PP_3361	Central intermediary metabolism	Other	hypothetical protein	0
NP_747317.1_5147	ppx	Central intermediary metabolism	Phosphorus compounds	Ppx/GppA phosphatase	-1
NP_747414.1_5244	hupA	DNA metabolism	DNA replication, recombination, and repair	histone family protein DNA-binding protein	0
NP_743507.1_1337	PP_1348	DNA metabolism	DNA replication, recombination, and repair	hypothetical protein	0

NP_746753.1_4583	radA	DNA metabolism	DNA replication, recombination, and repair	DNA repair protein RadA	0
NP_747055.1_4885	dbpA	DNA metabolism	DNA replication, recombination, and repair	ATP-dependent RNA helicase DbpA	0
NP_746626.1_4456	recQ	DNA metabolism	DNA replication, recombination, and repair	ATP-dependent DNA helicase RecQ	1
NP_744619.1_2449	ihfA	DNA metabolism	DNA replication, recombination, and repair	integration host factor subunit alpha	1
NP_743146.1_976	PP_0985	DNA metabolism	DNA replication, recombination, and repair	cold-shock domain-contain protein	1
NP_746118.1_3948	PP_3988	DNA metabolism	DNA replication, recombination, and repair	hypothetical protein	-1
NP_743134.1_964	ndpA	DNA metabolism	Other	nucleoid-associated protein NdpA	0
NP_746590.1_4420	aruG	Energy metabolism	Amino acids and amines	arginine N-succinyltransferase subunit beta	0
NP_745077.1_2907	PP_2933	Energy metabolism	Amino acids and amines	glutathione S-transferase	0
NP_746589.1_4419	astD	Energy metabolism	Amino acids and amines	succinylglutamic semialdehyde dehydrogenase	0
NP_743160.1_990	arcC	Energy metabolism	Amino acids and amines	carbamate kinase	0
NP_744681.1_2511	PP_2536	Energy metabolism	Amino acids and amines	glutathione S-transferase	0
NP_743162.1_992	arcA	Energy metabolism	Amino acids and amines	arginine deiminase	1
NP_746592.1_4422	argD	Energy metabolism	Amino acids and amines	bifunctional N-succinyldiaminopimelate-aminotransferase/acetylornithine transaminase	-1
NP_747518.1_5348	atpE	Energy metabolism	ATP-proton motive force interconversion	ATP synthase F0F1 subunit C	0
NP_747142.1_4972	glgP	Energy metabolism	Biosynthesis and degradation of polysaccharides	glycogen/starch/alpha-glucan phosphorylase	0
NP_743801.1_1631	wrbA	Energy metabolism	Electron transport	trp repressor binding protein	0
NP_746975.1_4805	PP_4870	Energy metabolism	Electron transport	azurin	0
NP_742974.1_804	cyoB	Energy metabolism	Electron transport	cytochrome o ubiquinol oxidase subunit I	0
NP_744574.1_2404	PP_2426	Energy metabolism	Fermentation	D-isomer specific 2-hydroxyacid dehydrogenase	0
NP_744486.1_2316	PP_2337	Energy metabolism	Fermentation	hypothetical protein	0
NP_746583.1_4413	csrA	Energy metabolism	Glycolysis/gluconeogenesis	carbon storage regulator	0
NP_746845.1_4675	PP_4737	Energy metabolism	Other	D-lactate dehydrogenase	0
NP_746604.1_4434	PP_4493	Energy metabolism	Other	FAD linked oxidase domain-containing protein	0
NP_747251.1_5081	rpiA	Energy metabolism	Pentose phosphate pathway	ribose-5-phosphate isomerase A	0
NP_743911.1_1741	fumC-2	Energy metabolism	TCA cycle	fumarate hydratase	0
NP_746309.1_4139	sdhD	Energy metabolism	TCA cycle	succinate dehydrogenase, hydrophobic membrane anchor protein	0
NP_744067.1_1897	fabD	Fatty acid and phospholipid metabolism	Biosynthesis	malonyl CoA-ACP transacylase	0
NP_747110.1_4940	PP_5008	Fatty acid and phospholipid metabolism	Biosynthesis	poly(hydroxyalkanoate) granule-associated protein	0
NP_747264.1_5094	PP_5163	no classification	no classification	acetyltransferase	0
NP_745216.1_3046	PP_3072	no classification	no classification	ecotin	0
NP_743636.1_1466	PP_1478	no classification	no classification	NADH:flavin oxidoreductase	0
NP_745333.1_3163	codA	no classification	no classification	N-isopropylammelide isopropylaminohydrolase	0
NP_744817.1_2647	PP_2673	no classification	no classification	pentapeptide repeat-containing protein	0
NP_743385.1_1215	PP_1225	no classification	no classification	radical SAM domain-containing protein	0

NP_742929.1_759	PP_0768	no classification	no classification	response regulator/hypothetical protein	0
NP_742342.1_172	PP_0173	no classification	no classification	transcriptional factor-like protein	0
NP_745091.1_2921	PP_2947	no classification	no classification	transcriptional regulator MvaT, P16 subunit	0
NP_743989.1_1819	PP_1834	no classification	no classification	ankyrin domain-containing protein	0
NP_746965.1_4795	PP_4860	no classification	no classification	N-acylglucosamine 2-epimerase	0
NP_746650.1_4480	ppiA	no classification	no classification	peptidyl-prolyl cis-trans isomerase A	0
NP_746414.1_4244	hyi	no classification	no classification	hydroxypyruvate isomerase	0
NP_745895.1_3725	PP_3765	no classification	no classification	transcriptional regulator MvaT, P16 subunit	0
NP_746190.1_4020	ivd	no classification	no classification	acyl-CoA dehydrogenase	0
NP_742427.1_257	nudE	no classification	no classification	ADP-ribose diphosphatase NudE	0
NP_742733.1_563	PP_0570	no classification	no classification	alpha-2-macroglobulin	0
NP_746147.1_3977	PP_4017	no classification	no classification	cupin	0
NP_742798.1_628	PP_0637	no classification	no classification	ISPpu15, transposase Orf2	0
NP_746672.1_4502	PP_4563	no classification	no classification	RNA binding S1 domain-containing protein	0
NP_742984.1_814	selD	no classification	no classification	selenophosphate synthetase	0
NP_747052.1_4882	PP_4949	no classification	no classification	TldD/PmbA family protein	0
NP_743946.1_1776	PP_1790	no classification	no classification	acylneuraminate cytidylyltransferase	0
NP_742363.1_193	algP	no classification	no classification	alginate regulatory protein AlgP	0
NP_742660.1_490	ansB	no classification	no classification	asparaginase	0
NP_746143.1_3973	PP_4013	no classification	no classification	NUDIX hydrolase	0
NP_744649.1_2479	PP_2501	no classification	no classification	phage integrase	0
NP_743587.1_1417	algN	no classification	no classification	sigma E regulatory protein MucB/RseB	0
NP_743456.1_1286	estB	no classification	no classification	carboxylesterase	0
NP_744456.1_2286	PP_2307	no classification	no classification	CHAD domain-containing protein	0
NP_744116.1_1946	PP_1964	no classification	no classification	deoxynucleotide monophosphate kinase	0
NP_745785.1_3615	PP_3649	no classification	no classification	GntR family transcriptional regulator	1
NP_742701.1_531	ppa	no classification	no classification	inorganic pyrophosphatase	1
NP_746054.1_3884	PP_3924	no classification	no classification	sterol-binding domain-containing protein	1
NP_743019.1_849	PP_0858	no classification	no classification	aminotransferase	-1
NP_746744.1_4574	PP_4635	no classification	no classification	trans-2-enoyl-CoA reductase	-1
NP_743398.1_1228	PP_1238	no classification	no classification	lipoprotein	-1
NP_745916.1_3746	PP_3786	no classification	no classification	aminotransferase	-1
NP_742914.1_744	PP_0753	no classification	no classification	lipoprotein	-1
NP_744471.1_2301	oprI	no classification	no classification	outer membrane lipoprotein OprI	-1
NP_747254.1_5084	PP_5153	no classification	no classification	fumarylacetoacetate (FAA) hydrolase	-1
NP_745569.1_3399	PP_3431	Protein fate	Degradation of proteins, peptides, and glycopeptides	ThiJ/PfpI domain-containing protein	0

NP_747102.1_4932	hslV	Protein fate	Degradation of proteins, peptides, and glycopeptides	ATP-dependent protease peptidase subunit	0
NP_746139.1_3969	clpS	Protein fate	Degradation of proteins, peptides, and glycopeptides	ATP-dependent Clp protease adaptor protein ClpS	-1
NP_742934.1_764	PP_0773	Protein fate	Protein and peptide secretion and trafficking	OmpA/MotB domain-containing protein	0
NP_742176.1_6	yidC	Protein fate	Protein and peptide secretion and trafficking	inner membrane protein translocase component YidC	0
NP_747178.1_5008	PP_5077	Protein fate	Protein and peptide secretion and trafficking	sporulation domain-containing protein	1
NP_745251.1_3081	PP_3107	Protein fate	Protein and peptide secretion and trafficking	hypothetical protein	-1
NP_743006.1_836	hscB	Protein fate	Protein folding and stabilization	co-chaperone HscB	0
NP_743871.1_1701	fklB-2	Protein fate	Protein folding and stabilization	FKBP-type peptidylprolyl isomerase	0
NP_744727.1_2557	PP_2583	Protein synthesis	Ribosomal proteins: synthesis and modification	acetyltransferase	0
NP_743982.1_1812	PP_1827	Protein synthesis	Ribosomal proteins: synthesis and modification	N5-glutamine S-adenosyl-L-methionine-dependent methyltransferase	0
NP_747383.1_5213	rpmB	Protein synthesis	Ribosomal proteins: synthesis and modification	50S ribosomal protein L28	1
NP_743476.1_1306	rpsI	Protein synthesis	Ribosomal proteins: synthesis and modification	30S ribosomal protein S9	1
NP_743620.2_1450	rpsP	Protein synthesis	Ribosomal proteins: synthesis and modification	30S ribosomal protein S16	1
NP_744065.1_1895	rpmF	Protein synthesis	Ribosomal proteins: synthesis and modification	50S ribosomal protein L32	1
NP_746818.2_4648	rpsO	Protein synthesis	Ribosomal proteins: synthesis and modification	30S ribosomal protein S15	1
NP_742894.1_724	prfA	Protein synthesis	Translation factors	peptide chain release factor 1	1
NP_742658.1_488	selA	Protein synthesis	tRNA aminoacylation	selenocysteine synthase	0
NP_743238.1_1068	PP_1077	Protein synthesis	tRNA aminoacylation	ybaK/ebsC protein	0
NP_743471.1_1301	PP_1311	Protein synthesis	tRNA aminoacylation	tryptophanyl-tRNA synthetase	-1
NP_747215.1_5045	PP_5114	Protein synthesis	tRNA and rRNA base modification	hypothetical protein	0
NP_742994.1_824	tgt	Protein synthesis	tRNA and rRNA base modification	queuine tRNA-ribosyltransferase	0
NP_746763.1_4593	trmA	Protein synthesis	tRNA and rRNA base modification	tRNA (uracil-5-)-methyltransferase	0
NP_743337.1_1167	nrdB	Purines, pyrimidines, nucleosides, and nucleotides	2'-Deoxyribonucleotide metabolism	ribonucleotide-diphosphate reductase subunit beta	-1
NP_742371.1_201	PP_0202	Purines, pyrimidines, nucleosides, and nucleotides	Purine ribonucleotide biosynthesis	hypothetical protein	0
NP_747099.1_4929	pyrR	Purines, pyrimidines, nucleosides, and nucleotides	Salvage of nucleosides and nucleotides	bifunctional pyrimidine regulatory protein PyrR/uracil phosphoribosyltransferase	0
NP_743374.1_1204	PP_1214	Regulatory functions	DNA interactions	hypothetical protein	0
NP_744119.1_1949	PP_1968	Regulatory functions	DNA interactions	TetR family transcriptional regulator	0
NP_744950.1_2780	PP_2806	Regulatory functions	DNA interactions	TetR family transcriptional regulator	0
NP_743346.1_1176	phoP	Regulatory functions	DNA interactions	winged helix family two component transcriptional regulator	0
NP_744934.1_2764	PP_2790	Regulatory functions	DNA interactions	Fis family transcriptional regulator	0
NP_744695.1_2525	PP_2550	Regulatory functions	DNA interactions	hypothetical protein	0
NP_744745.1_2575	PP_2601	Regulatory functions	DNA interactions	IclR family transcriptional regulator	-1
NP_747426.1_5256	phoU	Regulatory functions	Other	phosphate transporter PhoU	0
NP_746997.1_4827	hfq	Regulatory functions	Other	RNA-binding protein Hfq	1
NP_746478.1_4308	PP_4364	Regulatory functions	Protein interactions	anti-sigma-factor antagonist	0

NP_743586.1_1416	mucA	Regulatory functions	Protein interactions	anti sigma-E protein, RseA	0
NP_746509.1_4339	flgM	Regulatory functions	Protein interactions	anti-sigma-28 factor FlgM	-1
NP_742940.1_770	PP_0779	Regulatory functions	Small molecule interactions	methyl-accepting chemotaxis transducer/sensory box protein	0
NP_743621.1_1451	rimM	Transcription	RNA processing	16S rRNA-processing protein RimM	0
NP_744369.1_2199	PP_2220	Transcription	Transcription factors	C4-type zinc finger DksA/TraR family protein	0
NP_746802.1_4632	PP_4693	Transcription	Transcription factors	C4-type zinc finger DksA/TraR family protein	1
NP_746597.1_4427	PP_4486	Transport and binding proteins	Amino acids, peptides and amines	amino acid ABC transporter substrate-binding protein	0
NP_742578.1_408	PP_0412	Transport and binding proteins	Amino acids, peptides and amines	polyamine ABC transporter substrate-binding protein	0
NP_747281.1_5111	potF-1	Transport and binding proteins	Amino acids, peptides and amines	spermidine/putrescine ABC transporter substrate-binding protein	0
NP_743883.1_1713	PP_1726	Transport and binding proteins	Amino acids, peptides and amines	ABC transporter substrate-binding protein	0
NP_742450.1_280	PP_0283	Transport and binding proteins	Amino acids, peptides and amines	amino acid ABC transporter ATP-binding protein	0
NP_747125.1_4955	PP_5024	Transport and binding proteins	Amino acids, peptides and amines	amino acid ABC transporter substrate-binding protein	0
NP_743229.1_1059	PP_1068	Transport and binding proteins	Amino acids, peptides and amines	amino acid ABC transporter ATP-binding protein	0
NP_746972.1_4802	PP_4867	Transport and binding proteins	Amino acids, peptides and amines	ABC transporter substrate-binding protein	1
NP_747297.1_5127	PP_5196	Transport and binding proteins	Amino acids, peptides and amines	iron ABC transporter substrate-binding protein	1
NP_744602.1_2432	rbsB	Transport and binding proteins	Carbohydrates, organic alcohols, and acids	monosaccharide-transporting ATPase	0
NP_747384.1_5214	PP_5283	Transport and binding proteins	Cations and iron carrying compounds	peptide ABC transporter substrate-binding protein	0
NP_742648.1_478	PP_0482	Transport and binding proteins	Cations and iron carrying compounds	bacterioferritin	0
NP_743243.1_1073	bfr	Transport and binding proteins	Cations and iron carrying compounds	bacterioferritin	1
NP_743935.1_1765	PP_1779	Transport and binding proteins	Other	lipopolysaccharide ABC export system, ATP-binding protein	0
NP_742189.1_19	PP_0019	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_742257.1_87	PP_0087	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_743016.1_846	PP_0855	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_743069.1_899	PP_0908	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_743700.1_1530	PP_1543	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_744310.1_2140	PP_2161	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_744426.1_2256	PP_2277	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_744590.1_2420	PP_2442	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_746645.1_4475	PP_4535	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_746679.1_4509	PP_4570	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_746725.1_4555	PP_4616	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_744460.1_2290	PP_2311	Unknown function	Enzymes of unknown specificity	TatD family hydrolase	0
NP_747330.1_5160	PP_5229	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_747454.1_5284	PP_5353	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_742466.1_296	PP_0299	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_742589.1_419	PP_0423	Unknown function	Enzymes of unknown specificity	hypothetical protein	0

NP_742834.1_664	PP_0673	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_743384.1_1214	PP_1224	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_743820.1_1650	PP_1663	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_744255.1_2085	PP_2105	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_746039.1_3869	PP_3909	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_746623.1_4453	PP_4513	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_747302.1_5132	PP_5201	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_742234.1_64	PP_0064	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_743990.1_1820	PP_1835	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_744064.1_1894	PP_1910	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_745094.1_2924	PP_2950	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_745957.1_3787	PP_3827	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_743675.1_1505	PP_1518	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_746671.1_4501	PP_4562	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_747083.1_4913	PP_4981	Unknown function	Enzymes of unknown specificity	hypothetical protein	1
NP_742903.1_733	PP_0742	Unknown function	Enzymes of unknown specificity	hypothetical protein	-1
NP_744570.1_2400	PP_2422	Unknown function	General	alkylhydroperoxidase	0

a: Up/down regulation of protein expression, the value 0: exclusively expressed in H_2O_2 stress; Value 1: up regulated in H_2O_2 stress; Value -1: down regulated in H_2O_2 stress.

S-2: Differential expressed proteins under imipenem stress condition

accession	gene	TIGRFAM main role	TIGRFAM subrole	protein description	Imp-M9[a]
NP_747115.1_4945	hisI	Amino acid biosynthesis	Histidine family	phosphoribosyl-AMP cyclohydrolase	0
NP_743161.1_991	argI	Amino acid biosynthesis	Glutamate family	ornithine carbamoyltransferase	0
NP_743682.1_1512	dapE	Amino acid biosynthesis	Aspartate family	succinyl-diaminopimelate desuccinylase	0
NP_742244.1_74	aroE	Amino acid biosynthesis	Aromatic amino acid family	shikimate 5-dehydrogenase	0
NP_743926.1_1756	PP_1770	Amino acid biosynthesis	Aromatic amino acid family	bifunctional cyclohexadienyl dehydrogenase/ 3-phosphoshikimate 1-carboxyvinyltransferase	-1
NP_744266.1_2096	pdxB	Amino acid biosynthesis	Serine family	erythronate-4-phosphate dehydrogenase	-1
NP_742923.1_753	hprA	Amino acid biosynthesis	Serine family	glycerate dehydrogenase	0
NP_743806.1_1636	ldhA	Amino acid biosynthesis	Serine family	D-lactate dehydrogenase	0
NP_743001.1_831	cysE	Amino acid biosynthesis	Serine family	serine O-acetyltransferase	0
NP_745074.1_2904	PP_2930	Amino acid biosynthesis	Pyruvate family	L-serine dehydratase	0
NP_745584.1_3414	ilvA-1	Amino acid biosynthesis	Pyruvate family	threonine dehydratase	0
NP_746164.1_3994	PP_4037	Amino acid biosynthesis	Glutamate family	oxidoreductase	0
NP_744186.1_2016	PP_2036	Amino acid biosynthesis	Aspartate family	dihydrodipicolinate synthetase	0
NP_742819.1_649	PP_0658	Amino acid biosynthesis	Aspartate family	homocysteine S-methyltransferase	0
NP_743985.1_1815	aroC	Amino acid biosynthesis	Aromatic amino acid family	chorismate synthase	0
NP_742587.1_417	trpD	Amino acid biosynthesis	Aromatic amino acid family	anthranilate phosphoribosyltransferase	0
NP_742252.1_82	trpA	Amino acid biosynthesis	Aromatic amino acid family	tryptophan synthase subunit alpha	-1
NP_742682.1_512	thiL	Biosynthesis of cofactors, prosthetic groups, and carriers	Thiamine	thiamine monophosphate kinase	0
NP_747386.1_5216	coaBC	Biosynthesis of cofactors, prosthetic groups, and carriers	Pantothenate and coenzyme A	bifunctional phosphopantothenoylcysteine decarboxylase/phosphopantothenate synthase	0
NP_742546.1_376	pqqB	Biosynthesis of cofactors, prosthetic groups, and carriers	Other	pyrroloquinoline quinone biosynthesis protein PqqB	0
NP_744271.1_2101	moaB-1	Biosynthesis of cofactors, prosthetic groups, and carriers	Molybdopterin	molybdenum cofactor biosynthesis protein B	0
NP_747298.1_5128	ubiF	Biosynthesis of cofactors, prosthetic groups, and carriers	Menaquinone and ubiquinone	2-octaprenyl-3-methyl-6-methoxy-1,4-benzoquinol hydroxylase	0
NP_743921.1_1751	ubiG	Biosynthesis of cofactors, prosthetic groups, and carriers	Menaquinone and ubiquinone	3-demethylubiquinone-9 3-methyltransferase	0
NP_747113.1_4943	ubiE	Biosynthesis of cofactors, prosthetic groups, and carriers	Menaquinone and ubiquinone	ubiquinone/menaquinone biosynthesis methyltransferase	0
NP_746320.1_4150	PP_4203	Biosynthesis of cofactors, prosthetic groups, and carriers	Chlorophyll and bacteriochlorphyll	electron-transferring-flavoprotein dehydrogenase	0
NP_742533.1_363	bioD	Biosynthesis of cofactors, prosthetic groups, and carriers	Biotin	dithiobiotin synthetase	0
NP_746768.1_4598	ggt-2	Biosynthesis of cofactors, prosthetic groups, and carriers	Glutathione and analogs	gamma-glutamyltransferase	1
NP_743003.1_833	iscS	Biosynthesis of cofactors, prosthetic groups, and carriers	Other	cysteine desulfurase	-1
NP_746887.1_4717	thiD	Biosynthesis of cofactors, prosthetic groups, and carriers	Thiamine	phosphomethylpyrimidine kinase	0

NP_742569.1_399	pdxA	Biosynthesis of cofactors, prosthetic groups, and carriers	Pyridoxine	4-hydroxythreonine-4-phosphate dehydrogenase	0
NP_746808.1_4638	panB	Biosynthesis of cofactors, prosthetic groups, and carriers	Pantothenate and coenzyme A	3-methyl-2-oxobutanoate hydroxymethyltransferase	0
NP_747213.1_5043	PP_5112	Biosynthesis of cofactors, prosthetic groups, and carriers	Other	peptidase M16 domain-containing protein	0
NP_742896.1_726	PP_0735	Biosynthesis of cofactors, prosthetic groups, and carriers	Molybdopterin	molybdopterin biosynthesis protein MoeB	0
NP_746281.1_4111	PP_4164	Biosynthesis of cofactors, prosthetic groups, and carriers	Menaquinone and ubiquinone	alpha/beta hydrolase	0
NP_744071.1_1901	pabC	Biosynthesis of cofactors, prosthetic groups, and carriers	Folic acid	4-amino-4-deoxychorismate lyase	0
NP_744329.1_2159	PP_2180	Biosynthesis of cofactors, prosthetic groups, and carriers	Biotin	aminotransferase	0
NP_747069.1_4899	PP_4966	Biosynthesis of cofactors, prosthetic groups, and carriers	Biotin	ArsR family transcriptional regulator	0
NP_744272.1_2102	moeA	Biosynthesis of cofactors, prosthetic groups, and carriers	Molybdopterin	molybdopterin biosynthesis MoeA protein	-1
NP_747266.1_5096	PP_5165	Cell envelope	Other	NLPA lipoprotein	0
NP_743603.1_1433	oprB-2	Cell envelope	Other	porin	0
NP_743752.1_1582	uppS	Cell envelope	Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides	UDP diphosphate synthase	0
NP_744057.1_1887	kdsB	Cell envelope	Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides	3-deoxy-manno-octulosonate cytidylyltransferase	0
NP_742510.1_340	waaG	Cell envelope	Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides	lipopolysaccharide core biosynthesis protein WaaG	0
NP_742664.1_494	PP_0500	Cell envelope	Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides	dTDP-4-dehydrorhamnose reductase	0
NP_743758.1_1588	lpxD	Cell envelope	Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase	0
NP_742508.1_338	waaF	Cell envelope	Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides	lipopolysaccharide heptosyltransferase II	0
NP_743949.1_1779	PP_1793	Cell envelope	Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides	glycosyl transferase family protein	0
NP_746903.1_4733	PP_4798	Cell envelope	Biosynthesis and degradation of murein sacculus and peptidoglycan	lytic murein transglycosylase	0
NP_747000.1_4830	PP_4897	Cell envelope	Biosynthesis and degradation of murein sacculus and peptidoglycan	N-acetylmuramoyl-L-alanine amidase	0
NP_742710.1_540	mpl	Cell envelope	Biosynthesis and degradation of murein sacculus and peptidoglycan	UDP-N-acetylmuramate	0
NP_743497.1_1327	murC	Cell envelope	Biosynthesis and degradation of murein sacculus and peptidoglycan	UDP-N-acetylmuramate--L-alanine ligase	1
NP_743940.1_1770	rmlD	Cell envelope	Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides	dTDP-4-dehydrorhamnose reductase	-1
NP_742563.1_393	PP_0396	Cellular processes	Sporulation and germination	hypothetical protein	0
NP_747404.1_5234	PP_5303	Cellular processes	Other	endoribonuclease	0
NP_743774.1_1604	PP_1617	Cellular processes	Detoxification	S-formylglutathione hydrolase	0
NP_742938.1_768	PP_0777	Cellular processes	Detoxification	glutathione peroxidase	0
NP_747481.1_5311	copA	Cellular processes	Detoxification	copper resistance protein A	0
NP_746257.1_4087	PP_4138	Cellular processes	Detoxification	NADPH-dependent FMN reductase	0
NP_746767.1_4597	PP_4658	Cellular processes	Cell division	methyl-accepting chemotaxis sensory transducer	0
NP_747403.1_5233	spoT	Cellular processes	Adaptations to atypical conditions	(p)ppGpp synthetase I SpoT/RelA	0
NP_744635.1_2465	PP_2487	Cellular processes	Adaptations to atypical conditions	aldehyde dehydrogenase	0
NP_743780.1_1610	rpoS	Cellular processes	Adaptations to atypical conditions	RNA polymerase sigma factor RpoS	1

NP_742354.1_184	pprA	Cellular processes	Sporulation and germination	LytTR family two component transcriptional regulator	1
NP_742403.1_233	lsfA	Cellular processes	Detoxification	peroxidase	1
NP_745565.1_3395	oprN	Cellular processes	Detoxification	NodT family RND efflux system outer membrane lipoprotein	-1
NP_746431.1_4261	PP_4315	Cellular processes	Toxin production and resistance	PhzF family phenazine biosynthesis protein	-1
NP_742425.1_255	PP_0258	Cellular processes	Sporulation and germination	LysM domain/BON superfamily protein	0
NP_744080.1_1910	PP_1927	Cellular processes	Detoxification	arsenical resistance protein ArsH	0
NP_746494.1_4324	flgL	Cellular processes	Chemotaxis and motility	flagellar hook-associated protein FlgL	0
NP_746489.1_4319	fliS	Cellular processes	Chemotaxis and motility	flagellar protein FliS	0
NP_746991.1_4821	PP_4888	Cellular processes	Chemotaxis and motility	methyl-accepting chemotaxis sensory transducer	0
NP_746482.2_4312	fliG	Cellular processes	Chemotaxis and motility	flagellar motor switch protein G	0
NP_746458.1_4288	flhF	Cellular processes	Chemotaxis and motility	flagellar biosynthesis regulator FlhF	0
NP_742484.1_314	PP_0317	Cellular processes	Chemotaxis and motility	methyl-accepting chemotaxis sensory transducer	0
NP_744787.1_2617	PP_2643	Cellular processes	Chemotaxis and motility	methyl-accepting chemotaxis sensory transducer	0
NP_747121.1_4951	PP_5020	Cellular processes	Chemotaxis and motility	methyl-accepting chemotaxis sensory transducer	0
NP_743888.1_1718	minE	Cellular processes	Cell division	cell division topological specificity factor MinE	0
NP_742747.1_577	PP_0584	Cellular processes	Cell division	methyl-accepting chemotaxis transducer	0
NP_746490.1_4320	fliD	Cellular processes	Chemotaxis and motility	flagellar cap protein FliD	1
NP_746471.1_4301	fliN	Cellular processes	Chemotaxis and motility	flagellar motor switch protein	1
NP_745501.1_3331	PP_3361	Central intermediary metabolism	Other	hypothetical protein	0
NP_742935.1_765	pta	Central intermediary metabolism	Other	phosphate acetyltransferase	0
NP_743049.1_879	PP_0888	Central intermediary metabolism	Nitrogen metabolism	Fis family transcriptional regulator	0
NP_743792.1_1622	PP_1635	Central intermediary metabolism	Nitrogen metabolism	LuxR family transcriptional regulator	0
NP_743100.1_930	PP_0939	Central intermediary metabolism	Polyamine biosynthesis	nitrilase/cyanide hydratase and apolipoprotein N-acyltransferase	0
NP_745885.1_3715	paaH	Central intermediary metabolism	Other	3-hydroxybutyryl-CoA dehydrogenase	0
NP_746219.1_4049	gacA	Central intermediary metabolism	Nitrogen metabolism	DNA-binding response regulator GacA	0
NP_747149.1_4979	ntrC	Central intermediary metabolism	Nitrogen metabolism	Fis family transcriptional regulator	0
NP_745557.1_3387	PP_3419	Central intermediary metabolism	Nitrogen metabolism	Fis family transcriptional regulator	0
NP_744260.1_2090	aer-2	Central intermediary metabolism	Nitrogen fixation	aerotaxis receptor Aer-2	0
NP_743908.1_1738	PP_1752	Central intermediary metabolism	Phosphorus compounds	hypothetical protein	1
NP_744453.1_2283	PP_2304	Central intermediary metabolism	Nitrogen fixation	PpiC-type peptidyl-prolyl cis-trans isomerase	-1
NP_743134.1_964	ndpA	DNA metabolism	Other	nucleoid-associated protein NdpA	0
NP_747414.1_5244	hupA	DNA metabolism	DNA replication, recombination, and repair	histone family protein DNA-binding protein	0
NP_743507.1_1337	PP_1348	DNA metabolism	DNA replication, recombination, and repair	hypothetical protein	0
NP_746753.1_4583	radA	DNA metabolism	DNA replication, recombination, and repair	DNA repair protein RadA	0
NP_747055.1_4885	dbpA	DNA metabolism	DNA replication, recombination, and repair	ATP-dependent RNA helicase DbpA	0
NP_743146.1_976	PP_0985	DNA metabolism	DNA replication, recombination, and repair	cold-shock domain-contain protein	1
NP_746118.1_3948	PP_3988	DNA metabolism	DNA replication, recombination, and repair	hypothetical protein	-1
NP_743524.1_1354	sbcB	DNA metabolism	DNA replication, recombination, and repair	exonuclease I	0

NP_743706.1_1536	PP_1549	DNA metabolism	DNA replication, recombination, and repair	Cro/CI family transcriptional regulator	0
NP_742900.1_730	phrB	DNA metabolism	DNA replication, recombination, and repair	deoxyribodipyrimidine photo-lyase	0
NP_743929.1_1759	ihfB	DNA metabolism	DNA replication, recombination, and repair	integration host factor subunit beta	0
NP_746132.1_3962	PP_4002	DNA metabolism	DNA replication, recombination, and repair	recombination factor protein RarA	0
NP_742651.1_481	ssb	DNA metabolism	DNA replication, recombination, and repair	single-stranded DNA-binding protein	0
NP_747365.1_5195	rep	DNA metabolism	DNA replication, recombination, and repair	ATP-dependent DNA helicase Rep	0
NP_743911.1_1741	fumC-2	Energy metabolism	TCA cycle	fumarate hydratase	0
NP_746309.1_4139	sdhD	Energy metabolism	TCA cycle	succinate dehydrogenase, hydrophobic membrane anchor protein	0
NP_747251.1_5081	rpiA	Energy metabolism	Pentose phosphate pathway	ribose-5-phosphate isomerase A	0
NP_746604.1_4434	PP_4493	Energy metabolism	Other	FAD linked oxidase domain-containing protein	0
NP_744574.1_2404	PP_2426	Energy metabolism	Fermentation	D-isomer specific 2-hydroxyacid dehydrogenase	0
NP_744486.1_2316	PP_2337	Energy metabolism	Fermentation	hypothetical protein	0
NP_743801.1_1631	wrbA	Energy metabolism	Electron transport	trp repressor binding protein	0
NP_742974.1_804	cyoB	Energy metabolism	Electron transport	cytochrome o ubiquinol oxidase subunit I	0
NP_747142.1_4972	glgP	Energy metabolism	Biosynthesis and degradation of polysaccharides	glycogen/starch/alpha-glucan phosphorylase	0
NP_746589.1_4419	astD	Energy metabolism	Amino acids and amines	succinylglutamic semialdehyde dehydrogenase	0
NP_745077.1_2907	PP_2933	Energy metabolism	Amino acids and amines	glutathione S-transferase	0
NP_746590.1_4420	aruG	Energy metabolism	Amino acids and amines	arginine N-succinyltransferase subunit beta	0
NP_743160.1_990	arcC	Energy metabolism	Amino acids and amines	carbamate kinase	0
NP_744681.1_2511	PP_2536	Energy metabolism	Amino acids and amines	glutathione S-transferase	0
NP_746592.1_4422	argD	Energy metabolism	Amino acids and amines	bifunctional N-succinyldiaminopimelate-aminotransferase/acetylornithine transaminase	-1
NP_744261.1_2091	acnA	Energy metabolism	TCA cycle	aconitate hydratase	0
NP_743538.1_1368	pcaB	Energy metabolism	Other	3-carboxy-cis,cis-muconate cycloisomerase	0
NP_745305.1_3135	benA	Energy metabolism	Other	benzoate dioxygenase subunit alpha	0
NP_745846.1_3676	catA	Energy metabolism	Other	catechol 1,2-dioxygenase	0
NP_745876.1_3706	glcD	Energy metabolism	Other	glycolate oxidase subunit GlcD	0
NP_745307.1_3137	benC	Energy metabolism	Other	oxidoreductase FAD/NAD(P)-binding domain-containing protein	0
NP_743539.1_1369	pcaD	Energy metabolism	Other	3-oxoadipate enol-lactonase	0
NP_745306.1_3136	benB	Energy metabolism	Other	benzoate dioxygenase subunit beta	0
NP_746810.1_4640	pgi	Energy metabolism	Glycolysis/gluconeogenesis	glucose-6-phosphate isomerase	0
NP_746776.1_4606	mmsA-2	Energy metabolism	Amino acids and amines	methylmalonate-semialdehyde dehydrogenase	0
NP_744601.1_2431	ansA	Energy metabolism	Amino acids and amines	type II L-asparaginase	1
NP_746245.1_4075	nuoI	Energy metabolism	Electron transport	NADH dehydrogenase subunit I	1
NP_746141.1_3971	icd	Energy metabolism	TCA cycle	isocitrate dehydrogenase	1
NP_746307.1_4137	sdhB	Energy metabolism	Aerobic	succinate dehydrogenase iron-sulfur subunit	-1
NP_745522.1_3352	PP_3382	Energy metabolism	Electron transport	gluconate 2-dehydrogenase	-1
NP_746243.1_4073	nuoG	Energy metabolism	Electron transport	NADH dehydrogenase subunit G	-1
NP_743411.2_1241	mqo-2	Energy metabolism	TCA cycle	malate:quinone oxidoreductase	-1
NP_744067.1_1897	fabD	Fatty acid and phospholipid metabolism	Biosynthesis	malonyl CoA-ACP transacylase	0

NP_747110.1_4940	PP_5008	Fatty acid and phospholipid metabolism	Biosynthesis	poly(hydroxyalkanoate) granule-associated protein	0
NP_744364.1_2194	fadAx	Fatty acid and phospholipid metabolism	Other	acetyl-CoA acetyltransferase	0
NP_746868.1_4698	tesB	Fatty acid and phospholipid metabolism	Biosynthesis	acyl-CoA thioesterase	0
NP_744362.1_2192	fadDx	Fatty acid and phospholipid metabolism	Biosynthesis	AMP-dependent synthetase/ligase	0
NP_744098.1_1928	PP_1946	Fatty acid and phospholipid metabolism	Biosynthesis	short chain dehydrogenase/reductase oxidoreductase	0
NP_742744.1_574	fabG	Fatty acid and phospholipid metabolism	Biosynthesis	3-ketoacyl-ACP reductase	0
NP_742929.1_759	PP_0768	no classification	no classification	response regulator/hypothetical protein	0
NP_745091.1_2921	PP_2947	no classification	no classification	transcriptional regulator MvaT, P16 subunit	0
NP_747264.1_5094	PP_5163	no classification	no classification	acetyltransferase	0
NP_745216.1_3046	PP_3072	no classification	no classification	ecotin	0
NP_743636.1_1466	PP_1478	no classification	no classification	NADH:flavin oxidoreductase	0
NP_745333.1_3163	codA	no classification	no classification	N-isopropylammelide isopropylaminohydrolase	0
NP_744817.1_2647	PP_2673	no classification	no classification	pentapeptide repeat-containing protein	0
NP_743385.1_1215	PP_1225	no classification	no classification	radical SAM domain-containing protein	0
NP_742342.1_172	PP_0173	no classification	no classification	transcriptional factor-like protein	0
NP_746147.1_3977	PP_4017	no classification	no classification	cupin	0
NP_742798.1_628	PP_0637	no classification	no classification	ISPpu15, transposase Orf2	0
NP_746414.1_4244	hyi	no classification	no classification	hydroxypyruvate isomerase	0
NP_745895.1_3725	PP_3765	no classification	no classification	transcriptional regulator MvaT, P16 subunit	0
NP_746190.1_4020	ivd	no classification	no classification	acyl-CoA dehydrogenase	0
NP_742427.1_257	nudE	no classification	no classification	ADP-ribose diphosphatase NudE	0
NP_742733.1_563	PP_0570	no classification	no classification	alpha-2-macroglobulin	0
NP_742984.1_814	selD	no classification	no classification	selenophosphate synthetase	0
NP_747052.1_4882	PP_4949	no classification	no classification	TldD/PmbA family protein	0
NP_743456.1_1286	estB	no classification	no classification	carboxylesterase	0
NP_744456.1_2286	PP_2307	no classification	no classification	CHAD domain-containing protein	0
NP_744116.1_1946	PP_1964	no classification	no classification	deoxynucleotide monophosphate kinase	0
NP_746672.1_4502	PP_4563	no classification	no classification	RNA binding S1 domain-containing protein	0
NP_742701.1_531	ppa	no classification	no classification	inorganic pyrophosphatase	1
NP_746744.1_4574	PP_4635	no classification	no classification	trans-2-enoyl-CoA reductase	-1
NP_743398.1_1228	PP_1238	no classification	no classification	lipoprotein	-1
NP_744471.1_2301	oprI	no classification	no classification	outer membrane lipoprotein OprI	-1
NP_747035.1_4865	PP_4932	no classification	no classification	aldo/keto reductase	0
NP_743181.1_1011	PP_1020	no classification	no classification	aldose 1-epimerase	0
NP_746162.1_3992	PP_4034	no classification	no classification	allantoate amidohydrolase	0
NP_743473.1_1303	PP_1313	no classification	no classification	AraC family transcriptional regulator	0
NP_747467.1_5297	lpd3	no classification	no classification	dihydrolipoamide dehydrogenase	0
NP_746165.1_3995	PP_4038	no classification	no classification	dihydropyrimidine dehydrogenase	0
NP_743639.1_1469	PP_1481	no classification	no classification	gamma-aminobutyraldehyde dehydrogenase	0
NP_743957.1_1787	wbpZ	no classification	no classification	glycosyl transferase WbpZ	0

NP_745727.1_3557	PP_3591	no classification	no classification	malate/L-lactate dehydrogenase	0
NP_744092.1_1922	PP_1940	no classification	no classification	methyl-accepting chemotaxis transducer	0
NP_744278.1_2108	PP_2129	no classification	no classification	MOSC domain-containing protein	0
NP_747406.1_5236	PP_5305	no classification	no classification	NAD-dependent epimerase/dehydratase	0
NP_742755.1_585	PP_0592	no classification	no classification	short chain dehydrogenase	0
NP_744025.1_1855	tpm	no classification	no classification	thiopurine S-methyltransferase	0
NP_747040.1_4870	PP_4937	no classification	no classification	toluene tolerance protein	0
SV=4224473	SV=4224473	no classification	no classification	putative sORF	0
NP_747374.1_5204	PP_5273	no classification	no classification	oxidoreductase	0
NP_746081.1_3911	pcaI	no classification	no classification	3-oxoadipate CoA-transferase subunit A	0
NP_745264.1_3094	PP_3120	no classification	no classification	aldo/keto reductase	0
NP_743536.1_1366	pcaF	no classification	no classification	beta-ketoadipyl CoA thiolase	0
NP_744640.1_2470	PP_2492	no classification	no classification	butanol dehydrogenase, NADH-dependent	0
NP_746510.1_4340	PP_4396	no classification	no classification	FlgN family protein	0
NP_746192.1_4022	PP_4066	no classification	no classification	gamma-carboxygeranoyl-CoA hydratase	0
NP_743977.1_1807	PP_1822	no classification	no classification	glutaredoxin	0
NP_745259.1_3089	PP_3115	no classification	no classification	ISPpu13, transposase Orf3	0
NP_745694.1_3524	PP_3557	no classification	no classification	methyl-accepting chemotaxis transducer	0
NP_745397.1_3227	PP_3254	no classification	no classification	nucleosidase	0
NP_746545.1_4375	PP_4432	no classification	no classification	peptidase, M24 family protein	0
NP_743434.1_1264	PP_1274	no classification	no classification	short chain dehydrogenase	0
NP_742757.1_587	PP_0594	no classification	no classification	TetR family transcriptional regulator	0
NP_742565.1_395	glpE	no classification	no classification	thiosulfate sulfurtransferase	0
NP_746924.1_4754	PP_4819	no classification	no classification	Zinc finger-domain-containing protein	0
NP_744624.1_2454	PP_2476	no classification	no classification	zinc-containing alcohol dehydrogenase	0
NP_744262.1_2092	PP_2113	no classification	no classification	RNA 2'-O-ribose methyltransferase	0
NP_742236.1_66	sun	no classification	no classification	sun protein	0
NP_744158.1_1988	fadH	no classification	no classification	NADH:flavin oxidoreductase	1
NP_742787.1_617	ndh	no classification	no classification	FAD-dependent pyridine nucleotide-disulfide oxidoreductase	1
NP_743259.1_1089	PP_1098	no classification	no classification	ATP-binding Mrp/Nbp35 family protein	1
NP_746876.1_4706	PP_4770	no classification	no classification	periplasmic ligand-binding sensor protein	1
NP_746491.1_4321	PP_4377	no classification	no classification	flagellin FlaG	1
NP_744239.1_2069	oprF	no classification	no classification	OmpF family protein	1
NP_743295.1_1125	PP_1134	no classification	no classification	FAD dependent oxidoreductase	-1
NP_746900.1_4730	PP_4795	no classification	no classification	rare lipoprotein B	-1
NP_747129.1_4959	pip	no classification	no classification	proline iminopeptidase	-1
NP_743006.1_836	hscB	Protein fate	Protein folding and stabilization	co-chaperone HscB	0
NP_743871.1_1701	fklB-2	Protein fate	Protein folding and stabilization	FKBP-type peptidylprolyl isomerase	0
NP_742934.1_764	PP_0773	Protein fate	Protein and peptide secretion and trafficking	OmpA/MotB domain-containing protein	0
NP_742176.1_6	yidC	Protein fate	Protein and peptide secretion and trafficking	inner membrane protein translocase component YidC	0
NP_745569.1_3399	PP_3431	Protein fate	Degradation of proteins, peptides, and	ThiJ/PfpI domain-containing protein	0

			glycopeptides		
NP_747178.1_5008	PP_5077	Protein fate	Protein and peptide secretion and trafficking	sporulation domain-containing protein	1
NP_746139.1_3969	clpS	Protein fate	Degradation of proteins, peptides, and glycopeptides	ATP-dependent Clp protease adaptor protein ClpS	-1
NP_742767.1_597	fkpB	Protein fate	Protein folding and stabilization	FKBP-type peptidylprolyl isomerase	0
NP_742335.1_165	PP_0166	Protein fate	Protein and peptide secretion and trafficking	HlyD family type I secretion membrane fusion protein	0
NP_745213.1_3043	PP_3069	Protein fate	Protein and peptide secretion and trafficking	outer membrane autotransporter	0
NP_743481.1_1311	sspB	Protein fate	Degradation of proteins, peptides, and glycopeptides	ClpXP protease specificity-enhancing factor	0
NP_742503.1_333	msrA	Protein fate	Protein modification and repair	methionine sulfoxide reductase A	1
NP_744133.1_1963	ibpA	Protein fate	Protein folding and stabilization	heat shock protein Hsp20	1
NP_742784.1_614	PP_0622	Protein fate	Protein and peptide secretion and trafficking	competence lipoprotein ComL	1
NP_747215.1_5045	PP_5114	Protein synthesis	tRNA and rRNA base modification	hypothetical protein	0
NP_742994.1_824	tgt	Protein synthesis	tRNA and rRNA base modification	queuine tRNA-ribosyltransferase	0
NP_746763.1_4593	trmA	Protein synthesis	tRNA and rRNA base modification	tRNA (uracil-5-)-methyltransferase	0
NP_743238.1_1068	PP_1077	Protein synthesis	tRNA aminoacylation	ybaK/ebsC protein	0
NP_742658.1_488	selA	Protein synthesis	tRNA aminoacylation	selenocysteine synthase	0
NP_744727.1_2557	PP_2583	Protein synthesis	Ribosomal proteins: synthesis and modification	acetyltransferase	0
NP_743982.1_1812	PP_1827	Protein synthesis	Ribosomal proteins: synthesis and modification	N5-glutamine S-adenosyl-L-methionine-dependent methyltransferase	0
NP_747383.1_5213	rpmB	Protein synthesis	Ribosomal proteins: synthesis and modification	50S ribosomal protein L28	1
NP_742894.1_724	prfA	Protein synthesis	Translation factors	peptide chain release factor 1	1
NP_744065.1_1895	rpmF	Protein synthesis	Ribosomal proteins: synthesis and modification	50S ribosomal protein L32	1
NP_743620.2_1450	rpsP	Protein synthesis	Ribosomal proteins: synthesis and modification	30S ribosomal protein S16	1
NP_743476.1_1306	rpsI	Protein synthesis	Ribosomal proteins: synthesis and modification	30S ribosomal protein S9	1
NP_743471.1_1301	PP_1311	Protein synthesis	tRNA aminoacylation	tryptophanyl-tRNA synthetase	-1
NP_747204.1_5034	trmB	Protein synthesis	tRNA and rRNA base modification	tRNA (guanine-N(7)-)-methyltransferase	0
NP_747006.1_4836	PP_4903	Protein synthesis	Translation factors	ribosome-associated GTPase	0
NP_742729.1_559	PP_0566	Protein synthesis	Translation factors	translation initiation factor Sui1	0
NP_745545.1_3375	PP_3406	Protein synthesis	Ribosomal proteins: synthesis and modification	acetyltransferase	0
NP_742881.1_711	pth	Protein synthesis	Other	peptidyl-tRNA hydrolase	0
NP_747188.2_5018	rpmE	Protein synthesis	Ribosomal proteins: synthesis and modification	50S ribosomal protein L31	1
NP_742413.1_243	PP_0245	Protein synthesis	Ribosomal proteins: synthesis and modification	S1 RNA-binding domain-containing protein	1
NP_743357.1_1187	rimO	Protein synthesis	Ribosomal proteins: synthesis and modification	30S ribosomal protein S12 methylthiotransferase	-1
NP_742294.1_124	engB	Protein synthesis	Other	ribosome biogenesis GTP-binding protein YsxC	-1
NP_742765.1_595	ileS	Protein synthesis	tRNA aminoacylation	isoleucyl-tRNA synthetase	-1
NP_747001.1_4831	PP_4898	Protein synthesis	tRNA and rRNA base modification	hypothetical protein	-1
NP_746893.1_4723	PP_4788	Protein synthesis	Other	metalloprotease	-1
NP_743093.2_923	gatC	Protein synthesis	tRNA aminoacylation	aspartyl/glutamyl-tRNA amidotransferase subunit C	-1
NP_742641.1_471	rpmJ	Protein synthesis	Ribosomal proteins: synthesis and modification	50S ribosomal protein L36	-1
NP_747099.1_4929	pyrR	Purines, pyrimidines, nucleosides, and nucleotides	Salvage of nucleosides and nucleotides	bifunctional pyrimidine regulatory protein PyrR/uracil phosphoribosyltransferase	0

NP_743337.1_1167	nrdB	Purines, pyrimidines, nucleosides, and nucleotides	2'-Deoxyribonucleotide metabolism	ribonucleotide-diphosphate reductase subunit beta	-1
NP_746382.1_4212	apt	Purines, pyrimidines, nucleosides, and nucleotides	Salvage of nucleosides and nucleotides	adenine phosphoribosyltransferase	0
NP_744245.1_2075	pyrD	Purines, pyrimidines, nucleosides, and nucleotides	Pyrimidine ribonucleotide biosynthesis	dihydroorotate dehydrogenase 2	-1
NP_746478.1_4308	PP_4364	Regulatory functions	Protein interactions	anti-sigma-factor antagonist	0
NP_743586.1_1416	mucA	Regulatory functions	Protein interactions	anti sigma-E protein, RseA	0
NP_747426.1_5256	phoU	Regulatory functions	Other	phosphate transporter PhoU	0
NP_743374.1_1204	PP_1214	Regulatory functions	DNA interactions	hypothetical protein	0
NP_744119.1_1949	PP_1968	Regulatory functions	DNA interactions	TetR family transcriptional regulator	0
NP_743346.1_1176	phoP	Regulatory functions	DNA interactions	winged helix family two component transcriptional regulator	0
NP_744950.1_2780	PP_2806	Regulatory functions	DNA interactions	TetR family transcriptional regulator	0
NP_744695.1_2525	PP_2550	Regulatory functions	DNA interactions	hypothetical protein	0
NP_746997.1_4827	hfq	Regulatory functions	Other	RNA-binding protein Hfq	1
NP_746509.1_4339	flgM	Regulatory functions	Protein interactions	anti-sigma-28 factor FlgM	-1
NP_744653.1_2483	PP_2505	Regulatory functions	Small molecule interactions	GAF domain/GGDEF domain-containing protein	0
NP_743652.1_1482	PP_1494	Regulatory functions	Small molecule interactions	response regulator/GGDEF domain-containing protein	0
NP_742953.1_783	fruR	Regulatory functions	DNA interactions	DNA-binding transcriptional regulator FruR	0
NP_744293.1_2123	PP_2144	Regulatory functions	DNA interactions	TetR family transcriptional regulator	0
NP_745051.1_2881	PP_2907	Regulatory functions	DNA interactions	winged helix family two component transcriptional regulator	0
NP_743218.1_1048	PP_1057	Regulatory functions	DNA interactions	PadR family transcriptional regulator	0
NP_744605.1_2435	rbsR	Regulatory functions	DNA interactions	ribose operon repressor	0
NP_743585.1_1415	algU	Regulatory functions	DNA interactions	RNA polymerase sigma factor AlgU	0
NP_742303.1_133	algB	Regulatory functions	DNA interactions	Fis family transcriptional regulator	1
NP_745520.1_3350	ptxS	Regulatory functions	DNA interactions	PtxS family transcriptional regulator	1
NP_743621.1_1451	rimM	Transcription	RNA processing	16S rRNA-processing protein RimM	0
NP_746802.1_4632	PP_4693	Transcription	Transcription factors	C4-type zinc finger DksA/TraR family protein	1
NP_747395.1_5225	rph	Transcription	RNA processing	ribonuclease PH	0
NP_743935.1_1765	PP_1779	Transport and binding proteins	Other	lipopolysaccharide ABC export system, ATP-binding protein	0
NP_747384.1_5214	PP_5283	Transport and binding proteins	Cations and iron carrying compounds	peptide ABC transporter substrate-binding protein	0
NP_744602.1_2432	rbsB	Transport and binding proteins	Carbohydrates, organic alcohols, and acids	monosaccharide-transporting ATPase	0
NP_742578.1_408	PP_0412	Transport and binding proteins	Amino acids, peptides and amines	polyamine ABC transporter substrate-binding protein	0
NP_747281.1_5111	potF-1	Transport and binding proteins	Amino acids, peptides and amines	spermidine/putrescine ABC transporter substrate-binding protein	0
NP_743883.1_1713	PP_1726	Transport and binding proteins	Amino acids, peptides and amines	ABC transporter substrate-binding protein	0
NP_746597.1_4427	PP_4486	Transport and binding proteins	Amino acids, peptides and amines	amino acid ABC transporter substrate-binding protein	0
NP_743229.1_1059	PP_1068	Transport and binding proteins	Amino acids, peptides and amines	amino acid ABC transporter ATP-binding protein	0
NP_747297.1_5127	PP_5196	Transport and binding proteins	Amino acids, peptides and amines	iron ABC transporter substrate-binding protein	-1
NP_743545.1_1375	ttgA	Transport and binding proteins	Unknown substrate	RND family efflux transporter MFP subunit	0
NP_745555.1_3385	gntP	Transport and binding proteins	Carbohydrates, organic alcohols, and acids	gluconate transporter	0

NP_743119.1_949	ttg2A	Transport and binding proteins	Anions	toluene tolerance ABC efflux transporter ATP-binding protein	0
NP_747280.1_5110	potG	Transport and binding proteins	Amino acids, peptides and amines	spermidine/putrescine ABC transporter ATPase	0
NP_743046.1_876	PP_0885	Transport and binding proteins	Cations and iron carrying compounds	peptide ABC transporter substrate-binding protein	1
NP_742189.1_19	PP_0019	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_744310.1_2140	PP_2161	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_746679.1_4509	PP_4570	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_744460.1_2290	PP_2311	Unknown function	Enzymes of unknown specificity	TatD family hydrolase	0
NP_742257.1_87	PP_0087	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_743069.1_899	PP_0908	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_743700.1_1530	PP_1543	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_744426.1_2256	PP_2277	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_744590.1_2420	PP_2442	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_746645.1_4475	PP_4535	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_743384.1_1214	PP_1224	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_747302.1_5132	PP_5201	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_742466.1_296	PP_0299	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_742589.1_419	PP_0423	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_743820.1_1650	PP_1663	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_743675.1_1505	PP_1518	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_744255.1_2085	PP_2105	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_746039.1_3869	PP_3909	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_746623.1_4453	PP_4513	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_743016.1_846	PP_0855	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_746725.1_4555	PP_4616	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_747454.1_5284	PP_5353	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_742834.1_664	PP_0673	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_742358.1_188	hemY	Unknown function	General	HemY domain-containing protein	0
NP_742855.1_685	PP_0694	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_745583.1_3413	PP_3445	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_745763.1_3593	PP_3627	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_745825.1_3655	PP_3692	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_746702.1_4532	PP_4593	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_747053.1_4883	PP_4950	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_747413.1_5243	PP_5312	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_743350.1_1180	PP_1190	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_743956.1_1786	wbpY	Unknown function	Enzymes of unknown specificity	glycosyl transferase WbpY	0
NP_743080.1_910	PP_0919	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_743110.1_940	PP_0949	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_743645.1_1475	PP_1487	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_744156.1_1986	PP_2006	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_744203.1_2033	PP_2053	Unknown function	Enzymes of unknown specificity	hypothetical protein	0

NP_745549.1_3379	PP_3411	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_745827.1_3657	PP_3694	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_746272.1_4102	PP_4154	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_746317.1_4147	PP_4200	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_747013.1_4843	PP_4910	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_744249.1_2079	PP_2099	Unknown function	Enzymes of unknown specificity	hypothetical protein	0
NP_744711.1_2541	PP_2566	Unknown function	Enzymes of unknown specificity	hypothetical protein	1
NP_747396.2_5226	PP_5295	Unknown function	Enzymes of unknown specificity	hypothetical protein	-1

a: Up/down regulation of protein expression, the value 0: exclusively expressed in imipenem stress; Value 1: up regulated in H_2O_2 stress; Value -1: down regulated in imipenem stress.

S-3: Differential expressed proteins in LB medium

accession	gene	TIGRFAM main role	TIGRFAM subrole	protein description	LB-M9[a]
NP_743924.1_1754	serC	Amino acid biosynthesis	Serine family	phosphoserine aminotransferase	-1
NP_743421.1_1251	PP_1261	Amino acid biosynthesis	Serine family	2-hydroxyacid dehydrogenase	-1
NP_743274.1_1104	PP_1113	Amino acid biosynthesis	Serine family	pyridoxal-phosphate dependent enzyme family protein	-1
NP_744266.1_2096	pdxB	Amino acid biosynthesis	Serine family	erythronate-4-phosphate dehydrogenase	-1
NP_745516.1_3346	kguD	Amino acid biosynthesis	Serine family	2-ketogluconate 6-phosphate reductase	-1
NP_744136.1_1966	leuC	Amino acid biosynthesis	Pyruvate family	isopropylmalate isomerase large subunit	-1
NP_743186.1_1016	leuA	Amino acid biosynthesis	Pyruvate family	2-isopropylmalate synthase	-1
NP_745648.1_3478	ilvE	Amino acid biosynthesis	Pyruvate family	branched-chain amino acid aminotransferase	-1
NP_747229.1_5059	ilvD	Amino acid biosynthesis	Pyruvate family	dihydroxy-acid dehydratase	-1
NP_746789.1_4619	ilvB	Amino acid biosynthesis	Pyruvate family	acetolactate synthase 3 catalytic subunit	-1
NP_745301.1_3131	PP_3157	Amino acid biosynthesis	Histidine family	inositol monophosphatase	-1
NP_747116.1_4946	hisE	Amino acid biosynthesis	Histidine family	phosphoribosyl-ATP pyrophosphatase	-1
NP_747196.1_5026	proC-2	Amino acid biosynthesis	Glutamate family	pyrroline-5-carboxylate reductase	-1
NP_745908.1_3738	proC-1	Amino acid biosynthesis	Glutamate family	pyrroline-5-carboxylate reductase	-1
NP_742852.1_682	proB	Amino acid biosynthesis	Glutamate family	gamma-glutamyl kinase	-1
NP_747284.1_5114	PP_5183	Amino acid biosynthesis	Glutamate family	glutamine synthetase	-1
NP_743161.1_991	argI	Amino acid biosynthesis	Glutamate family	ornithine carbamoyltransferase	0
NP_747287.1_5117	argE	Amino acid biosynthesis	Glutamate family	acetylornithine deacetylase	-1
NP_743629.1_1459	thrC	Amino acid biosynthesis	Aspartate family	threonine synthase	-1
NP_742291.1_121	thrB	Amino acid biosynthesis	Aspartate family	homoserine kinase	-1
NP_744143.1_1973	PP_1992	Amino acid biosynthesis	Aspartate family	aspartate-semialdehyde dehydrogenase	-1
NP_747199.1_5029	metW	Amino acid biosynthesis	Aspartate family	methionine biosynthesis protein MetW	0
NP_747070.1_4900	metK	Amino acid biosynthesis	Aspartate family	S-adenosylmethionine synthetase	-1

NP_747329.1_5159	dapF	Amino acid biosynthesis	Aspartate family	diaminopimelate epimerase	0
NP_745920.1_3750	dapF	Amino acid biosynthesis	Aspartate family	diaminopimelate epimerase	-1
NP_743682.1_1512	dapE	Amino acid biosynthesis	Aspartate family	succinyl-diaminopimelate desuccinylase	0
NP_743397.1_1227	dapA	Amino acid biosynthesis	Aspartate family	dihydrodipicolinate synthase	-1
NP_744140.1_1970	asd	Amino acid biosynthesis	Aspartate family	aspartate-semialdehyde dehydrogenase	-1
NP_742588.1_418	trpC	Amino acid biosynthesis	Aromatic amino acid family	indole-3-glycerol phosphate synthase	-1
NP_742252.1_82	trpA	Amino acid biosynthesis	Aromatic amino acid family	tryptophan synthase subunit alpha	-1
NP_743926.1_1756	PP_1770	Amino acid biosynthesis	Aromatic amino acid family	bifunctional cyclohexadienyl dehydrogenase/ 3-phosphoshikimate 1-carboxyvinyltransferase	-1
NP_747180.1_5010	aroK	Amino acid biosynthesis	Aromatic amino acid family	shikimate kinase	-1
NP_745898.2_3728	aroE	Amino acid biosynthesis	Aromatic amino acid family	shikimate 5-dehydrogenase	-1
NP_743985.1_1815	aroC	Amino acid biosynthesis	Aromatic amino acid family	chorismate synthase	0
NP_747179.1_5009	aroB	Amino acid biosynthesis	Aromatic amino acid family	3-dehydroquinate synthase	-1
NP_747025.1_4855	thiC	Biosynthesis of cofactors, prosthetic groups, and carriers	Thiamine	thiamine biosynthesis protein ThiC	-1
NP_747371.1_5201	dadA-2	Biosynthesis of cofactors, prosthetic groups, and carriers	Thiamine	D-amino acid dehydrogenase small subunit	0
NP_746974.1_4804	nadE	Biosynthesis of cofactors, prosthetic groups, and carriers	Pyridine nucleotides	NAD synthetase	-1
NP_743584.1_1414	nadB	Biosynthesis of cofactors, prosthetic groups, and carriers	Pyridine nucleotides	L-aspartate oxidase	-1
NP_746808.1_4638	panB	Biosynthesis of cofactors, prosthetic groups, and carriers	Pantothenate and coenzyme A	3-methyl-2-oxobutanoate hydroxymethyltransferase	0
NP_747386.1_5216	coaBC	Biosynthesis of cofactors, prosthetic groups, and carriers	Pantothenate and coenzyme A	bifunctional phosphopantothenoylcysteine decarboxylase/phosphopantothenate synthase	0
NP_742545.1_375	pqqC	Biosynthesis of cofactors, prosthetic groups, and carriers	Other	pyrroloquinoline quinone biosynthesis protein PqqC	-1
NP_747213.1_5043	PP_5112	Biosynthesis of cofactors, prosthetic groups, and carriers	Other	peptidase M16 domain-containing protein	0
NP_743008.1_838	PP_0847	Biosynthesis of cofactors, prosthetic groups, and carriers	Other	(2Fe-2S) ferredoxin	-1
NP_743004.1_834	iscU	Biosynthesis of cofactors, prosthetic groups, and carriers	Other	scaffold protein	-1
NP_743003.1_833	iscS	Biosynthesis of cofactors, prosthetic groups, and carriers	Other	cysteine desulfurase	-1
NP_743005.1_835	iscA	Biosynthesis of cofactors, prosthetic groups, and carriers	Other	iron-sulfur cluster assembly protein IscA	-1
NP_744272.1_2102	moeA	Biosynthesis of cofactors, prosthetic groups, and carriers	Molybdopterin	molybdopterin biosynthesis MoeA protein	-1
NP_744271.1_2101	moaB-1	Biosynthesis of cofactors, prosthetic groups, and carriers	Molybdopterin	molybdenum cofactor biosynthesis protein B	0
NP_743921.1_1751	ubiG	Biosynthesis of cofactors, prosthetic groups, and carriers	Menaquinone and ubiquinone	3-demethylubiquinone-9 3-methyltransferase	0
NP_747113.1_4943	ubiE	Biosynthesis of cofactors, prosthetic groups, and carriers	Menaquinone and ubiquinone	ubiquinone/menaquinone biosynthesis methyltransferase	0
NP_744138.1_1968	PP_1987	Biosynthesis of cofactors, prosthetic groups, and carriers	Menaquinone and ubiquinone	UbiE/COQ5 family methlytransferase	-1

NP_742924.1_754	PP_0763	Biosynthesis of cofactors, prosthetic groups, and carriers	Menaquinone and ubiquinone	long-chain-fatty-acid--CoA ligase	-1
NP_746748.1_4578	PP_4639	Biosynthesis of cofactors, prosthetic groups, and carriers	Heme, porphyrin, and cobalamin	CobW/P47K family protein	-1
NP_746697.1_4527	PP_4588	Biosynthesis of cofactors, prosthetic groups, and carriers	Heme, porphyrin, and cobalamin	nitroreductase	-1
NP_742905.1_735	hemH	Biosynthesis of cofactors, prosthetic groups, and carriers	Heme, porphyrin, and cobalamin	ferrochelatase	-1
NP_742893.1_723	hemA	Biosynthesis of cofactors, prosthetic groups, and carriers	Heme, porphyrin, and cobalamin	glutamyl-tRNA reductase	0
NP_743836.1_1666	cobT	Biosynthesis of cofactors, prosthetic groups, and carriers	Heme, porphyrin, and cobalamin	nicotinate-nucleotide--dimethylbenzimidazole phosphoribosyltransferase	-1
NP_746933.1_4763	cobH	Biosynthesis of cofactors, prosthetic groups, and carriers	Heme, porphyrin, and cobalamin	precorrin-8X methylmutase	0
NP_747095.1_4925	gshB	Biosynthesis of cofactors, prosthetic groups, and carriers	Glutathione and analogs	glutathione synthetase	-1
NP_742411.1_241	gshA	Biosynthesis of cofactors, prosthetic groups, and carriers	Glutathione and analogs	glutamate--cysteine ligase	-1
NP_746768.1_4598	ggt-2	Biosynthesis of cofactors, prosthetic groups, and carriers	Glutathione and analogs	gamma-glutamyltransferase	-1
NP_744147.1_1977	folC	Biosynthesis of cofactors, prosthetic groups, and carriers	Folic acid	bifunctional folylpolyglutamate synthase/ dihydrofolate synthase	-1
NP_746320.1_4150	PP_4203	Biosynthesis of cofactors, prosthetic groups, and carriers	Chlorophyll and bacteriochlorphyll	electron-transferring-flavoprotein dehydrogenase	0
NP_747283.1_5113	PP_5182	Biosynthesis of cofactors, prosthetic groups, and carriers	Biotin	aminotransferase	-1
NP_743681.1_1511	PP_1524	Biosynthesis of cofactors, prosthetic groups, and carriers	Biotin	rRNA (guanine-N(1)-)-methyltransferase	0
NP_742531.1_361	bioH	Biosynthesis of cofactors, prosthetic groups, and carriers	Biotin	carboxylesterase	-1
NP_742529.1_359	bioB	Biosynthesis of cofactors, prosthetic groups, and carriers	Biotin	biotin synthase	-1
NP_747266.1_5096	PP_5165	Cell envelope	Other	NLPA lipoprotein	0
NP_745894.1_3724	PP_3764	Cell envelope	Other	porin	0
NP_742960.1_790	PP_0799	Cell envelope	Other	porin	0
NP_742402.1_232	oprE	Cell envelope	Other	porin	1
NP_743366.1_1196	oprD	Cell envelope	Other	porin	-1
NP_743180.1_1010	oprB-1	Cell envelope	Other	porin	-1
NP_742510.1_340	waaG	Cell envelope	Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides	lipopolysaccharide core biosynthesis protein WaaG	0
NP_742508.1_338	waaF	Cell envelope	Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides	lipopolysaccharide heptosyltransferase II	0
NP_743949.1_1779	PP_1793	Cell envelope	Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides	glycosyl transferase family protein	0
NP_743932.1_1762	PP_1776	Cell envelope	Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides	mannose-6-phosphate isomerase	0
NP_743117.1_947	PP_0956	Cell envelope	Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides	YrbI family phosphatase	-1

NP_742664.1_494	PP_0500	Cell envelope	Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides	dTDP-4-dehydrorhamnose reductase	0
NP_743758.1_1588	lpxD	Cell envelope	Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase	0
NP_747000.1_4830	PP_4897	Cell envelope	Biosynthesis and degradation of murein sacculus and peptidoglycan	N-acetylmuramoyl-L-alanine amidase	0
NP_746903.1_4733	PP_4798	Cell envelope	Biosynthesis and degradation of murein sacculus and peptidoglycan	lytic murein transglycosylase	0
NP_743492.1_1322	murF	Cell envelope	Biosynthesis and degradation of murein sacculus and peptidoglycan	UDP-N-acetylmuramoylalanyl-D-glutamyl-2,6-diaminopimelate--D-alanyl-D-alanyl ligase	-1
NP_743491.1_1321	murE	Cell envelope	Biosynthesis and degradation of murein sacculus and peptidoglycan	UDP-N-acetylmuramoylalanyl-D-glutamate--2,6-diaminopimelate ligase	-1
NP_743494.1_1324	murD	Cell envelope	Biosynthesis and degradation of murein sacculus and peptidoglycan	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate synthetase	-1
NP_743497.1_1327	murC	Cell envelope	Biosynthesis and degradation of murein sacculus and peptidoglycan	UDP-N-acetylmuramate--L-alanine ligase	1
NP_742710.1_540	mpl	Cell envelope	Biosynthesis and degradation of murein sacculus and peptidoglycan	UDP-N-acetylmuramate	0
NP_746910.1_4740	mltB	Cell envelope	Biosynthesis and degradation of murein sacculus and peptidoglycan	lytic murein transglycosylase B	0
NP_743544.1_1374	ttgB	Cellular processes	Toxin production and resistance	hydrophobe/amphiphile efflux-1 (HAE1) family transporter	0
NP_746648.1_4478	PP_4539	Cellular processes	Toxin production and resistance	LysR family transcriptional regulator	-1
NP_746431.1_4261	PP_4315	Cellular processes	Toxin production and resistance	PhzF family phenazine biosynthesis protein	-1
NP_745564.1_3394	mexF	Cellular processes	Toxin production and resistance	hydrophobe/amphiphile efflux-1 (HAE1) family transporter	-1
NP_745563.1_3393	mexE	Cellular processes	Toxin production and resistance	RND family efflux transporter MFP subunit	-1
NP_742354.1_184	pprA	Cellular processes	Sporulation and germination	LytTR family two component transcriptional regulator	-1
NP_745633.1_3463	PP_3496	Cellular processes	Other	endoribonuclease L-PSP	-1
NP_744292.1_2122	lexA-1	Cellular processes	DNA interactions	LexA repressor	0
NP_743076.1_906	sodB	Cellular processes	Detoxification	superoxide dismutase	-1
NP_745804.1_3634	PP_3668	Cellular processes	Detoxification	catalase/peroxidase HPI	-1
NP_743773.1_1603	PP_1616	Cellular processes	Detoxification	D-isomer specific 2-hydroxyacid dehydrogenase	-1
NP_743633.1_1463	PP_1475	Cellular processes	Detoxification	L-carnitine dehydratase/bile acid-inducible protein F	-1
NP_744014.1_1844	ohr	Cellular processes	Detoxification	OsmC family protein	-1
NP_742403.1_233	lsfA	Cellular processes	Detoxification	peroxidase	-1
NP_742647.1_477	katA	Cellular processes	Detoxification	catalase	0
NP_746262.1_4092	gloB	Cellular processes	Detoxification	hydroxyacylglutathione hydrolase	-1
NP_743802.1_1632	arsC-1	Cellular processes	Detoxification	arsenate reductase	-1
NP_743530.1_1360	PP_1371	Cellular processes	Chemotaxis and motility	methyl-accepting chemotaxis transducer	0
NP_746471.1_4301	fliN	Cellular processes	Chemotaxis and motility	flagellar motor switch protein	1
NP_746482.2_4312	fliG	Cellular processes	Chemotaxis and motility	flagellar motor switch protein G	0
NP_746490.1_4320	fliD	Cellular processes	Chemotaxis and motility	flagellar cap protein FliD	-1
NP_746492.1_4322	fliC	Cellular processes	Chemotaxis and motility	flagellin FliC	1
NP_746457.1_4287	fleN	Cellular processes	Chemotaxis and motility	flagellar number regulator FleN	-1
NP_747191.1_5021	PP_5090	Cellular processes	Cell division	sporulation domain-containing protein	0
NP_746767.1_4597	PP_4658	Cellular processes	Cell division	methyl-accepting chemotaxis sensory transducer	0
NP_743889.1_1719	minD	Cellular processes	Cell division	septum site-determining protein MinD	-1
NP_743890.1_1720	minC	Cellular processes	Cell division	septum formation inhibitor	-1
NP_746826.1_4656	ftsH	Cellular processes	Cell division	ATP-dependent metalloprotease FtsH	1

NP_742725.1_555	ctpL	Cellular processes	Cell division	methyl-accepting chemotaxis sensory transducer	0
NP_747403.1_5233	spoT	Cellular processes	Adaptations to atypical conditions	(p)ppGpp synthetase I SpoT/RelA	0
NP_743780.1_1610	rpoS	Cellular processes	Adaptations to atypical conditions	RNA polymerase sigma factor RpoS	-1
NP_747379.1_5209	PP_5278	Cellular processes	Adaptations to atypical conditions	aldehyde dehydrogenase	1
NP_745581.1_3411	PP_3443	Cellular processes	Adaptations to atypical conditions	glyceraldehyde-3-phosphate dehydrogenase	-1
NP_744824.1_2654	PP_2680	Cellular processes	Adaptations to atypical conditions	aldehyde dehydrogenase	-1
NP_744635.1_2465	PP_2487	Cellular processes	Adaptations to atypical conditions	aldehyde dehydrogenase	0
NP_742708.1_538	PP_0545	Cellular processes	Adaptations to atypical conditions	aldehyde dehydrogenase	-1
NP_747165.1_4995	betA	Cellular processes	Adaptations to atypical conditions	choline dehydrogenase	0
NP_745774.1_3604	PP_3638	Central intermediary metabolism	Sulfur metabolism	acyl-CoA dehydrogenase	-1
NP_743463.1_1293	cysD	Central intermediary metabolism	Sulfur metabolism	sulfate adenylyltransferase subunit 2	-1
NP_742730.1_560	speA	Central intermediary metabolism	Polyamine biosynthesis	arginine decarboxylase	1
NP_742433.1_263	PP_0266	Central intermediary metabolism	Polyamine biosynthesis	agmatine deiminase	0
NP_743908.1_1738	PP_1752	Central intermediary metabolism	Phosphorus compounds	hypothetical protein	0
NP_742935.1_765	pta	Central intermediary metabolism	Other	phosphate acetyltransferase	0
NP_744483.1_2313	prpB	Central intermediary metabolism	Other	2-methylisocitrate lyase	-1
NP_744636.1_2466	PP_2488	Central intermediary metabolism	Other	succinic-semialdehyde dehydrogenase	-1
NP_743304.1_1134	PP_1143	Central intermediary metabolism	Other	3-hydroxyisobutyrate dehydrogenase	-1
NP_746415.1_4245	glxR	Central intermediary metabolism	Other	2-hydroxy-3-oxopropionate reductase	-1
NP_742382.1_212	gabT	Central intermediary metabolism	Other	4-aminobutyrate aminotransferase	1
NP_742507.1_337	glnE	Central intermediary metabolism	Nitrogen metabolism	bifunctional glutamine-synthetase adenylyltransferase/deadenyltransferase	-1
NP_743746.1_1576	glnD	Central intermediary metabolism	Nitrogen metabolism	PII uridylyl-transferase	-1
NP_742381.1_211	gabD	Central intermediary metabolism	Nitrogen metabolism	succinate-semialdehyde dehydrogenase I	1
NP_746926.1_4756	fis	Central intermediary metabolism	Nitrogen metabolism	Fis family transcriptional regulator	1
NP_746507.1_4337	cheV-3	Central intermediary metabolism	Nitrogen metabolism	chemotaxis protein CheV	-1
NP_746848.1_4678	hsdR	DNA metabolism	Restriction/modification	type I restriction-modification system, R subunit	-1
NP_743134.1_964	ndpA	DNA metabolism	Other	nucleoid-associated protein NdpA	0
NP_747453.1_5283	uvrD	DNA metabolism	DNA replication, recombination, and repair	DNA-dependent helicase II	1
NP_742651.1_481	ssb	DNA metabolism	DNA replication, recombination, and repair	single-stranded DNA-binding protein	0
NP_743524.1_1354	sbcB	DNA metabolism	DNA replication, recombination, and repair	exonuclease I	0
NP_743376.1_1206	ruvA	DNA metabolism	DNA replication, recombination, and repair	Holliday junction DNA helicase RuvA	0
NP_747365.1_5195	rep	DNA metabolism	DNA replication, recombination, and repair	ATP-dependent DNA helicase Rep	0
NP_743786.1_1616	recA	DNA metabolism	DNA replication, recombination, and repair	recombinase A	1
NP_747082.1_4912	PP_4980	DNA metabolism	DNA replication, recombination, and repair	DEAD/DEAH box helicase	1
NP_745824.1_3654	PP_3691	DNA metabolism	DNA replication, recombination, and repair	DNA helicase-like protein	-1
NP_744619.1_2449	ihfA	DNA metabolism	DNA replication, recombination, and repair	integration host factor subunit alpha	-1
NP_746385.1_4215	dnaX	DNA metabolism	DNA replication, recombination, and repair	DNA polymerase III subunits gamma and tau	0

NP_743286.1_1116	dinG	DNA metabolism	DNA replication, recombination, and repair	ATP-dependent DNA helicase DinG	0
NP_747055.1_4885	dbpA	DNA metabolism	DNA replication, recombination, and repair	ATP-dependent RNA helicase DbpA	0
NP_746140.1_3970	cspD	DNA metabolism	DNA replication, recombination, and repair	cold-shock protein CspD	-1
NP_743679.1_1509	cspA-1	DNA metabolism	DNA replication, recombination, and repair	cold shock protein CspA	-1
NP_746309.1_4139	sdhD	Energy metabolism	TCA cycle	succinate dehydrogenase, hydrophobic membrane anchor protein	0
NP_746310.1_4140	sdhC	Energy metabolism	TCA cycle	succinate dehydrogenase, cytochrome b556 subunit	1
NP_743411.2_1241	mqo-2	Energy metabolism	TCA cycle	malate:quinone oxidoreductase	1
NP_746141.1_3971	icd	Energy metabolism	TCA cycle	isocitrate dehydrogenase	-1
NP_742523.1_353	glcB	Energy metabolism	TCA cycle	malate synthase G	-1
NP_746235.1_4065	aceA	Energy metabolism	TCA cycle	isocitrate lyase	-1
NP_745518.1_3348	kguK	Energy metabolism	Sugars	PfkB domain-containing protein	-1
NP_743183.1_1013	zwf-1	Energy metabolism	Pentose phosphate pathway	glucose-6-phosphate 1-dehydrogenase	-1
NP_744317.1_2147	tal	Energy metabolism	Pentose phosphate pathway	transaldolase B	-1
NP_747251.1_5081	rpiA	Energy metabolism	Pentose phosphate pathway	ribose-5-phosphate isomerase A	0
NP_742581.1_411	rpe	Energy metabolism	Pentose phosphate pathway	ribulose-phosphate 3-epimerase	-1
NP_743184.1_1014	pgl	Energy metabolism	Pentose phosphate pathway	6-phosphogluconolactonase	-1
NP_747255.1_5085	PP_5154	Energy metabolism	Other	FAD linked oxidase domain-containing protein	-1
NP_746755.1_4585	PP_4646	Energy metabolism	Other	oxidoreductase FAD/NAD(P)-binding domain-containing protein	-1
NP_743795.1_1625	fpr	Energy metabolism	Other	oxidoreductase FAD/NAD(P)-binding domain-containing protein	-1
NP_746598.1_4428	acsA	Energy metabolism	Other	acetyl-CoA synthetase	-1
NP_743170.1_1000	gap-1	Energy metabolism	Glycolysis/gluconeogenesis	glyceraldehyde-3-phosphate dehydrogenase, type I	-1
NP_744232.1_2062	ppsA	Energy metabolism	Glycolysis/gluconeogenesis	phosphoenolpyruvate synthase	-1
NP_743172.1_1002	glk	Energy metabolism	Glycolysis/gluconeogenesis	glucokinase	-1
NP_743769.1_1599	eno	Energy metabolism	Glycolysis/gluconeogenesis	phosphopyruvate hydratase	-1
NP_744574.1_2404	PP_2426	Energy metabolism	Fermentation	D-isomer specific 2-hydroxyacid dehydrogenase	0
NP_743171.1_1001	edd	Energy metabolism	Entner-Doudoroff	phosphogluconate dehydratase	-1
NP_747170.1_5000	trxC	Energy metabolism	Electron transport	thioredoxin	-1
NP_745759.1_3589	PP_3623	Energy metabolism	Electron transport	gluconate 2-dehydrogenase	-1
NP_745522.1_3352	PP_3382	Energy metabolism	Electron transport	gluconate 2-dehydrogenase	-1
NP_746245.1_4075	nuoI	Energy metabolism	Electron transport	NADH dehydrogenase subunit I	1
NP_742974.1_804	cyoB	Energy metabolism	Electron transport	cytochrome o ubiquinol oxidase subunit I	0
NP_742973.1_803	cyoA	Energy metabolism	Electron transport	ubiquinol oxidase subunit II	0
NP_746181.1_4011	glgX	Energy metabolism	Biosynthesis and degradation of polysaccharides	glycogen debranching protein GlgX	-1
NP_746184.1_4014	glgB	Energy metabolism	Biosynthesis and degradation of polysaccharides	glycogen branching protein	-1
NP_743234.1_1064	glpD	Energy metabolism	Anaerobic	glycerol-3-phosphate dehydrogenase	0
NP_745288.1_3118	sda-3	Energy metabolism	Amino acids and amines	iron-sulfur-dependent L-serine dehydratase	0
NP_747050.1_4880	putA	Energy metabolism	Amino acids and amines	trifunctional transcriptional regulator/proline dehydrogenase/pyrroline-5-carboxylate dehydrogenase	1
NP_746703.1_4533	PP_4594	Energy metabolism	Amino acids and amines	cystathionine gamma-lyase	-1
NP_745905.1_3735	PP_3775	Energy metabolism	Amino acids and amines	sarcosine oxidase	-1

NP_745873.1_3703	PP_3742	Energy metabolism	Amino acids and amines	glutathione S-transferase	-1
NP_743506.1_1336	PP_1347	Energy metabolism	Amino acids and amines	glutathione S-transferase	-1
NP_742352.1_182	PP_0183	Energy metabolism	Amino acids and amines	glutathione S-transferase	-1
NP_746602.1_4432	phhB	Energy metabolism	Amino acids and amines	pterin-4-alpha-carbinolamine dehydratase	0
NP_746601.1_4431	phhA	Energy metabolism	Amino acids and amines	phenylalanine 4-monooxygenase	0
NP_742760.1_590	mmsA-1	Energy metabolism	Amino acids and amines	methylmalonate-semialdehyde dehydrogenase	-1
NP_745571.1_3401	hpd	Energy metabolism	Amino acids and amines	4-hydroxyphenylpyruvate dioxygenase	1
NP_742836.1_666	gdhA	Energy metabolism	Amino acids and amines	glutamate dehydrogenase	-1
NP_743147.1_977	gcvT-1	Energy metabolism	Amino acids and amines	glycine cleavage system T protein	-1
NP_747295.1_5125	gcvT	Energy metabolism	Amino acids and amines	glycine cleavage system aminomethyltransferase T	-1
NP_743149.1_979	gcvP-1	Energy metabolism	Amino acids and amines	glycine dehydrogenase	-1
NP_747294.1_5124	gcvH	Energy metabolism	Amino acids and amines	glycine cleavage system protein H	-1
NP_746589.1_4419	astD	Energy metabolism	Amino acids and amines	succinylglutamic semialdehyde dehydrogenase	0
NP_747439.1_5269	aspA	Energy metabolism	Amino acids and amines	aspartate ammonia-lyase	1
NP_746590.1_4420	aruG	Energy metabolism	Amino acids and amines	arginine N-succinyltransferase subunit beta	0
NP_746592.1_4422	argD	Energy metabolism	Amino acids and amines	bifunctional N-succinyldiaminopimelate-aminotransferase/acetylornithine transaminase	-1
NP_744601.1_2431	ansA	Energy metabolism	Amino acids and amines	type II L-asparaginase	-1
NP_745884.1_3714	PP_3754	Fatty acid and phospholipid metabolism	Other	beta-ketothiolase	-1
NP_746868.1_4698	tesB	Fatty acid and phospholipid metabolism	Biosynthesis	acyl-CoA thioesterase	0
NP_747011.1_4841	psd	Fatty acid and phospholipid metabolism	Biosynthesis	phosphatidylserine decarboxylase	0
NP_747109.1_4939	PP_5007	Fatty acid and phospholipid metabolism	Biosynthesis	poly(hydroxyalkanoate) granule-associated protein	-1
NP_745917.1_3747	PP_3787	Fatty acid and phospholipid metabolism	Biosynthesis	hypothetical protein	-1
NP_745684.1_3514	PP_3547	Fatty acid and phospholipid metabolism	Biosynthesis	short chain dehydrogenase/reductase oxidoreductase	-1
NP_744457.1_2287	PP_2308	Fatty acid and phospholipid metabolism	Biosynthesis	acyl-CoA thioesterase	-1
NP_742917.1_747	PP_0756	Fatty acid and phospholipid metabolism	Biosynthesis	phospholipid/glycerol acyltransferase	0
NP_747105.1_4935	phaA	Fatty acid and phospholipid metabolism	Biosynthesis	poly(3-hydroxyalkanoate) polymerase 1	-1
NP_746659.1_4489	fadD2	Fatty acid and phospholipid metabolism	Biosynthesis	long-chain-fatty-acid--CoA ligase	-1
NP_746658.1_4488	fadD	Fatty acid and phospholipid metabolism	Biosynthesis	long-chain-fatty-acid--CoA ligase	-1
NP_744366.1_2196	fadB1x	Fatty acid and phospholipid metabolism	Biosynthesis	enoyl-CoA hydratase	-1
NP_744070.1_1900	fabF	Fatty acid and phospholipid metabolism	Biosynthesis	3-oxoacyl-ACP synthase	1
NP_744067.1_1897	fabD	Fatty acid and phospholipid metabolism	Biosynthesis	malonyl CoA-ACP transacylase	0
NP_747448.1_5278	accC-2	Fatty acid and phospholipid metabolism	Biosynthesis	pyruvate carboxylase subunit A	-1
NP_743764.1_1594	accA	Fatty acid and phospholipid metabolism	Biosynthesis	acetyl-CoA carboxylase carboxyltransferase subunit alpha	1
NP_742172.1_2	PP_0002	Mobile and extrachromosomal element functions	Plasmid functions	ParA family protein	0
NP_745163.1_2993	PP_3019	Protein fate	Protein modification and repair	nitrilase/cyanide hydratase and apolipoprotein N-acyltransferase	-1
NP_744387.1_2217	PP_2238	Protein fate	Protein modification and repair	peptidase M24	-1
NP_743137.1_967	PP_0976	Protein fate	Protein modification and repair	SAM-dependent methyltransferase	0
NP_743020.1_850	PP_0859	Protein fate	Protein modification and repair	nitrilase/cyanide hydratase and apolipoprotein N-acyltransferase	-1

NP_742922.1_752	PP_0761	Protein fate	Protein modification and repair	rRNA (guanine-N(2)-)-methyltransferase	0
NP_747301.1_5131	pepP	Protein fate	Protein modification and repair	peptidase M24	-1
NP_742503.1_333	msrA	Protein fate	Protein modification and repair	methionine sulfoxide reductase A	-1
NP_746954.1_4784	PP_4849	Protein fate	Protein folding and stabilization	molecular chaperone DnaK	0
NP_743007.1_837	hscA	Protein fate	Protein folding and stabilization	chaperone protein HscA	-1
NP_743871.1_1701	fklB-2	Protein fate	Protein folding and stabilization	FKBP-type peptidylprolyl isomerase	0
NP_746834.1_4664	dnaJ	Protein fate	Protein folding and stabilization	molecular chaperone DnaJ	1
NP_743588.1_1418	algY	Protein fate	Protein folding and stabilization	protease Do	-1
NP_742176.1_6	yidC	Protein fate	Protein and peptide secretion and trafficking	inner membrane protein translocase component YidC	0
NP_742640.1_470	secY	Protein fate	Protein and peptide secretion and trafficking	preprotein translocase subunit SecY	1
NP_746559.1_4389	PP_4448	Protein fate	Protein and peptide secretion and trafficking	hypothetical protein	-1
NP_745251.1_3081	PP_3107	Protein fate	Protein and peptide secretion and trafficking	hypothetical protein	-1
NP_745250.1_3080	PP_3106	Protein fate	Protein and peptide secretion and trafficking	hypothetical protein	-1
NP_745248.1_3078	PP_3104	Protein fate	Protein and peptide secretion and trafficking	hypothetical protein	-1
NP_745239.1_3069	PP_3095	Protein fate	Protein and peptide secretion and trafficking	chaperone-associated ATPase	-1
NP_745238.1_3068	PP_3094	Protein fate	Protein and peptide secretion and trafficking	hypothetical protein	-1
NP_745233.1_3063	PP_3089	Protein fate	Protein and peptide secretion and trafficking	hypothetical protein	-1
NP_745232.1_3062	PP_3088	Protein fate	Protein and peptide secretion and trafficking	hypothetical protein	-1
NP_743248.1_1078	PP_1087	Protein fate	Protein and peptide secretion and trafficking	OmpA family outer membrane protein	0
NP_743017.1_847	PP_0856	Protein fate	Protein and peptide secretion and trafficking	hypothetical protein	1
NP_742335.1_165	PP_0166	Protein fate	Protein and peptide secretion and trafficking	HlyD family type I secretion membrane fusion protein	0
NP_746133.1_3963	lolA	Protein fate	Protein and peptide secretion and trafficking	outer-membrane lipoprotein carrier protein	-1
NP_743481.1_1311	sspB	Protein fate	Degradation of proteins, peptides, and glycopeptides	ClpXP protease specificity-enhancing factor	0
NP_746692.1_4522	PP_4583	Protein fate	Degradation of proteins, peptides, and glycopeptides	oligopeptidase B	-1
NP_746682.1_4512	PP_4573	Protein fate	Degradation of proteins, peptides, and glycopeptides	ATPase AAA	-1
NP_743054.1_884	PP_0893	Protein fate	Degradation of proteins, peptides, and glycopeptides	Pfpi family intracellular protease	-1
NP_747102.1_4932	hslV	Protein fate	Degradation of proteins, peptides, and glycopeptides	ATP-dependent protease peptidase subunit	0
NP_747321.1_5151	elbB	Protein fate	Degradation of proteins, peptides, and glycopeptides	isoprenoid biosynthesis protein with amidotransferase-like domain	-1
NP_747204.1_5034	trmB	Protein synthesis	tRNA and rRNA base modification	tRNA (guanine-N(7)-)-methyltransferase	0
NP_746763.1_4593	trmA	Protein synthesis	tRNA and rRNA base modification	tRNA (uracil-5-)-methyltransferase	0
NP_742994.1_824	tgt	Protein synthesis	tRNA and rRNA base modification	queuine tRNA-ribosyltransferase	0
NP_742993.1_823	queA	Protein synthesis	tRNA and rRNA base modification	S-adenosylmethionine--tRNA ribosyltransferase-isomerase	0
NP_747215.1_5045	PP_5114	Protein synthesis	tRNA and rRNA base modification	hypothetical protein	0
NP_747001.1_4831	PP_4898	Protein synthesis	tRNA and rRNA base modification	hypothetical protein	-1
NP_746694.1_4524	PP_4585	Protein synthesis	tRNA and rRNA base modification	RNA methyltransferase	0
NP_744318.1_2148	PP_2169	Protein synthesis	tRNA and rRNA base modification	tRNA-dihydrouridine synthase A	0
NP_743798.1_1628	PP_1641	Protein synthesis	tRNA and rRNA base modification	C32 tRNA thiolase	1
NP_743486.1_1316	PP_1326	Protein synthesis	tRNA and rRNA base modification	uroporphyrin-III C/tetrapyrrole methyltransferase	-1

NP_743011.1_841	PP_0850	Protein synthesis	tRNA and rRNA base modification	radical SAM enzyme, Cfr family	0
NP_742174.1_4	gidA	Protein synthesis	tRNA and rRNA base modification	tRNA uridine 5-carboxymethylaminomethyl modification protein GidA	1
NP_743365.1_1195	proS	Protein synthesis	tRNA aminoacylation	prolyl-tRNA synthetase	1
NP_743238.1_1068	PP_1077	Protein synthesis	tRNA aminoacylation	ybaK/ebsC protein	0
NP_744618.1_2448	pheT	Protein synthesis	tRNA aminoacylation	phenylalanyl-tRNA synthetase subunit beta	1
NP_745048.1_2878	glnS	Protein synthesis	tRNA aminoacylation	glutaminyl-tRNA synthetase	1
NP_743112.1_942	rpoX	Protein synthesis	Translation factors	sigma 54 modulation protein/ribosomal protein S30EA	-1
NP_742894.1_724	prfA	Protein synthesis	Translation factors	peptide chain release factor 1	1
NP_747006.1_4836	PP_4903	Protein synthesis	Translation factors	ribosome-associated GTPase	0
NP_746914.1_4744	PP_4809	Protein synthesis	Translation factors	iojap family protein	1
NP_742729.1_559	PP_0566	Protein synthesis	Translation factors	translation initiation factor Sui1	0
NP_742556.1_386	rpsU	Protein synthesis	Ribosomal proteins: synthesis and modification	30S ribosomal protein S21	1
NP_742643.1_473	rpsK	Protein synthesis	Ribosomal proteins: synthesis and modification	30S ribosomal protein S11	1
NP_743476.1_1306	rpsI	Protein synthesis	Ribosomal proteins: synthesis and modification	30S ribosomal protein S9	1
NP_742641.1_471	rpmJ	Protein synthesis	Ribosomal proteins: synthesis and modification	50S ribosomal protein L36	-1
NP_747382.1_5212	rpmG	Protein synthesis	Ribosomal proteins: synthesis and modification	50S ribosomal protein L33	1
NP_744065.1_1895	rpmF	Protein synthesis	Ribosomal proteins: synthesis and modification	50S ribosomal protein L32	1
NP_747188.2_5018	rpmE	Protein synthesis	Ribosomal proteins: synthesis and modification	50S ribosomal protein L31	1
NP_742638.1_468	rpmD	Protein synthesis	Ribosomal proteins: synthesis and modification	50S ribosomal protein L30	1
NP_742610.1_440	rplA	Protein synthesis	Ribosomal proteins: synthesis and modification	50S ribosomal protein L1	1
NP_743982.1_1812	PP_1827	Protein synthesis	Ribosomal proteins: synthesis and modification	N5-glutamine S-adenosyl-L-methionine-dependent methyltransferase	0
NP_742413.1_243	PP_0245	Protein synthesis	Ribosomal proteins: synthesis and modification	S1 RNA-binding domain-containing protein	1
NP_742881.1_711	pth	Protein synthesis	Other	peptidyl-tRNA hydrolase	0
NP_743018.1_848	engA	Protein synthesis	Other	GTP-binding protein EngA	1
NP_742907.1_737	upp	Purines, pyrimidines, nucleosides, and nucleotides	Salvage of nucleosides and nucleotides	uracil phosphoribosyltransferase	1
NP_747099.1_4929	pyrR	Purines, pyrimidines, nucleosides, and nucleotides	Salvage of nucleosides and nucleotides	bifunctional pyrimidine regulatory protein PyrR/uracil phosphoribosyltransferase	0
NP_746382.1_4212	apt	Purines, pyrimidines, nucleosides, and nucleotides	Salvage of nucleosides and nucleotides	adenine phosphoribosyltransferase	0
NP_743767.1_1597	pyrG	Purines, pyrimidines, nucleosides, and nucleotides	Pyrimidine ribonucleotide biosynthesis	CTP synthetase	1
NP_743615.1_1445	purT	Purines, pyrimidines, nucleosides, and nucleotides	Purine ribonucleotide biosynthesis	phosphoribosylglycinamide formyltransferase 2	-1
NP_745734.1_3564	PP_3598	Purines, pyrimidines, nucleosides, and nucleotides	Purine ribonucleotide biosynthesis	peptidase C26	0
NP_743997.1_1827	PP_1842	Purines, pyrimidines, nucleosides, and nucleotides	Purine ribonucleotide biosynthesis	amidotransferase	0
NP_746397.1_4227	gad	Purines, pyrimidines, nucleosides, and nucleotides	Other	guanine deaminase	0
NP_743663.1_1493	adk	Purines, pyrimidines, nucleosides, and nucleotides	Nucleotide and nucleoside interconversions	adenylate kinase	-1

NP_743339.1_1169	nrdA	Purines, pyrimidines, nucleosides, and nucleotides	2'-Deoxyribonucleotide metabolism	ribonucleotide-diphosphate reductase subunit alpha	1
NP_747387.1_5217	dut	Purines, pyrimidines, nucleosides, and nucleotides	2'-Deoxyribonucleotide metabolism	deoxyuridine 5'-triphosphate nucleotidohydrolase	0
NP_742360.1_190	pfrA	Regulatory functions	Protein interactions	anti-RNA polymerase sigma 70 factor	-1
NP_746509.1_4339	flgM	Regulatory functions	Protein interactions	anti-sigma-28 factor FlgM	-1
NP_747145.1_4975	PP_5044	Regulatory functions	Other	GTP-binding protein TypA	1
NP_744234.1_2064	menG	Regulatory functions	Other	ribonuclease activity regulator protein RraA	-1
NP_746997.1_4827	hfq	Regulatory functions	Other	RNA-binding protein Hfq	1
NP_745520.1_3350	ptxS	Regulatory functions	DNA interactions	PtxS family transcriptional regulator	-1
NP_747425.1_5255	PP_5324	Regulatory functions	DNA interactions	response regulator receiver protein	1
NP_745909.1_3739	PP_3779	Regulatory functions	DNA interactions	LysR family transcriptional regulator	-1
NP_745553.1_3383	PP_3415	Regulatory functions	DNA interactions	LacI family transcriptional regulator	-1
NP_745443.1_3273	PP_3300	Regulatory functions	DNA interactions	TetR family transcriptional regulator	0
NP_744745.1_2575	PP_2601	Regulatory functions	DNA interactions	IclR family transcriptional regulator	-1
NP_744504.1_2334	PP_2355	Regulatory functions	DNA interactions	response regulator receiver protein	-1
NP_744119.1_1949	PP_1968	Regulatory functions	DNA interactions	TetR family transcriptional regulator	0
NP_743688.1_1518	PP_1531	Regulatory functions	DNA interactions	arsenate reductase	0
NP_743374.1_1204	PP_1214	Regulatory functions	DNA interactions	hypothetical protein	0
NP_743372.1_1202	PP_1212	Regulatory functions	DNA interactions	hypothetical protein	0
NP_747093.1_4923	pilH	Regulatory functions	DNA interactions	response regulator receiver protein	-1
NP_743346.1_1176	phoP	Regulatory functions	DNA interactions	winged helix family two component transcriptional regulator	0
NP_743224.1_1054	metR-1	Regulatory functions	DNA interactions	transcriptional activator MetR	0
NP_743173.1_1003	gltR-2	Regulatory functions	DNA interactions	winged helix family two component transcriptional regulator	-1
NP_742953.1_783	fruR	Regulatory functions	DNA interactions	DNA-binding transcriptional regulator FruR	0
NP_745891.1_3721	PP_3761	Signal transduction	Two-component systems	multi-sensor hybrid histidine kinase	-1
NP_744369.1_2199	PP_2220	Transcription	Transcription factors	C4-type zinc finger DksA/TraR family protein	0
NP_743246.1_1076	rnt	Transcription	RNA processing	ribonuclease T	0
NP_746260.1_4090	rnhA	Transcription	RNA processing	ribonuclease H	0
NP_743621.1_1451	rimM	Transcription	RNA processing	16S rRNA-processing protein RimM	0
NP_743098.1_928	cafA	Transcription	Degradation of RNA	Rne/Rng family ribonuclease	-1
NP_743545.1_1375	ttgA	Transport and binding proteins	Unknown substrate	RND family efflux transporter MFP subunit	0
NP_745593.1_3423	PP_3455	Transport and binding proteins	Unknown substrate	RND family efflux transporter MFP subunit	0
NP_745080.1_2910	PP_2936	Transport and binding proteins	Other	ABC transporter ATP-binding protein	1
NP_744050.1_1880	PP_1895	Transport and binding proteins	Other	ABC transporter ATP-binding protein	-1
NP_743935.1_1765	PP_1779	Transport and binding proteins	Other	lipopolysaccharide ABC export system, ATP-binding protein	0
NP_742688.1_518	PP_0525	Transport and binding proteins	Other	B12 family TonB-dependent receptor	0
NP_747447.1_5277	oadA	Transport and binding proteins	Cations and iron carrying compounds	pyruvate carboxylase subunit B	1
NP_743046.1_876	PP_0885	Transport and binding proteins	Cations and iron carrying compounds	peptide ABC transporter substrate-binding protein	-1
NP_746722.1_4552	fecA	Transport and binding proteins	Cations and iron carrying compounds	outer membrane iron(III) dicitrate receptor	0

NP_743243.1_1073	bfr	Transport and binding proteins	Cations and iron carrying compounds	bacterioferritin	-1
NP_744603.1_2433	rbsA	Transport and binding proteins	Carbohydrates, organic alcohols, and acids	ribose ABC transporter ATP-binding protein	0
NP_743176.1_1006	PP_1015	Transport and binding proteins	Carbohydrates, organic alcohols, and acids	sugar ABC transporter substrate-binding protein	-1
NP_743119.1_949	ttg2A	Transport and binding proteins	Anions	toluene tolerance ABC efflux transporter ATP-binding protein	0
NP_746421.1_4251	PP_4305	Transport and binding proteins	Anions	sulfate ABC transporter substrate-binding protein	-1
NP_746972.1_4802	PP_4867	Transport and binding proteins	Amino acids, peptides and amines	ABC transporter substrate-binding protein	-1
NP_746597.1_4427	PP_4486	Transport and binding proteins	Amino acids, peptides and amines	amino acid ABC transporter substrate-binding protein	0
NP_743239.1_1069	PP_1078	Transport and binding proteins	Amino acids, peptides and amines	ABC transporter ATP-binding protein	1
NP_743232.1_1062	PP_1071	Transport and binding proteins	Amino acids, peptides and amines	amino acid ABC transporter substrate-binding protein	-1
NP_743179.1_1009	PP_1018	Transport and binding proteins	Amino acids, peptides and amines	sugar ABC transporter ATP-binding protein	-1
NP_742450.1_280	PP_0283	Transport and binding proteins	Amino acids, peptides and amines	amino acid ABC transporter ATP-binding protein	0
NP_742449.1_279	PP_0282	Transport and binding proteins	Amino acids, peptides and amines	amino acid ABC transporter substrate-binding protein	-1
NP_742395.1_225	PP_0227	Transport and binding proteins	Amino acids, peptides and amines	cystine transporter subunit	-1
NP_747280.1_5110	potG	Transport and binding proteins	Amino acids, peptides and amines	spermidine/putrescine ABC transporter ATPase	0
NP_743302.1_1132	braC	Transport and binding proteins	Amino acids, peptides and amines	ABC transporter substrate-binding protein	-1
NP_743457.1_1287	aapJ	Transport and binding proteins	Amino acids, peptides and amines	amino acid ABC transporter substrate-binding protein	-1

a: Up/down regulation of protein expression, the value 0: exclusively expressed in LB; Value 1: up regulated in H_2O_2 stress; Value -1: down regulated in LB, also including proteins that were exclusively expressed in M9.

Manuscript 3

Ribosome profiling analysis and novel small ORF identification in *Pseudomonas putida* KT-2440

Xiaochen Yang, Lasse E. Pedersen, Anna Koza, Katherine S. Long

Ribosome profiling analysis and novel small ORF identification in *Pseudomonas putida* KT-2440

Xiaochen Yang, Lasse E. Pedersen, Anna Koza, Katherine S. Long*

Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Hørsholm, Denmark

Keywords: Small proteins, Pseudomonas putida, KT2440, sORFs, ribosome profiling, ribosome footprint, ORF prediction

Abstract:

Proteins of fifty amino acids or less have been overlooked due to difficulties in their identification and validation. In this study, ribosome profiling was used to identify putative small open reading frames (sORFs) in *Pseudomonas putida* strain KT2440. Three different growth conditions including two different media, LB and M9, as well as imipenem stress conditions were selected for investigated. The ribosome footprint data were obtained by RNA sequencing. The novel ORFs were predicted by analysis of the ribosome footprint sequences that were mapped to the intergenic regions of the *P.putida* KT2440 genome. A total number of 1221 putative novel ORFs were identified from three cultivation conditions, including 1163 sORFs encoding proteins with less than 51 amino acids.

1. Introduction

Due to the difficulties in their prediction and detection, small proteins with fifty amino acids or less have been understudied for decades (Storz, Wolf, & Ramamurthi, 2014). Knowledge on small proteins in bacteria has been limited in ribosomal proteins, leader peptides and toxic peptides until the small protein SpoVM was identified to be involved in the spore formation process in *Bacillus* (Levin, 1993). Additional small proteins were then discovered and found to be involved in many other biological processes, such as cell division (Handler, 2008; Modell, 2011), as well as the regulation of transport (Gaßel, 1999; Hobbs,2012; Vanderpool, 2004; Wadler, 2007) and

signal transduction (Burkholder, 2001) in bacteria. However, most well-characterized small proteins were discovered serendipitously. They were identified either by mutation or deletion of intergenic regions adjacent to functional genes (Sandman, 1987; Levin, 1993), or by re-identification of RNA sequences (Ebmeier, 2012). These strategies were limited in targeted research on small protein identification.

In order to obtain more information on small proteins in bacteria, genomic and proteomic approaches were adopted for small protein and sORF identification. However, both the genome scale Bioinformatic prediction and proteomic approaches had limitations on sORF identification. Bioinformatic prediction was one of the most popular approaches for genome-wide ORF annotation and identification. With bioinformatic approaches, hundreds of sORFs were annotated in the genome of *P.putida* KT-2440 (Belda, 2016; Nelson, 2002). However, most bioinformatic prediction tools were suitable for long ORF prediction and sORFs were easily missed especially when they were overlapped with long ORFs (Delcher, 2007; Salzberg, 1998). A proteomic approach has also been used for small protein identification (Kim, 2009). The major obstacles for proteomic small protein identification are the low concentration of small proteins in total protein samples and inappropriate protein databases for MS data analysis.

Ribosome profiling has been used for novel ORF identification in bacteria and many novel sORFs were identified in intergenic regions of the *Caulobacter crescentus* and *Mycobacterial smegmatis* genome (Schrader, 2014; Shell, 2015). This approach was established based on the binding and movement of ribosomes on mRNAs during the protein translation process. By deep sequencing the ribosome-protected RNA fragments, the regions that ribosome bind with were separated from other regions, and the coding and non-coding regions of genome can be identified (Ingolia,2014 A; Ingolia, 2014 B; Ingolia, 2009).

Pseudomonas putida is a ubiquitous Gram-negative bacterium that is tolerant to many antibiotics and environmental stress conditions. Due to its non-pathogenicity and capability to degrade organic molecules, combined with the availability of genetic tools, *P.putida* is an appropriate host for industrial applications (Marqués, 1993; Nikel, 2014). In this study, a ribosome profiling approach was used to analysis the gene expression of *P.putida* KT-2440 in three different growth conditions. By analyzing the ribosome footprints that were located in intergenic regions of

genome, more than 1200 novel ORFs with lengths longer than 48nt were obtained. More than 90% of ORFs were sORFs.

2. Material and methods

2.1 Bacteria and cultivation conditions

Pseudomonas putida KT-2440 single colonies were grown in LB medium (10 g L^{-1} of tryptone, 5 g L^{-1} of yeast extract and 5 g L^{-1} of NaCl) or M9 + glucose(0.5%(w/v)) medium with chloramphenicol (Cm) 30 ug ml^{-1} at 30 °C overnight. The overnight cultures were used to inoculate 100mL of LB or M9 medium cultures without antibiotics to an initial OD_{600} value of 0.05. The cultures without stress compound treatment were cultivated at 30°C with shaking at 250 rpm and harvested at OD value of 0.6. The stress-treatment cultures were prepared by incubation with a final concentration of 0.1ug/mL imipenem for 1 hour at an OD value of 0.6. The imipenem concentration and incubation were determined by CFU experiments described in another study (Bojanovic et al., submitted to Environmental Microbiology).

2.2 Sample preparation and RNA sequencing

The ribosome footprint samples were prepared by a gel-free protocol described previously (Latif, 2015) with modification. Briefly, by following the protocol, the cell lysate was first treated with Micrococcal Nuclease for DNA and RNA degradation. The ribosome-protected RNA fragments were then released by monosome recovery. The rRNA was removed with Ribo-Zero-rRNA Removal Kit for bacteria (Epicentre). The cDNA libraries without tRNA depletion were prepared with TruSeq RNA Library Preparation Kit (Illumina).

2.3 Material and methods for quality control, normalization, pooling and sequencing

The final concentration of each library was measured by Qubit® 2.0 Fluorimeter and Qubit DNA Broad range assay (Life Technologies). The average dsDNA library size was determined using the Agilent DNA HS kit on an Agilent 2100 Bioanalyzer (Agilent Technologies). Libraries were normalized and pooled in 10 mM Tris-Cl, pH 8.0, plus 0.05% Tween 20 to the final concentration

of 10 nM. The 9 pooled libraries were denatured in 0.2 N NaOH and neutralized in 200 mM Tris-HCl, pH 7.0. The 1.1 pM pool of 9 libraries was spiked with a 5% PhiX control (Illumina) and was loaded onto the flow cell provided in the NextSeq 500/550 High Output v2 Reagent kit (75 cycles) (Illumina). The libraries were sequenced on the NextSeq (Illumina) platform with a single read protocol and read lengths of 75nt.

2.4 Reads mapping and data analysis

The RNA sequencing data was first trimmed with trim galore to remove the reads with poor quality and barcode sequences of the Ilumina kit. The barcode sequence were from the Illumina adapters that were used for sequencing. The Trimmed reads were then mapped on the *P.putida KT-2440* genome with Bowtie (Langmead, 2012). To analyze the novel ORFs in intergenic regions, the reads that were mapped to annotated-genes on NCBI refseq annotation were removed by bedtools. The ribosome profiling reads that were located in intergenic regions were grouped by their overlapping regions. A custom-made python library was used to group the overlapping reads. Each group of overlapping reads was considered as a ribosome footprint sequence and the ORF was predicted by checking the alternative start codon (ATG, GTG and TTG) and stop codon (TAG, TAA and TGA) on the sequence. Only the longest ORFs were collected from each ribosome footprint sequence. The ORFs that encoded 15 amino acids or more were collected for further analysis.

3. Results and discussion

In this study, ribosome footprints of *P. putida* KT2440 under, imipenem stress, LB medium and M9 medium were sequenced with Nextseq techniques. For each cultivation condition, three biological replicates were prepared and sequenced. More than 300 million reads were obtained from all the samples. A bioinformatic pipeline was designed for novel ORFs prediction. With the pepline, more than 7000 grouped reads in the intergenic regions were obtained, and more than 1200 putative ORFs were predicted from grouped reads.

3.1 Quality of ribosome profiling library

In the classical ribosome profiling sample preparation protocol, the polyacrylamide gel was required for ribosome-protected fragments (RPFs) purification before library establishment (Schrader, 2014; Shell, 2015). The main purpose was to check the quality of fragments and remove incorrect fragments from the sample. However, this step was not necessary in the protocol used here(Latif, 2015). In this study, the gel-based purification step was replaced by a kit-based RNA purification step, which removed fragments larger than 200nt. Thus, impurities with less than 200 nt could easily escape the purification step, making quality checking of the sample important.

In general, the RPF is around 30nt in bacteria. In this study, the length of RPFs was analyzed with two methods. First, the length of RPFs in RNA libraries was tested by Agilent DNA HS kit on an Agilent 2100 Bioanalyzer before sequencing. Second, the length of RPFs was evaluated during the data analysis. (Figure1). The results showed that more than 80% of reads were with the length in between 27 and 44nt.

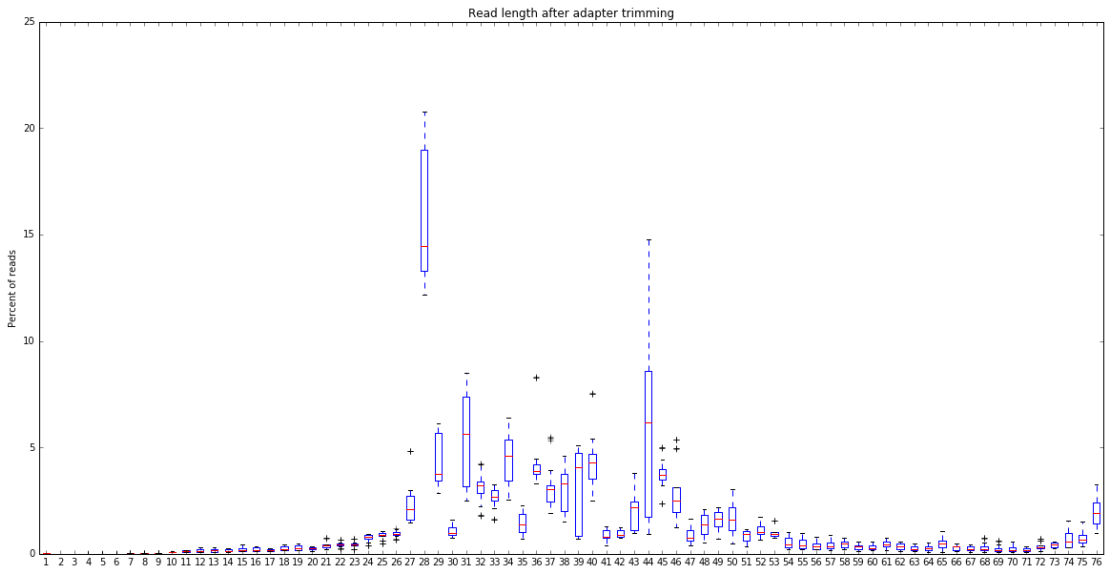


Figure 1: Length of reads in library. *The distribution of length of reads obtained from cDNA sequencing.*

3.2 Ribosome profiling reads mapping and ribosome footprint analysis

With RNA deep sequencing approach, around 380 million of reads were obtained from samples of three growth conditions, including 125.7 million reads from LB samples, 115.7 million from M9

samples and 138.3 million from imipenem samples. The reads were mapped on *P.putida* KT2440 genome, more than 50% of reads, on average, were mapped with tRNA. In comparison, fewer reads mapped to intergenic regions. In the LB sample, 0.37% of the reads mapped to intergenic regions. In the M9 and imipenem samples, 1.42% and 1.59% of the reads mapped to intergenic regions. Other reads were mapped with annotated genes.

The grouped sequences that were used for ORF prediction were obtained by grouping the ribosome profiling reads that overlapped with each other. The start and stop site of grouped sequences were determined by the start site of the first read and stop site of last read in a group. With read grouping, a total number of 7248 grouped sequences were obtained, including 1682 from LB sample, 2613 from M9 sample and 3183 from imipenem sample.

3.3 Novel ORF prediction

The ORFs were predicted by analyzing the start and stop codons on the ribosome footprint sequences. For the sequences that contained alternative ORFs, only the longest ORF was collected. In this study, the total number of 1221 novel ORFs encoding 15 amino acids or more were obtained, including 310 from LB, 647 from M9 and 833 from Imipenem samples. With the ORFs obtained from the different growth condition, around half of them were expressed exclusively in different growth condition. (Figure 2)

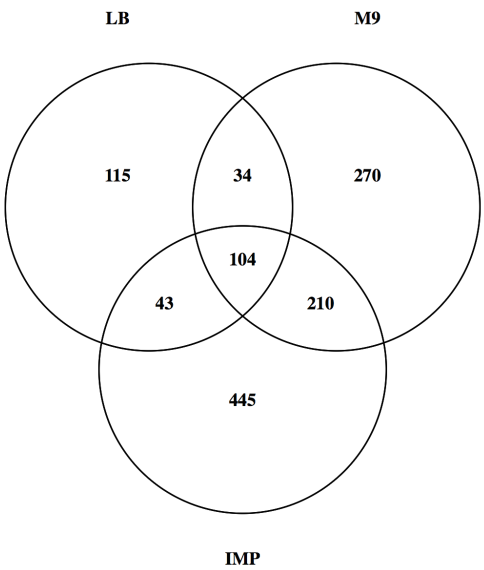


Figure 2: Novel ORFs obtained from different samples. LB: The ORFs obtained from LB medium culture. M9: The ORFs obtained from M9 medium culture. IMP: The ORFs obtained from imipenem treated sample.

The former studies indicated that the novel ORFs obtained by ribosome profiling were in general very short (Schrader; 2014, Shell, 2015; Brar, 2012; Liu, 2013). In this study, less than 10% of novel ORFs was larger than 150 bp. In addition, more than 70% of grouped reads contained putative ORFs encoding less than 15 amino acids, which was not included in our discussion. This indicated that the length of the novel ORFs obtained from ribosome profiling in this study were consistent with previous studies.

Acknowledgements

We thank Richard Szubin and Justin Tan from the Bioengineering Department, University of California San Diego, and Thomas Kallehauge from Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, for sharing their experiences with sample preparation. This work is supported by the Novo Nordisk Foundation and a PhD grant from the People Programme (Marie Curie Actions) of the European Union Seventh Framework Programme [FP7-People-2012-ITN], under grant agreement no. 317058,"BACTORY".

Reference

Belda, E., Heck, R. G. A. Van, Lopez-sanchez, M. J., & Cruveiller, S. (n.d.). The revisited genome of Pseudomonas putida KT2440 enlightens its value as a robust metabolic chassis, 1–49.

Brar, G. a., Yassour, M., Friedman, N., Regev, A., Ingolia, N. T., & Weissman, J. S. (2012). High-Resolution View of the Yeast Meiotic Program Revealed by Ribosome Profiling. *Science*, *335*(6068), 552–557.

Burkholder, W. F., Kurtser, I., & Grossman, A. D. (2001). Replication initiation proteins regulate a developmental checkpoint in Bacillus subtilis. *Cell*, *104*(2), 269–279.

Delcher, A. L., Bratke, K. A., Powers, E. C., & Salzberg, S. L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer, *23*(6), 673–679.

Ebmeier, S. E., Tan, I. S., Clapham, K. R., & Ramamurthi, K. S. (2012). Small proteins link coat and cortex assembly during sporulation in Bacillus subtilis. *Molecular Microbiology*, *84*(4), 682–

696.

Gaßel, M., Möllenkamp, T., Puppe, W., & Altendorf, K. (1999). The KdpF subunit is part of the K+-translocating Kdp complex of Escherichia coli and is responsible for stabilization of the complex in vitro. *Journal of Biological Chemistry*, *274*(53), 37901–37907.

Handler, A. A., Lim, J. E., & Losick, R. (2008). Peptide inhibitor of cytokinesis during sporulation in Bacillus subtilis. *Molecular Microbiology*, *68*(3), 588–599.

Hobbs, E. C., Yin, X., Paul, B. J., Astarita, J. L., & Storz, G. (2012). Conserved small protein associates with the multidrug efflux pump AcrB and differentially affects antibiotic resistance. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(41), 16696–701.

Ingolia, N. T. (2014). Ribosome profiling: new views of translation, from single codons to genome scale. *Nature Reviews. Genetics*, *15*(3), 205–13.

Ingolia, N. T., Brar, G. A., Stern-Ginossar, N., Harris, M. S., Talhouarne, G. J. S., Jackson, S. E., … Weissman, J. S. (2014). Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. *Cell Reports*, *8*(5), 1365–1379.

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., & Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science (New York, N.Y.)*, *324*(5924), 218–23.

Kim, W., Silby, M. W., Purvine, S. O., Nicoll, J. S., Hixson, K. K., Monroe, M., Levy, S. B. (2009). Proteomic detection of non-annotated protein-coding genes in Pseudomonas fluorescens Pf0-1. *PloS One*, *4*(12), e8455. doi:10.1371/journal.pone.0008455

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*, *9*(4), 357–359.

Latif, H., Szubin, R., Tan, J., Brunk, E., Lechner, A., Zengler, K., & Palsson, B. O. (2015). A streamlined ribosome profiling protocol for the characterization of microorganisms. *BioTechniques*, *58*(6), 329–32.

Levin, P. A., Fan, N., Ricca, E., Driks, A., Losick, R., & Cutting, S. (1993). An unusually small gene required for sporulation by Bacillus subtilis. *Molecular Microbiology*, *9*(4), 761–771.

Liu, X., Jiang, H., Gu, Z., & Roberts, J. W. (2013). High-resolution view of bacteriophage lambda gene expression by ribosome profiling. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(29), 11928–33.

Marqués, S., & Ramos, J. L. (1993). Transcriptional control of the *Pseudomonas putida* TOL

plasmid catabolic pathways. *Molecular Microbiology*, *9*(5), 923–929.

Modell, J. W., Hopkins, A. C., & Laub, M. T. (2011). A DNA damage checkpoint in Caulobacter crescentus inhibits cell division through a direct interaction with FtsW. *Genes & Development*, *25*(12), 1328–43.

Nelson, K. E., Weinel, C., Paulsen, I. T., Dodson, R. J., Hilbert, H., Santos, V. A. P. M., … Fraser, C. M. (2002). Complete genome sequence and comparative analysis of the metabolically versatile Pseudomonas putida KT2440, *4*, 799–808.

Nikel, P. I., Martínez-García, E., & de Lorenzo, V. (2014). Biotechnological domestication of pseudomonads using synthetic biology. *Nature Reviews. Microbiology*, *12*(5), 368–79.

Salzberg, S. L., Delcher, A. L., Kasif, S., & White, O. (1998). Microbial gene identification using interpolated Markov models, *26*(2), 544–548.

Schrader, J. M., Zhou, B., Li, G.-W., Lasker, K., Childers, W. S., Williams, B., … Shapiro, L. (2014). The coding and noncoding architecture of the Caulobacter crescentus genome. *PLoS Genetics*, *10*(7), e1004463. doi:10.1371/journal.pgen.1004463

Shell, S. S., Wang, J., Lapierre, P., Mir, M., Chase, M. R., Pyle, M. M., Gray, T. A. (2015). Leaderless Transcripts and Small Proteins Are Common Features of the Mycobacterial Translational Landscape, 1–31.

Storz, G., Wolf, Y. I., & Ramamurthi, K. S. (2014). Small proteins can no longer be ignored. *Annual Review of Biochemistry*, *83*, 753–77.

Vanderpool, C. K., & Gottesman, S. (2004). Involvement of a novel transcriptional activator and small RNA in post-transcriptional regulation of the glucose phosphoenolpyruvate phosphotransferase system. *Molecular Microbiology*, *54*(4), 1076–1089.

Wadler, C. S., & Vanderpool, C. K. (2007). A dual function for a bacterial small RNA: SgrS performs base pairing-dependent regulation and encodes a functional polypeptide. *Proceedings of the National Academy of Sciences*, *104*(51), 20454–20459.

Supplementary documents

S-1: Putative ORFs obtained in M9 medium
S-2: Putative ORFs obtained in LB medium
S-3: Putative ORFs obtained in Imipenem stress condition

Table description:

There are five columns for each table, the first three columns on the left side contain information of grouped reads, including start and stop site of grouped reads on genome, and the length of grouped reads. Two columns on the right side contain information of putative ORFs, including the protein sequences of putative ORFs and the length of the protein.

S-1: Putative ORFs obtained in M9 medium

Grouped reads			Putative ORFs	
Start	End	Length	Protein sequences	Length
3519505	3520246	741	['MVVIPDKFKGLLGILKSKLDMGLSESTGMQRITSEEVKWSKGPSLSILPYGLELAGVTP SPMLSAEPKNKKNCFKFYCSEQHGFYKSDIYGVNSTAIETEIYQVEGQSTYSVRFDDEGK AIRASGVIFENRVPDISCRLEDDGEYWCYEYRCEAARIVSMLAYASNSAPGTEIFIERDGE SVVGLYFYDKNSKVHLYKV']	199
4199231	4199710	479	['MNKPVYAYEEIPLIIPGLYDPNRNSYSVLIGDNGAGKSRLLSNISSMLAQKAKHRPDSAY TWRFNTNVRLDGMPKVITVSTSPFDTFKLPKSDSPYDPENSNYRYIGMRGAGSMSSGS ITLISSAAVGLLDKRLRMSAYTR']	141
2857321	2857711	390	['MSIFPNTPGFTVVIPKLHYGSYAFEQSDSVIRDLVLAAKKTALLIDRALPGVGRTGMILE GYGVDHLHAKLFPMHGTGVDSTFKKISSSVDKYFEKYEGYISSHDSVRADDKILSELAAH IKSCN']	125
6149245	6150028	783	['MDADRFGYGCGVPIQRCGHSVPAGFPDDFHAGWPHRSLLRSGCSHYLINVARPNA RAQGALTNLRRHQVAPGPCPQDGPSHRGRWSGGRYSAHPRAYGRPPARSPWREGA GRRGGPGGRKCYR']	123
4372453	4372824	371	['MDKSSMSHAPDTTAPKLSSLTIPSIVNLSSGRAGLTIDGQATDDLSGIKNIVVSFDRN LAYSYTLNDSYAEKTNIHCKWRG']	81
2064997	2065259	262	['MQCRCDRPEARRQGSQRVFRNNNKKAVRHVQTIENSPSWAHFIRHHVAIDLAEPD TFVWLTAAVAPTGNAQVTCWP']	76
1744947	1745212	265	['MDREPTNQEIAAALGIDEDQVDKYRQEAVLLGDGSWLVHFSYDMPRELRHSFTGS FTAIVACVESCVDGRAVD']	73
949859	950193	334	['MATYLQSPSISVRHCLRVAGLKEASALLFADFVKFRPTAIPGGHPSAFSRTKWTWSE VQACCLRNHKHR']	69
321873	322116	243	['MDMNAASESALRPSAVDRQSLRLLAKRLKHHGSIRVRTTDPRRLLAGRYPQGLISE AEMQALMAVWH']	67
869589	869912	323	['MWRANEQSGMCKSGRCVPNSPEQAADALIPYECLNRSITHARHLDHKSNCRTCGK RRSMHEPED']	64
2478976	2479184	208	['MGSGRSQLCRPRRMARQPGRGSDTYPGCLAMRLHRRRPALVDEQAHRSCINGSA GDQLPTLCG']	63
2672318	2672592	274	['MARAMHPSHARASPHQSPFCRRKTLKAASVRIPCDPGWAAHKAAAAHGCIECWI SVPDPL']	60
171118	171383	265	['MTGSEESVECAPRLRRKALNQTLFNKSNQAIRVGACEYGLIVKKIISITSDHARNHIVI']	59
176568	176818	250	['MTGSEESVECAPRLRRKALNQTLFNKSNQAIRVGACEYGLIVKKIISITSGHARNHIVI']	59
524698	524946	248	['MTGPEESVECAPRLRRKALNQTLFNKSNQAIRVGACEYGLIVKKIISITSGHARNHIVI']	59
697573	697815	242	['MTGSEESVECAPRLRRKALNQTLFNKSNQAIRVGACEYGLIVKKIISITSGHARNHIVI']	59
1325251	1325494	243	['MTGSEESVECAPRLRRKALNQTLFNKSNQAIRVGACEYGLIVKKIISITSDHARNHIVI']	59
2548439	2548680	241	['MTGSEESVECAPRLRRKALNQTLFNKSNQAIRVGACEYGLIVKKIISITSDHARNHIVI']	59
5168115	5168343	228	['MSIKGMTPAQPIAPPSSRPSLPKARSRAGENPLVHIILAFWALWHCRHARPPDTSLTPL']	59
5312690	5312932	242	['MTGSEESVECAPRLRRKALNQTLFNKSNQAIRVGACEYGLIVKKIISITSDHARNHIVI']	59
744501	744804	303	['MESKSKLLVFSPMKRRVEELVCLGFQRGCIWRLVWLAVRNDLFIWWIKLLGVICCCDL']	58
2674749	2674980	231	['MSTKPHEASTMSEFFRHGSVEERRSVYRMAATAAIDEQKDVIRSAKSGEFSMAKCKM']	57
5041130	5041344	214	['MVGLGYRPSQQTGRQRNRHKAVRGLARGLWLFSAGLRRADEEVMSSSQRCWYQTIT P']	57
4076355	4076559	204	['MERSGWAFGSGEATGHNRGYRYFLSTKIVRQWNNFTCPHRHPAAFMLNSVSFSNPA']	56
1745823	1746077	254	['MWPFACRDLVGAQYRRRYVWLGITATVMLRRALRMLTSGSSNPGRRRNGQVFITA']	55
5047160	5047465	305	['MTTIPSKGNDKQASDACDRKIGRKRPAFLNQDTLDRIAHEHADELKQLGSRQEKR']	55
6168913	6169098	185	['MVAEALCLQLTCLVLVLSLLPVAGAASTVVSLRVENRFKARRPTFSATIVWRLNG']	55
2621626	2621846	220	['MTETFIRGLHHGFVVPDTVRTVGKPQCVRHGLTADDDHWEPYLSTLFMGGRGAN']	54
1995230	1995570	340	['MHQAMLLQPKAPASIVMMTSMYEQETLLPLLWAAAAAVTYALLAKNVSLRLN']	52
22659	22855	196	['MRNALVSISVNTIASLLTALVLHWITRPLSDLSIQPITASNSGSKSSSCSS']	51
3595543	3595766	223	['MLAGGKGVAGDRESEGSPRQNAGLTNRKRIEAAQRGEVAQIAKAQYLHGTL']	51
4393169	4393338	169	['MLAGGKGVAGDRESEGSPRQNAGLTNRKRIEAAQRGEVAQIAKAQYLHGTL']	51
5140361	5140644	283	['MRLTRQGGIMNKPLSNWLHDLAVALGLIPPPLQPVPIPTDEEQRKRQPRRR']	51
3649347	3649811	464	['MPAFPIRWRPCAMAPSAGPWRGSNWLMVRTAALPRSPRLPAPTPPARPGR']	50
3465794	3466035	241	['MNYATYYYANTYAWRFSQSRSGQPAASDRSSTGGNAAANTNSTICRTPR']	49
2626685	2626912	227	['MTKYFGCRPPPAPRSPCWKQAWRCAVRALKVNRPALFSVVPCGVTSAA']	48
2488146	2488319	173	['MLDGSPLPGLPSGRPGNGLGSVSRAWCMQRWFQDPPDRQAVADDAGA']	47

4319837	4320137	300	['MKRTAIREGGRFTPHDLQAQPPPGWHSPGSRRNSVRSRLSDDEAGTG']	47
4551906	4552078	172	['MGAALCRDRAAKRPLDVCIAAKTAGAAVRPDRDTRPLPQGPRQLFRR']	47
5924468	5924623	155	['MLASRLKPLLHRYSTAFETGAAPVGAALAAKRPGLQTSFPAQSAIPS']	47
3023072	3023268	196	['MPTCPHPTTPFPLRASSPRQARALPLPRNPRNSRSSDRGALSRQVG']	46
6048585	6048815	230	['MSVPALSRVNPPLQATRPLWERVHPRRRHRRVCRYRPAKQRFPIGW']	46
6150354	6150558	204	['MSFPRIARSSDYPVAPSYRIPNRADYHPVAHAVASGHGRAGTIRCA']	46
2460804	2460976	172	['MWRGAYGGFTLGPPPLPWVRHDDIRRRRSPRILPHRRSDRTSNKPP']	46
4741278	4741514	236	['MHCCVVWVFCCRLWPGLKTGKWPRTRTGSRCPSVRWPGPSTRPIAV']	46
5152251	5152484	233	['MRMDTRMPSNPRLKGEEGGRSMTWISCATEHTDNIPKHQSGRLRRV']	46
25575	25745	170	['MTLADRLLYKPIDFLRSRCHCRSKNSPHASGRSQTCTAHQGVSCF']	45
589685	589849	164	['MVVSPFLVCRRKFLSVAVLSGETVESRPVASNGSGVGLPGQETAK']	45
2632932	2633083	151	['MAGGLSLTVFSAPCYKAARYCRHASQHTFHKLFITYLETNSIEK']	44
4224278	4224540	262	['MTDRRPINPNVNVPKPGGGEQPRSRNKDGQIRDKRSDAGKPRGK']	44
5974109	5974278	169	['MGHRNCDILFEIYRTVSVGCSEHAGKKLCPLSVYLRHFTTPPFG']	44
5400811	5401035	224	['MIDSLWPSVAPIVGCPSWQGGAWSNLSALNAKLGYSQNPGVVCV']	44
1348613	1348754	141	['MDEPMSTPIPLHPHLLPPLFPRPARVRLPVRSARLPRVHPWP']	42
4022680	4022890	210	['MRREVTGMDARQALGPRMGPAARSSREQARSEGTPERSAGAG']	42
4073719	4073933	214	['MVPNARERGRRPHWCAALLPNERVAQMASALIPVGTACCDHV']	42
5075337	5075475	138	['MSGVEVPALGASPMGVGLFCFSSTPLNMLRLIGRTRLPAAHS']	42
509115	509367	252	['MALARLTLSPAGPPGPSPFAFSFTARCGKAAPVAGAGRFES']	41
2406801	2406970	169	['MLPLPYSRVNPLLHGPGWTRISGSTQASTARHSIGGSAQPL']	41
5496734	5496871	137	['MKSRWSSILGDQVEIQDRVAFFTGSPAPTGSLQAAHAALYL']	41
2154013	2154143	130	['MGRPADRAPPGRVNDQGCWTAGPAYSRVNPRLQHTAGHPDR']	41
744784	744930	146	['MTMPLRIYILEARQAMLIDQMNSLQIDMAKFKEAVAANMA']	40
4963727	4963879	152	['MGWVFALGVAFGRGVRLEVFVWAALDLRGTTKLTANPSRS']	40
5966014	5966177	163	['MLGLPIACGQQGLMDLSGPARSRACWVPPPIRSADPRAVA']	40
967827	967961	134	['MAGRASSRVNPPPQVLHKAQILCRPCSGFTREEAISLTS']	39
6052440	6052569	129	['MMRGPCRIGFTREEDTAVDGTGFARVRGQARSHRDRASF']	39
1017861	1018027	166	['MIVELNPFVDEVAIEAVAQGDVCDGGAGLVALWALKGLL']	39
1388489	1388626	137	['MPPVRTVIAGCIGWPATCSRSISHAHRAASSSCDRPYQQ']	39
5743205	5743406	201	['MSRKLQGLRCTCGSERAREWTHNQTTGLAKAELPTYCAV']	39
526470	526743	273	['MSSHTNCLIHCRRRFSWSESGNEHSHRMLISGFCQIVL']	38
948986	949157	171	['MPPGVEPTSLRPVSATLDSPAQCFRDAPSACSQAMPSV']	38
1327024	1327297	273	['MSSHTNCLIHCRRRFSWSEPGNEHSHRMLISGFCQIVL']	38
3342201	3342433	232	['MFAVGIALLAVGLGTSNPGLWIPGAALMAARIVKRARS']	38
2034790	2034928	138	['MSAPALRSGCQPSPGTAQDLTAQPASEGSRQADTSRGP']	38
2436571	2436726	155	['MNNPACLAFSAQQNAKRRLGTPPLQGYAEPPLCRYRTD']	38
5560498	5560795	297	['MQRHPAWQLMHQAGCILNEKGCMMRQPGNFPAFLRLQG']	38
5596078	5596225	147	['MPGPGLSQLHTASGTPMSGTAPKEGASARFRFTPTFHC']	38
6159059	6159214	155	['MVPFIISGEMRVSMASVLGHVAAGRLAWLPTGRTVTGG']	38
194091	194243	152	['MSGRLTVTWILTPVTLTLSSKSVAPRDLRLLRCMKKC']	37
880309	880484	175	['MGEGDMRKLCRRRCKPAIRSLGPFLKGLTAITKKYME']	37
1274182	1274366	184	['MGNRYEPMQSAKIPFQTRVRTGWSEKPVKQHRCHRHA']	37
2158403	2158557	154	['MTGFVKKESYCEPKIRYKAATRFQAAGVVTSLKVRFA']	37
2412796	2412991	195	['MRREVTGMDARQALRPHGWGLQRVLPGASPERGNPGA']	37
4414891	4415030	139	['MKPPSSDPTVRLGGLHHAFTLFSRPPLHSEGSSASLR']	37
5340697	5340844	147	['MIDFPHLAHDYRRPVWHQSAAYPYLGPDGPVPEHHAR']	37
2303071	2303200	129	['MRASSRACPLPLTAYAENLAPYLWERACPRTPAQLPP']	37
5319455	5319607	152	['MRCTPAPTCYTPQRVTSRWVGNPRNTLGISQYLLSQC']	37
132436	132577	141	['MMKGRPGCPFFRPQPLHYRTRALPGLSQCRSRRGDP']	36
867902	868065	163	['MPASSRARPLPQESRRTCGHYRTCGSGHAREEGGTG']	36
1067647	1067802	155	['MCEPMNRGVVPVGAGLGAKRPVLAAENLKDISFYRA']	36
1999907	2000059	152	['MMAVLAFMLNNPTKIIIASGLGRLDVILHVSAVWRS']	36
2069156	2069348	192	['MRQAWRVARKRNPLGCGGQVGDLEVKVLYTPGKGKC']	36
2153948	2154141	193	['MWCLADQAPPARRIARGLAHRSGCPAVCCRRGFTRE']	36
3030117	3030315	198	['MDTVSRRCLTSAEWPGQALEPEHRSRHPGRASRARA']	36
1149824	1150007	183	['MGAALPAKPTPWWMARASPVFAGKPAPPGIAPAFRN']	36
1273215	1273373	158	['MGAGAPANTGEARAIHRVACFAGLPAPTVSVPASGS']	36
1747508	1747770	262	['MTFALPLVREVDGSERHDQCSPHRDQKWTRSLSAHG']	36
1967989	1968146	157	['MTGHRPFVGAGSPANAPARTTLAPDRTHSRVNPLLR']	36
3971625	3971798	173	['MSTCRSPLSRLPYSRVNPLLCMDSPTLSSCFFEHRK']	36
4190864	4190999	135	['MLDKLGSQIYFRVAHYFRDVDSPRKSGPNSRLTVSR']	36
4348883	4349068	185	['MRQAWRVARKRNPLGCGGQVGDLEVKVLYTPGKGKC']	36
3525215	3525475	260	['MFAQQLAKGLQVQGVSVGEGTVKVEQQGGQHGISR']	35
4372204	4372336	132	['MITTKVTLFGTSRCIQTNSKEPVKNGVPIYIVNGR']	35
4998866	4999022	156	['MASTFRWCMASSTATLPAPSPVSGKIGRMLCREFV']	35
1275675	1275846	171	['MPFSVRDVRWHSSRVTNALEKSVIPLMRALVAALR']	35
1784904	1785028	124	['MPAVDNSMAAFIPTDRRPALLESGLSRLFAPCTAP']	35
2099898	2100067	169	['MPAPTGLAFSLGLSCTCGNGQARGGAPSGPRLAIN']	35
2150901	2151092	191	['MRCFTCLPDSFRCTSARESLPRIARRMLRKYLPWT']	35
3764467	3764630	163	['MVKSSSTSSRTISSVQWRKESCCLSCSSRSCSALG']	35

4224526	4224705	179	['MEGTALSRLGNSQPSRISLLHRTFLGAFPGLHGQS']	35
5047024	5047220	196	['MMERLSCRRFLTPSVSEMTRWKLMDTPKGKRSAPA']	35
5573963	5574119	156	['MGAGRRRSGFTRERAGPAGQFPPAVHRCSRHRIRG']	35
5756566	5756707	141	['MCPRRGRSGFTREYGGGGNAERQVEIGQQARPLRG']	35
850901	851041	140	['MLPGGPRCRPHPRAQRLTGIHARHLPAQGLRSAS']	34
2280944	2281062	118	['MSGLREPDTCTRGSGHARETGDAVDGTGIAGVRG']	34
2984107	2984218	111	['MAQFAGWRPGCENGSPGKVALLAFWAAGRPKPVD']	34
3520907	3521073	166	['MNSGKLSNCLGVNVAAGSLCLRFDPSSSPVRPDG']	34
194317	194493	176	['MYKVNSLAVRQFQKVIDGILTSKTSRDHRHSSA']	33
3352114	3352231	117	['MSPSLRYLKRLVLFRGSDSVASSKQGVHRGWAE']	33
4656028	4656184	156	['MIRLRHLVTACATPSRRSTLISDCRAGSLYKDF']	33
2406813	2406955	142	['MNPLLQRLCRTTDAVPGRRCLCRTTDARPARSM']	33
30149	30280	131	['MTAGALPLLPPSKICRAAGSAGRPTIRQHAFV']	32
130295	130440	145	['MHRSVPGVLKFHETIANSFMGFGTKPVKRNSE']	32
131276	131474	198	['MSFPFGWRKYIMPQFLCAHSPSEGERAWGVEQ']	32
1013175	1013300	125	['MQPFAGTPAPTGITRPGGHVIPVGAGVPAKGR']	32
1373148	1373254	106	['MQVARPESGANLTGAFGASDQKVRNSSERKGA']	32
4861073	4861184	111	['MGWCVGFQVSVDCASLFAGKPAPTRASHALGR']	32
4022431	4022587	156	['MPTSKKQEARSKKQDANSSVLTQRTVRLRWRR']	32
5008027	5008201	174	['MIPSGRPSFDASGARRRSAQARGWTPVGPRRR']	32
6155262	6155395	133	['METLRKFCSTSAPDFPASRKYNSAPVSNAYSE']	32
1151147	1151269	122	['MSCLHRCSGLLTMSLGRNRLEGKCSVNRPYN']	31
1307641	1307790	149	['MRRGHANALAHLPPEHVPELYRYSVERCSSA']	31
2446494	2446611	117	['MLTSACMALPVANGACLPVSAAPRWASLCLQ']	31
2856845	2857152	307	['MKQKWVSWASCGATWCFCCCASKCWQQYHRI']	31
4169997	4170204	207	['MLTRCPVAEFAGPRERLAELGQRSNRVAVGA']	31
4252102	4252215	113	['MGCYNFAQLGVWRYTRAFVMTALWPVRSRTA']	31
6060403	6060525	122	['MDCGSGLAREHRRSRCHAPRRLFRGHARSHR']	31
1712015	1712129	114	['MRHLRLTGQAVAPGATALPVLVPGLYGRHQT']	31
5466642	5466837	195	['MPVAAMYNDRFCPQRHKAARSRARPLAIKLD']	31
5773098	5773316	218	['MRNADSAVRPLALRQALGAGGVTCDECAHLR']	31
6133852	6133976	124	['MTCAFEQVPKTCGIPKEAGPAIERPGYQRII']	31
890718	890836	118	['MKALSWNFSLPMRSKAGIEPLGICRIYEFQ']	30
1168483	1168709	226	['MCALRLVGRLVGAPFFCLGYLQWLCWPFRG']	30
1246654	1246827	173	['MFRGGHLRRSISQAQNVLRWTPGGARSHRR']	30
1828846	1828948	102	['MASRSGPFAGKPAPTTVCVQQPDQVQEQTA']	30
2050532	2050662	130	['MGRCAACAGLFAGEPAPTGMSQSPKTAGCL']	30
2528916	2529118	202	['MRREVTGMDARQALRPHGWGLQRVLPGARP']	30
752286	752454	168	['MMLDDAALKTAAECSVTASKLRFLGCFRLS']	30
1191449	1191553	104	['MLAVPASSRASPLLRGVHWFQNQRRGGFAR']	30
3811479	3811641	162	['MAATSRMLPRLPRRCSPCPAQYGRACYAPG']	30
3865648	3865834	186	['MQRVFLRSRRRGAWHWLRRCSRVNPLLQAL']	30
3948641	3948837	196	['MWGTGWVCLFSVCQLLAGTLVKSVIYSGAT']	30
4900427	4900536	109	['MHLDCDERACARARHLVGAALASQGCGRGG']	30
4909821	4910039	218	['MRREVTGMDARQALRPHGWGLQRVLPGARP']	30
5728638	5728879	241	['MSGSVWGIITPSRGPNSARQPGICRWPLVS']	30
9181	9318	137	['MDNLRWPMNYGVYRVLQLCSDPVLRLFQS']	29
1105283	1105503	220	['MQERKQARNGRGRARLCWLQMSIPSLFLT', 'MRKVAERRSCKNENRRGTAGVAPDFAGCK']	29
2233048	2233206	158	['MAMPASSRVNPLLHRIGGGHTSVVEHRPS']	29
2821374	2821550	176	['MRDLGRGGGSNRRSKFYLSLSDQSTGMAS']	29
2910816	2910929	113	['MPRRGQSFQHTSAALLSPVHPTRFCARQH']	29
5149076	5149208	132	['MANSELHNGRQYLWEMSYKKGAKSDFFSR']	29
5203738	5203902	164	['MHPPAPSARTRQSSSTTGFGGSGRNPGNG']	29
410802	410966	164	['MPTNMNVFPHKPRCDQTSHKCSFTTTRVV']	29
975152	975288	136	['MLFCAWWFCVACTDLFAGKPAPTWIALPL']	29
1747826	1748038	212	['MCDLNAQTAHNWKWSARLAYAGVAPAVTR']	29
2621642	2621785	143	['MVIISSQAMTDTLRFPYSPNSVRHHETVM']	29
2900638	2900746	108	['MVNMDHRRRHTDRCLLPWHNLWRNLRKQQ']	29
5534815	5534939	124	['MCGCSGLFAGKPAPTRCGDLLLLWELAYP']	29
5574121	5574245	124	['MGVGNCATHHLNAGAITVGAGLPAKPPAR']	29
236920	237053	133	['MFHGPSTVGQRTGSESAGPSESLAIAPV']	28
518471	518609	138	['MVRRPKGHPKSCYRRPAPLLSCAFQTAR']	28
2884425	2884529	104	['MDQFFSLSVALSRPLDCPGSACQSPLSP']	28
3497900	3498027	127	['MRRPPVRCGMGARVRMMAMRRVSSRPAG']	28
3500822	3500956	134	['MELMRQVARQFEWELSGMGSDRYLKNLL']	28
3971739	3971923	184	['MASRPGSFAGKPACMVGLEGRRRAITRF']	28
5699344	5699522	178	['MGVASSLHDFAGRQNPRSTAVPYHWSPE']	28
1501723	1501831	108	['MTTLPYKNNDNDVYPTLSVAAARQHPGH']	28
1503006	1503101	95	['MGARVGHANVRRGSSCVMVHHFRHAERG']	28
3343081	3343208	127	['MPATVGSIWTPVVCGYWNDDSDPAERTR']	28

4228256	4228404	148	['MSHNATCRHRRRQLGSTRKLVGQITVPF']	28
58416	58554	138	['MAAEEAGSSEVVTAGIASRILNPRSPD']	27
709142	709232	90	['MAGQASVFRIPRRPWALLPVTSARVWL']	27
753071	753194	123	['MLRHQCAYSPTTMSSRPGSIPVTVSAQ']	27
1298324	1298508	184	['MRGKEAVDRTGVLLCSRMREWPKWHRR']	27
1367434	1367568	134	['MAFSWRRPGDGLCKSTCCSRRLYSRLL']	27
1441831	1442022	191	['MRGKEAVDRTGVLLCSRMREWPKWHRR']	27
1819837	1820067	230	['MRNPLGCGGQVGDLEVKVLYTPGKGKC']	27
1983635	1983809	174	['MPRWARIFIAAAPDRSGCQILRYMQVS']	27
2036237	2036351	114	['MRSKGAQLSTKFYLFTLSGIVRFSFGA']	27
2137784	2137897	113	['MCSLALHHGCTHCASDFTWRVLHLFLR']	27
2472099	2472242	143	['MPASSRVNPLLRRPAPKHQCVSTPNQV']	27
2925571	2925754	183	['MRGKEAVDRTGVLLCSRMREWPKWHRR']	27
3392050	3392210	160	['MHRTTPWFMKMSSASSSPNAATLLPVP']	27
3518858	3519066	208	['MGNVFNRYIQVSMLTRWALIWIRQGNT']	27
3722474	3722574	100	['MQNEGVPAGSAGEAQGVNAMARLRRLR']	27
5000765	5000870	105	['MLTAPILAPLPRFCTKSRETESTPLGS']	27
5973696	5973784	88	['MRNSLYWAVLFWRHYRSWCWCIYRIVI']	27
85022	85143	121	['MQPIATQGRSHRGSAYDVLCGNGLASR']	27
475063	475288	225	['MRTASWATCLKRQSWSSLLYDQTCTSD']	27
2034188	2034536	348	['MDDKICELPTLLTVPPLRRNAACYSPP']	27
2710899	2711008	109	['MPASSRVNPRLQRPVSNPRPVSTLVQA']	27
2867348	2867442	94	['MSQPRSTRHRPGSCSASATGLAHWPCG']	27
3525191	3525472	281	['MTVPSPTLTPCTCRPFANCCANTTAAS']	27
3826234	3826464	230	['MRGKEAVDRTGVLLCSRMREWPKWHRR']	27
4302343	4302527	184	['MRGKEAVDRTGVLLCSRMREWSKWHRS']	27
4440343	4440457	114	['MGKLALFLGGFLLLTILIGILGTIPPS']	27
4861056	4861284	228	['MLAGSGIVVACAGLFAGKPAPTGTSPP']	27
5222593	5222793	200	['MRGKEAVDRTGVLLCSRMREWPKWHRR']	27
5545358	5545524	166	['MRGKEAVDRTGVLLCSRMREWPKWHRR']	27
6158887	6159074	187	['MRGRLFSQRIERFVGDVLKQSNVEDYS']	27
172966	173196	230	['MSDRSLKIRICDRYRLNTSFTAGGSG']	26
178405	178632	227	['MSDRSLKIRICDRYRLNTSFTAGGSG']	26
475066	475167	101	['MCARRQPLRARQSDVHVWSYSKEDQL']	26
699608	699824	216	['MSDRSLKIRICDRYRLNTSFTAGGSG']	26
1429759	1429957	198	['MPLCYRASKGDGMSVQKSAEAVVSGG']	26
2034780	2034940	160	['MLCRVKVGIQIAKLALTRGAILVEIS']	26
2550535	2550705	170	['MSDRSLKIRICDRYRLNTSFTAGGSG']	26
3827989	3828212	223	['MRITLLRGCADTHQCWLRSLSRAHEP']	26
4075433	4075616	183	['MRITLLRGCADTHQCWLRSLSRAHEP']	26
4619623	4619717	94	['MEETLYEATRIHAPSSTENQDSKRAT']	26
5280874	5280982	108	['MPGLLCSPFAGKPAPTLIGVGVRRSL']	26
5973914	5974041	127	['MAMRCLLVALWNIMLVSFLIFVCASG']	26
435298	435471	173	['MPRLLLHVVSRGADYKTSFGCLPRRL']	26
733916	734019	103	['MPCRRGFTCECDGGCTAAFAGEPAPT']	26
1058887	1058992	105	['MRRKEYRLADAYRAARDRPCANRKQT']	26
1296576	1296790	214	['MRITLLRGCADTHQCWLRSLSRAHEP']	26
1440091	1440305	214	['MRITLLRGCADTHQCWLRSLSRAHEP']	26
1811393	1811516	123	['MLLVLASSRVNPLLRRSVSNHRCNPL']	26
3754337	3754433	96	['MDDKRTYQMERVRLNSARLGLPADCF']	26
4382844	4382937	93	['MDDVLRNGKQLAADDAGDPIAVERDA']	26
4519917	4520098	181	['MRKIWLTARIFIDHGGFLLVVTPFCR']	26
5081542	5081639	97	['MEGLVATQATPVHPFMLDCFLGPCMG']	26
5310679	5310809	130	['MSDRSLKIRICDRYRLNTSFTAGGSG']	26
5453105	5453319	214	['MRITLLRGCADTHQCWLRSLSRAHEP']	26
5793339	5793442	103	['MASLELQVMAVRGLPGLSDRLAACSV']	26
5874536	5874649	113	['MALSRACPLPQVLRRPQGLRYLGRSG']	26
27155	27282	127	['MADSLYQAVIDFQLLKAGKPPSTIQ']	25
105989	106133	144	['MLAGGPRLRACQGQTCPRPGEWPCD']	25
499164	499287	123	['MPETCRSLCGSRHARERAWPADLYG']	25
1817559	1817675	116	['MRRGVFLCRWRILVASPCLSGGAKR']	25
1818649	1818754	105	['MYPACSCDVEGTGSSAAREKGAARV']	25
1983550	1983642	92	['MQLLKRRLNSRLKMRSLALLGGQAT']	25
2271618	2271710	92	['MGAWHRSEATRHMPWADPSFNRAQA']	25
2271793	2271925	132	['MKQTQEKRSRYRFFCACPEAIAPEG']	25
2373027	2373154	127	['MRLSTFCHICSAYFPGTGHKRPAKS']	25
3537384	3537558	174	['MNSPTVAIFPVARRTNIYKALQPRL']	25
5798154	5798239	85	['MWPAVRDRKACRARAQQASRKLEFA']	25
6156238	6156560	322	['MPKYRGRAGFLSQLYGGYAVATNLG']	25
1170409	1170563	154	['MKTFSTLILLRQHRSLCLVSARLPG']	25
1373970	1374062	92	['MISTYTAHPLIKKNHTLTQASPCTT']	25

1689543	1689674	131	['MPVPAFSRASPLPQVLHRLQGWRAL']	25
2643696	2643790	94	['MSPKLQATSRKQEQAEYWPALACGS']	25
4102525	4102661	136	['MPTSVTRLRSAVCGPGRLVPGSGPQ']	25
4393338	4393445	107	['MCARHGALRVSATRLVVVAKWVTWR']	25
5203780	5203867	87	['MPGPPKDQPLPGLRPEPPKPVVEDD']	25
5338014	5338108	94	['MFVVTRSAPVRPSAARAALGLTDAA']	25
5907968	5908068	100	['MRHRSCEVKLRAGFACAAALSLPRH']	25
5912223	5912446	223	['MPAQCHRVHPSFLQASSMAWSPWPQ']	25
6148413	6148509	96	['MGSTMPPRWPGLMSGSPWEPAPMLP']	25
341403	341478	75	['MQRRLGHPLARSVPAVGVVHSLQR']	24
1427681	1427899	218	['MINKKPALKNTLYRGQPQALTLFS']	24
1915605	1915807	202	['MFLTQIVCAAGSHGVKRLCKKRPI']	24
2069525	2069620	95	['MCYRASKGDGMSVQKSAEAVVSGG']	24
3784621	3784728	107	['MLETVAGAVALAMGTLSLHAHNAC']	24
4196408	4196524	116	['MTQNETAKKLGLPTTTVHLYWHRQ']	24
5261720	5261819	99	['MSHSDHHLAHPDMTRLAFCRTMPP']	24
5873956	5874048	92	['MPVFSRARPLPQESPQALRAALYL']	24
379458	379567	109	['MSPMDTVRGVFGRTLHGKRTSCFI']	24
997708	997805	97	['MKGCACAGPFAAEAAPALPVQERL']	24
1126559	1126694	135	['MGGGLCVGAVYPRMRWWLYLRIRG']	24
1999808	1999906	98	['MTLSNPRKRTRLPDAHPALPCPAN']	24
3749353	3749479	126	['MAWRKPGIRWARAGVSFAIAPARP']	24
4075707	4075852	145	['MLASSRVNPLLQRAAQFSKAVTSP']	24
4896370	4896480	110	['MLLPVPASSRVNPLLRGTAHGPRL']	24
4933047	4933134	87	['MCKVAMPRGTQNCPTCGLSPSETR']	24
5078755	5078895	140	['MWKTTMDLPSSYSLTRFIQKDLTD']	24
5122749	5122881	132	['MVKEACHERKNNVSQLGIRDREGT']	24
5377564	5377664	100	['MSSLQVLPQLAKVVAATKFKFRES']	24
5759411	5759546	135	['MLPARASSRVNPLLQGIAPFPRPA']	24
5873964	5874081	117	['MLIVPAYSRVSPLLQVQRSPQGLW']	24
6148815	6148923	108	['MSRALGSVYVARWTANGCSLATRP']	24
266211	266290	79	['MALWLSEDWLWKLRRTVCQVIQE']	23
492423	492545	122	['MPSARVWGRTPLDGLKKMLDNKA']	23
889265	889394	129	['MASQGDVGHFLLEVAGLLFGQGR']	23
1914729	1914819	90	['MPVAIGHGRLISSRLYPDCPHGA']	23
2034316	2034670	354	['MIPRRRLGIICSVLILPASLAEV']	23
2065174	2065281	107	['MASLSSRFGSIAAALHGAGLLAS']	23
2087313	2087448	135	['MGPQRLTGIHARHLPPQCLHSAS']	23
2896547	2896670	123	['MRPVRGQARSRRFFTMLNRVRTL']	23
3070626	3070771	145	['MWTKPAYTDLRIGFEVTMYFANR']	23
3803949	3804039	90	['MRRRNGHPAYGRAQQALCRFRPA']	23
5669402	5669521	119	['MKLQVAGKGIPCRCHLQLRKHRS']	23
5759395	5759531	136	['MPVRASSRVNPLLRRPRERRNPL']	23
691679	691767	88	['MGPPVDKLGALSCRLVISRPPEL']	23
1013205	1013302	97	['MGPPSGPSRARPLPQVSHDLQVW']	23
1828849	1828938	89	['MLLHLIGLLYADRRRSGFTRERA']	23
1830125	1830215	90	['MLPMAAGRQAAAPNTSARSGARV']	23
2035385	2035563	178	['METPAMMPPGPDRCAKRRLIQGE']	23
2427912	2428070	158	['MKSSTLYRRLHAHLDPCGKMPRG']	23
2655288	2655455	167	['MARASPVFAGKPAPTGIAPAFRH']	23
3342353	3342444	91	['MREPPPPLAKGRSVNEWTRAVLL']	23
3351489	3351579	90	['MKTASECSNTTRKLRFLGCFRLA']	23
3498105	3498217	112	['MAGRSGLFAGKPAPTTWRHHRSM']	23
4163470	4163640	170	['MPGRSYTPDAAARNGPAGRQGRQ']	23
4189966	4190051	85	['MHSTISRLHAARVVQRAVQGHNL']	23
4492932	4493047	115	['MEGSEQPRSVLECDMDAQRLARP']	23
4514994	4515146	152	['MLKVVVRWSRSGRSLLDAHHAQE', 'MQLYEFAPGQTPIRRAAMAAGGQ']	23
4856064	4856149	85	['MARASPVFAGKPAPTGIAPAFRN']	23
5338113	5338202	89	['MAGSKCRNFVTVRLGDCYGGKLG']	23
5545611	5545731	120	['MRSPVGDHPALPGCAVAQQTGCA']	23
5971306	5971419	113	['MRRGFSGQGFSLRSHAVRQKWIQ']	23
410685	410787	102	['MRLALWADVVNSWITGVSSGAL']	22
733188	733381	193	['MSGTKRNTSNIGDRKTTYKSVS']	22
1276476	1276556	80	['MKRALLNHEIRRQIERSKRNQK']	22
1999601	1999694	93	['MPVLTVPICSRWKRRSRCSGGR']	22
2063589	2063669	80	['MVPWQALRWWRAYLFCGLAPPQ']	22
2087973	2088083	110	['MAAMPPLGLLKSQSALPELPRT']	22
2281017	2281135	118	['MNPLLRRAVPGHRCLCRATLAL']	22
2598532	2598661	129	['MKARHYLDVCIESYEELRESTF']	22
2848705	2848780	75	['MAGGGRYRMLLDERVASKARIY']	22
3517645	3517785	140	['MAGGLHKIASIVSGINIRLCIW']	22

3521162	3521461	299	['MSVLDMIDSEIAHGIYSGLDKA', 'MNLEQRRSCLLVCLRLYPEREN']	22
5044742	5044882	140	['MEGRASVMLPELLRRAFPDTKL']	22
5955776	5955966	190	['MPAPTGPAQASRPRGTCGATAR']	22
5988712	5988869	157	['MARALRVVVPVRLGAARAALRG']	22
537476	537564	88	['MVCATHMGVLQLVIGTGLTGHQ']	22
875891	875977	86	['MTRTPYPGRGGCTCHSRVDPPP']	22
2048032	2048144	112	['MATVRLKAMAAVWQGSTLLAVV']	22
2486805	2486900	95	['MRKVPDVKCTMAAPIEHRAGGA']	22
2657689	2657759	70	['MIQALRSTCRNCLVLRMGQTQC']	22
3502571	3502681	110	['MAGHQYTRQPGYCCQLLQVTGL']	22
3826092	3826182	90	['MAYAIFLMRFNISCLRARPTAC']	22
4675993	4676103	110	['MKQTWKFSPCGKKTQAQENGHL']	22
5331608	5331749	141	['MRRMLHLRRPPALRLRAAELTR']	22
108645	108771	126	['MPLVAVYTPVERDSLMRPDCF']	21
668036	668186	150	['MGAALRRDRAAKRPRQFKRRS']	21
786579	786701	122	['MPALSRARPLPQESRILQALW']	21
915032	915177	145	['MGAALCRERAAKRPRQSKLRS']	21
958395	958506	111	['MQGSWPVVNLPATGCKAARDA']	21
1324060	1324130	70	['MGSPWPVSEQDQGHELTVAAR']	21
1455982	1456082	100	['MNPLLQRPVSNHRCVSTTNLV']	21
1627157	1627272	115	['MLGGENFCAKPESMVASLNDG']	21
1804595	1804764	169	['MPRGRFAALVIDGTGFAGVRG']	21
2429871	2429974	103	['MLPVPASSRVNPLLQRPLKPL']	21
2573157	2573237	80	['MERILMIRFRWPSMNWVPACH']	21
2643815	2643942	127	['MRGAGKPHLRAVRNTAKALVR']	21
3401705	3401803	98	['MRGWAGAVSAGCIRALERELL']	21
3647569	3647685	116	['MAWDIALAALRRSPGKGGASA']	21
4965360	4965505	145	['MLDIIPTEWSVFGRLKFENSH']	21
5214265	5214374	109	['MLDSSRATSQDHLNLKRCGVP']	21
6000896	6000996	100	['MPASSRVNPLLQVVHRLRGLW']	21
81398	81479	81	['MAEDAGRRIFPKCPAGQTFWQ']	21
668042	668187	145	['MGAALRRDRAAQQPQAFSFAA', 'MNCRGRFAALSRRKAAPTLSA']	21
1274868	1275000	132	['MPIRTGGRERILLAHHRRCGP']	21
1431441	1431529	88	['MNPLLPGMAISLRVALGLAIE']	21
1744874	1745211	337	['MPRAAAISWFVGSRSIGYSSR']	21
2085367	2085454	87	['MPVTSLRKPCARPPEVAICIA']	21
2087613	2087700	87	['MPVTSLRKPCARPPEVAICVA']	21
2256117	2256197	80	['MPVTSLRKPCARPPEVAICVA']	21
2807783	2807868	85	['MERIPPPSNTRVRCSSFVPGP']	21
2940654	2940763	109	['MAGWDHGPHFTLGCRSAGAER']	21
3329501	3329623	122	['MPCRTPSAFQVQDRHSPETST']	21
3412016	3412105	89	['MEQTLWSDYRCPSTAFTAATE']	21
3647214	3647303	89	['MPGMAWGIALASFGAYIRADV']	21
4740936	4741158	222	['MFGWQKSTASGSWCPRVTALK']	21
4747791	4748035	244	['MEQGFAGLGRFVVEFTRHDRK']	21
4890662	4890743	81	['MVAGSTALASSRASPLLHAQL']	21
5947071	5947210	139	['MGPPCGPSRASPLLHRSLRPL']	21
6150045	6150193	148	['MRKTRTRNSRTSPAASGGPCP']	21
386619	386716	97	['MPAPSRARPPPQWSCKRQSL']	20
743412	743476	64	['MCARYVKQPSVALHSGHTAS']	20
977834	977928	94	['MVACGACAHAPRLLQSRPPP']	20
1156413	1156512	99	['MSLPGDNPVVLRDIRHDPAG']	20
1275685	1275790	105	['MLGCLPSPERRNQGTHERYH']	20
1296613	1296791	178	['MRGTEIVASTGVCLHTRVGE']	20
1420332	1420520	188	['MDSLDALRRTGLPLRPSLDG']	20
1440132	1440307	175	['MRGTEIVASTGVCLHTRVGE']	20
1567282	1567440	158	['MNTVTSLLRPHPHAVSLYPG']	20
1777302	1777377	75	['MGSKRTKKAQFRCAIFALRH']	20
2236949	2237027	78	['MAGTGCAGVRGHARSHRDRI']	20
2416468	2416579	111	['MTLQVGVVTCWGGQAHWLGQ']	20
2470299	2470426	127	['MPGRCSCPTTQVECDHKEFP']	20
2598379	2598509	130	['MTLSRLEVRTISSPFSRSTS']	20
2840588	2840663	75	['MRIKVVMAVWLCAASVGVRR']	20
3398858	3398931	73	['MTVGVVDVVLPTHPAWMESQ']	20
3450598	3450679	81	['MHEQYSEKAILSNVYGSRPA']	20
4041800	4041906	106	['MPVFSLPCTTSGQKPRACVG']	20
4080857	4080936	79	['MCKVVYATLHDRDEGWLFRG']	20
4405153	4405292	139	['MTEFGGTGQKSEHLTSRRYP']	20
4832027	4832176	149	['MQIRPITRSAKNAASYGRYL']	20
5288373	5288496	123	['MGLFAGLPAPTSTAKSLKPA']	20
5390528	5390626	98	['MRSFFQGFSLGGSLLINDRL']	20

5453121	5453320	199	['MRGTEIVASTGVCLHTRVGE']	20
5477089	5477187	98	['MPVPAYSRVNPLLRTGSRLS']	20
6093379	6093458	79	['MRVSPAQKKGSIRSPEVNHL']	20
344016	344093	77	['MQSYPKRQQETRLVRVFCCL']	20
589258	589395	137	['MMLLPGAVALSRSRCCLLPV']	20
747186	747281	95	['MGNLITRTKVPEGFVTFFWV']	20
1332531	1332637	106	['MARSAGWLCVEPIVMISCSP']	20
2050544	2050681	137	['MRASSRVNPLLQAPRRLRAL']	20
2059576	2059683	107	['MFREPASSRARPLPQGYHRA']	20
2700519	2700632	113	['MPETRHWLPVRRETGNLGTG']	20
2939074	2939244	170	['MRGTEIVASTGVCLHTRVGE']	20
3477969	3478058	89	['MGRFGGLCSAARRVFRRAVH']	20
3827997	3828173	176	['MRGTEIVASTGVCLHTRVGE']	20
4075434	4075605	171	['MRGTEIVASTGVCLHTRVGE']	20
4273888	4273958	70	['MLTGPTSFFGQYCLFTVLSG']	20
4375993	4376078	85	['MCLVATQKTKPWMLVSRASS']	20
4522055	4522137	82	['MLSVLESRGAQGRRIALAAP']	20
4731450	4731535	85	['MPGLASQEARPRGSRLVACS']	20
5080147	5080300	153	['MPRPSRFCSPRRSCPALCGE']	20
5110830	5110904	74	['MQGYTVRRVLPVTLLRTAHS']	20
5190270	5190389	119	['MNAIPPGLFLPRRDRFMMAG']	20
5326431	5326512	81	['MEATLSCDRPAPGAGHLKCG']	20
5400680	5400744	64	['MPQVSVLVCAQECIVRDEQG']	20
5690343	5690422	79	['MRRAAFLQLAACNLRLQSYW']	20
5758169	5758262	93	['MRGFVPSRGAAGTGTVHSDD']	20
5828852	5828960	108	['MELVPSFQVNVGSAHKVRVL']	20
5975659	5975759	100	['MLKSDDQNSALFCVEPLRGT']	20
6150197	6150555	358	['MTTRNRMSNRMVVSAIWYAV']	20
815906	815972	66	['MLKRRRSVLFAFTFSPHLA']	19
917664	917740	76	['MDPCRSGFTREHRRSLLRG']	19
1626854	1626951	97	['MVTRGWHCETIVKRVGMPL']	19
1799832	1799941	109	['MGPLRGPSRARPLSQDSSV']	19
2034189	2034268	79	['MILARREERPSGAGGRRKH']	19
2719521	2719651	130	['MPECPGFSTGPSRGPAPSL']	19
2871120	2871195	75	['MKPTCRLSVLAQAVSGRRP']	19
3403499	3403577	78	['MRCTDPAWAPSCSKTLVDW']	19
3521045	3521155	110	['MNLGEHRPTQSCKHSHQCL']	19
3522537	3522680	143	['MSPRRPALARPRHPAAAGR']	19
3531353	3531427	74	['MPVFAGMPAPTGIAPAFRN']	19
4073872	4074007	135	['MRSTASFPRIGNHSKVFGG']	19
4357684	4357806	122	['MPASTLADKPLTVCFGTCN']	19
4379421	4379576	155	['MGVVGISETYPAMPLLRTS']	19
4970246	4970376	130	['MDTAASGPGDSRVWPREPS']	19
5172985	5173114	129	['MQTGSSGSGPRVTALSGVA']	19
5222743	5222853	110	['MRSTASFPRIGNHSKVFGG']	19
5525710	5525788	78	['MAEKNNFCHNACSPPGPVH']	19
5679809	5679878	69	['MGAGEPANKGEALARHRGA']	19
5841332	5841431	99	['MECTVERHVKIRTMVQIRV']	19
5974438	5974518	80	['MRLTKLSRIAGRRLLSNGY']	19
6166412	6166511	99	['MPFLTYPQGPKVALGPLSI']	19
161509	161614	105	['MSGTAVVVISTQHCVIRLA']	19
475506	475585	79	['MKTDVLQHTFVLPGARLPR']	19
837965	838043	78	['MIGLQCRAHCNPPTPRKSP']	19
1610388	1610469	81	['MPEQRRWRASIRDASIDID']	19
1644790	1644862	72	['MDSGAAAQPNRDTRLLPNH']	19
2821345	2821422	77	['MLPPPLPRSLIQSTPSLPR']	19
3414648	3414720	72	['MRRGTGKFLGGTRPPGGAH']	19
3899256	3899327	71	['MKLDQKSARSRLPKRLQAS']	19
4320856	4320929	73	['MTPDRRGHLVSSSDFVFSK']	19
4967677	4967757	80	['MSLDSDCVSLIHRQAIGIR']	19
5429126	5429215	89	['MNVGGRVAYVERTRIVPLN']	19
5855372	5855486	114	['MQPFAGEPAPTGGRRVFRF']	19
6149076	6149256	180	['MACSSETPVLSRTSARSTR']	19
557368	557441	73	['MTGSPEPGFLPKRFGWLS']	18
741483	741720	237	['MRVSATRLVVVAKWVTWR']	18
1076491	1076617	126	['MRCSMSNDLLSRPRGTAH']	18
1120430	1120541	111	['MVLDVLSHCLAPLHGSPP']	18
1351861	1351969	108	['MCSGLYLARPLGGARTIE']	18
1429552	1429638	86	['MRVCATRLVVVAKWVTWR']	18
1569060	1569180	120	['MPRADHFRTHPNRAVRHA']	18
2099930	2100062	132	['MAPLRGLARSHRYSLAPG']	18

2258002	2258197	195	['MTGIHARHLPPQGLRSAS']	18
2676143	2676240	97	['MGIGAALCCEGPLQITHL']	18
3030313	3030414	101	['MPCAAAGPRLAAASARLP']	18
3515962	3516076	114	['MNLLILYFAGVRSSLITC']	18
3520735	3520808	73	['MVCVLLALLPLSGPSKRV']	18
3982920	3982984	64	['MAWCAYNSSPSDRALARQ']	18
4430158	4430227	69	['MFQPWRTIRLPNSTCYRE']	18
4467621	4467730	109	['MQLGKRIATPGAATTCVG']	18
4624409	4624529	120	['MTLKCWLAKFTHGFAMTN']	18
4655850	4655953	103	['MILIIITGFAIRYPLQET']	18
5075420	5075552	132	['MKNTEKAANLVLQMVVMD']	18
5167951	5168041	90	['MRILCCLHAPCGSGLVSR']	18
5342557	5342633	76	['MGCLAAHSRASPLLHVYL']	18
5883906	5884024	118	['MFATCAGLFAGKPDPTAA']	18
82033	82141	108	['MDGTGFAGVRGHARSHRY']	18
505437	505505	68	['MASSNRSGKRTGRHTPAG']	18
1011112	1011219	107	['MGTGFISIFGPDGVRLMA']	18
1274386	1274486	100	['MQLLPYSSGWPLTLPRCN']	18
1349732	1349809	77	['MGILMYWLKTRGVRRHSF']	18
1388667	1388768	101	['MRLKRVPTTVATLAACRM']	18
1477230	1477299	69	['MTAAVPVSALLRAGSEPG']	18
1944533	1944620	87	['MRGAWAEHQPVWRPSPTA']	18
2087927	2088099	172	['MEALRPSQTDCTPDQNRR']	18
2271193	2271306	113	['MIACAGLFAGKPAPTVGH']	18
3370472	3370630	158	['MPPRCCPCSRKRCWSACC']	18
3377831	3377906	75	['MDSHLRYYTVQKLSGPVR']	18
3595750	3595857	107	['MRVSATRLVVVAKWVTWR']	18
3654687	3654752	65	['MSHGSTLGAHLKRGIDHE']	18
4272433	4272492	59	['MVLRGQIRRGPCRSFKGN']	18
4428855	4428961	106	['MRLTGCQRKNIQAFEKVV']	18
4484552	4484640	88	['MTDMPPTSSAWPLRPARV']	18
4740810	4740936	126	['MDAPFNTLKGLRERAEKN']	18
171052	171111	59	['MLYDSHPATRDRSEPSV']	17
176478	176556	78	['MLYYSPPATRDRSELSV']	17
335536	335595	59	['MSGMASSRVNPLLHGGG']	17
524548	524671	123	['MLYDSPPATRDRSEPSV']	17
589426	589491	65	['MLNHVPAAGLQRAPFSS']	17
697403	697546	143	['MLYDSPPATRDRSEPSV']	17
1163590	1163693	103	['MGVGDRPRQRYLKQQEL']	17
1330527	1330597	70	['MDLLESGDGTGFAGVRG']	17
1387816	1387895	79	['MMLPSGAARPLWALVCR']	17
1429638	1429781	143	['MSRVTANLKEARGKMLA']	17
1635654	1635817	163	['MCRDRATQQPRRFMLRS']	17
1883698	1883911	213	['MHDRNVPVTPLLRDERT']	17
2059574	2059659	85	['MFAGMPAPSGTPAHQAL']	17
2256319	2256395	76	['MVPFCHDKKGPAVRAER']	17
2380225	2380336	111	['MSLLFVVGLLNYHQMGI']	17
2474099	2474189	90	['MAMAACSRASPPLHRSV']	17
3572563	3572657	94	['MHRSAQNIEGPKPPLAV']	17
3826099	3826166	67	['MRGFVNRPLVGHEDRKC']	17
4120928	4120997	69	['MAHVPTRGCPQQRPRST']	17
4564130	4564342	212	['MPDCRPCNSDNYSSQTF']	17
4852428	4852541	113	['MCRERAAKQPRHSVLRS']	17
4886127	4886242	115	['MESLWERACPRMGQNRQ']	17
4960938	4961048	110	['MRRDRAAERPRRSLRRS']	17
5122774	5122837	63	['MPSRSLMPNWLTLFFRS']	17
5344429	5344539	110	['MPASSRVNPPTRQRVTS']	17
5425322	5425409	87	['MDNLRERRCSPLLARPA']	17
5692532	5692647	115	['MPASSRARPLPRVAHRR']	17
5914146	5914230	84	['MSPSSRARPLPQGNFEG']	17
16190	16266	76	['MIAGAIWIRSEALASNS']	17
105046	105137	91	['MTRTGHRSQSVWTCYLR']	17
278703	278814	111	['MRHVCRSGLPAKSPAPL']	17
400282	400372	90	['MVSPSLPLASSNTTVAL']	17
445037	445103	66	['MSDRTHSRASPLLQGRR']	17
507004	507104	100	['MNPLLRRPRMPEACATL']	17
1387776	1387895	119	['MLHLKAASYLWGLACRR']	17
1427786	1427862	76	['MINSSQENRVRACGWPR']	17
1818647	1818749	102	['MLSSNSDSSGTFLPGSA']	17
1859080	1859171	91	['MSNPKMLRHCRSIVGLP']	17
1883693	1883807	114	['MIAPPLAALFQSGSFVP']	17

1883799	1884035	236	['MIPSPALRCASTNHMTD']	17
2034538	2034670	132	['MTIILRLVKPAGLIPNK']	17
4292586	4292663	77	['MADNNNDGSATLFNGLF']	17
4808178	4808262	84	['MDGVGILRCLPEDFQRL']	17
5063578	5063637	59	['MTERDRDVRSDFRRLAQ']	17
5149260	5149319	59	['MHSDSRRAHSYKASALD']	17
5312957	5313080	123	['MLYDSPPATRDRSEPSV']	17
5401083	5401174	91	['MEESLMPSRSACSDWPG']	17
5571379	5571447	68	['MLPASRARPLPQGWRML']	17
5750035	5750106	71	['MAAVFAIALPTAGPAEP']	17
6168654	6168741	87	['MEGGLCYPGTKAAVVSM']	17
413986	414086	100	['MAYSRASPLPQVLHRP']	16
532226	532333	107	['MVELEAVSKQLSFRFD']	16
748808	748867	59	['MSVTTRSVQSMEMFSY']	16
1276578	1276640	62	['MPLGERIRGLQAKTSQ']	16
1395326	1395381	55	['MWSYRWHAPGALAIRH']	16
1583445	1583541	96	['MLRIVRECFPFVLRGY']	16
1831864	1831976	112	['MTELAARVGETAGRLA']	16
1999698	1999886	188	['MQLTETNLGADKSRGP']	16
2263910	2263989	79	['MEWRAGSPGGRSTIVS']	16
2667985	2668070	85	['MRPFAGLPAPTGNHRV']	16
2717610	2717768	158	['MVLIDCKGGCFAALRG']	16
3350921	3351003	82	['MHGPDEWRARLGRLAH']	16
3506533	3506682	149	['MGRHCAGRASRIFDKA']	16
3637275	3637328	53	['MAGEPMPVALCAAPCG']	16
3853363	3853424	61	['MYGACHAQPVWFFEPL']	16
4350015	4350153	138	['MPPITTVAIIEISQCR']	16
4987510	4987586	76	['MAKCRYLRVSLLPGEQ']	16
5075667	5075760	93	['MGIPPLQISRPTRWDC']	16
5125074	5125207	133	['MPAPSRVNPLLQDAPA']	16
5229160	5229230	70	['MAASSRVNPPLRGVYP']	16
5248715	5248857	142	['MKGSPPRLSSLPGSPT']	16
5404033	5404108	75	['MGSAAVAIRPVAKPLR']	16
5574135	5574212	77	['MPFTALEASRVNPLLQ']	16
5644549	5644644	95	['MLGAVIGAAARPNNRP', 'MLARAWRRYWGRCAPQ']	16
6104549	6104625	76	['MRDLRQENDVLATQGR']	16
368120	368216	96	['MKDQLQRNNMPLMPPA']	16
753534	753587	53	['MRRLVETQKIYAPARK']	16
1123917	1124020	103	['MKVLGSIDIFVFEPAQ']	16
1691748	1691815	67	['MRRSAARAALDLTGAQ']	16
2035018	2035176	158	['MGLSCVAESTNTSVDL']	16
3648888	3649080	192	['MAAACWRVSPRATCRR']	16
4022590	4022687	97	['MTRRKGGRVTMRHLRE']	16
4129228	4129293	65	['MKHVVNLWVYPKKGLH']	16
4221167	4221296	129	['MRLRLRLRLGHLSDAP']	16
4909732	4909831	99	['MTRRKGGKVTMRPLRE']	16
5095891	5095961	70	['MVQTAEALRFAPFFIL']	16
5276696	5276816	120	['MGAVGDWGHKPGRRII']	16
5366557	5366748	191	['MNIMRGTAASCAGLRG']	16
5401396	5401473	77	['MQRISVLVVNRLGLGC']	16
5496788	5496897	109	['MQHVQPAVILWERASP']	16
5809288	5809367	79	['MWLPGSLRILGGPSEQ']	16
5926156	5926261	105	['MRPCGYNARHSRGGCE']	16
4591	4698	107	['MSLCREKLNPVPGVP']	15
136926	137009	83	['MRCEEASWVDAVVST']	15
144215	144306	91	['MSQGYLARPDFRGAM']	15
150973	151059	86	['MGCLVIMCCVYAWQQ']	15
796370	796459	89	['MLELKDGTAFRMETA']	15
818682	818807	125	['MACQRLAAQEPAFAR']	15
1028935	1029046	111	['MFLGRGKRRMHPNGA']	15
1159154	1159284	130	['MTPDLWERACPRSGA']	15
1278232	1278386	154	['MCRTDTKDCAMSYGQ']	15
1502986	1503181	195	['MFGCALALGKLCTIQ']	15
1582725	1582811	86	['MPPERSPHALRVRAR']	15
1959077	1959200	123	['MPLEHIATTAQNALL']	15
1976173	1976302	129	['MGLPCSPSRACPLPQ']	15
2345508	2345582	74	['MREGATYAQEQIGET']	15
2427826	2427908	82	['MAKGYREHQVAPDCC', 'MHPIAAKGFPDQGLG']	15
2624272	2624386	114	['MGILPRNIFAFGNVP']	15
2857714	2857844	130	['MQRTKDSPIKSAQAI']	15
2954584	2954726	142	['MMLPHPAGTRNHPLG']	15

3091322	3091440	118	['MGRCMNDIEVISQVF']	15
3412015	3412159	144	['MPYSVAAVKAVLGHL']	15
3742590	3742646	56	['MAQPEVDLAQVPAVS']	15
3866926	3867083	157	['MVAATTRSPPGVHTA']	15
4192583	4192693	110	['MCPKDRLRPRRWTTA']	15
4292759	4292833	74	['MPRGTQIKDTHLYRI']	15
4501203	4501299	96	['MTEGKKLEPQKRKKP', 'MSDAETIRELRKRPG']	15
4547576	4547642	66	['MSGLTGLTALPALHP']	15
4588016	4588107	91	['MPAPTGGGAGSQAMV']	15
4624516	4624855	339	['MFQALTTQNSLTSFN']	15
5474849	5474938	89	['MADGRKWMGQCVGKM']	15
5697291	5697366	75	['MKGLFHADSPAQGPY']	15
6100727	6100777	50	['MPQPACASRDLAPDR']	15
6159380	6159458	78	['MKCVEIRHYSPAQQQ']	15
244179	244244	65	['MHAWAKLLSCSCVKV']	15
753027	753219	192	['MTGIEPGREDMVVGE']	15
869952	870132	180	['MHRLSCLLFLALHTA']	15
1039395	1039506	111	['MLVGYVERLFECSRL']	15
1122740	1122874	134	['MELPSGPFLMPARYL']	15
1168560	1168701	141	['MMRAFFVGLDWPFRG']	15
1480205	1480292	87	['MGAALCGPIATQAPL']	15
2063608	2063672	64	['MRRRQPTEQVGPPPA']	15
2367012	2367079	67	['MNVRRHAQGSPKLND']	15
2469030	2469111	81	['MAASATHRAAIKAAV']	15
2539054	2539144	90	['MSQPVCRILRQLACW']	15
2831563	2831667	104	['MRRAGAQRSRRSLPI']	15
3088553	3088614	61	['MSGSRSSEENPPLLR']	15
3275526	3275587	61	['MRLAGHLRTDRTRTL']	15
3741932	3742011	79	['MTPGPSDGYTAMSES']	15
3750435	3750517	82	['MAPRQRLNGKKSADV']	15
4073506	4073585	79	['MATSGKGKPLVKALC']	15
4156792	4156867	75	['MDRRVCFARQPSVSS']	15
4194798	4194943	145	['MARFTTADPRGRARL']	15
4361483	4361540	57	['MLSLRHCCPSLDPRV']	15
4965362	4965476	114	['MTLLVAVFEFQTSEH']	15
4986533	4986606	73	['MPPVGKPANPCSDTV']	15
5041433	5041495	62	['MKAEVMGQFCQGDYI']	15
5159757	5159832	75	['MPVPAFSRVNPLLQR']	15
5224317	5224481	164	['MSAHPRRRVIRTVSA']	15
5604440	5604513	73	['MYIERLGAYCGSVVQ']	15
5628577	5628635	58	['MAAPALRVPPISSQA']	15
5655590	5655696	106	['MSPSFAGKPAPTEVA']	15
6147419	6147556	137	['MDGLRSLRTESSRGR']	15

S-2: Putative ORFs obtained in LB medium

Grouped reads			Putative ORFs	
Start	**End**	**Length**	**Protein sequences**	**Length**
4624413	4624902	489	['MSPEKKIMATIRKGYRRPHDIEMLASEIYTWLCNDKLASREILMEFVSSVNNSKFPD VIQLTFEYLKRLSTHESELLYEESEKIGHLFDSINIMTTLGLHHDDNIIKESDELIINALKSK RFTNPPKQINTEKPWWSRISDKLLKEHIPNKSL']	153
2857289	2857716	427	['MSFLSIFPNTPGFTVVIPKLHYGSYAFEQSDSVIRDLVLAAKKTALLIDRALPGVGRTG MILEGYGVDHLHAKLFPMHGTGVDSTFKKISSSVDKYFEKYEGYISSHDSVRADDKILS ELAAHIKSCN']	128
4740842	4741158	316	['MLHITTIKEADGSLTRNLKEEPTYLPEEKGFVRVAKVDGFWKLVPKGDSTEVTYQVH TEPGGSVPAMVANKFVVDAPFNTLKGLRERAEKN']	91
1744947	1745214	267	['MDREPTNQEIAAALGIDEDQVDKYRQEAVLLGDGSWLVHFSYDMPRELRHSFTGS FTAIVACVESCVDGRAVD']	73
4199415	4199705	290	['MDGMPKVITVSTSPFDTFKLPKSDSPYDPENSNYRYIGMRGAGSMSSGSITLISSAA VGLLDKRLRMSAYTR']	72
4224258	4224683	425	['MSLSLAAGNNLAPVTKTARYVISGAMRENPAENKPRWPSSAPVHCWLPRVKAYE LKTGLGIAPICSGDDTEP']	72
4379421	4379719	298	['MDGAAMSKGLLAGLVVVLLSGCNHYFNSPYGSRWDQRNIPCDATPPNQLGCYND AHQEGLLNRLFDDD']	68
321873	322161	288	['MDMNAASESALRPSAVDRQSLRLLAKRLKHHGSIRVRTTDPRRLLAGRYPQGLISE	67

			AEMQALMAVWH']	
697575	697807	232	['MTGSEESVECAPRLRRKALNQTLFNKSNQAIRVGACEYGLIVKKIISITSGHARNHIVI']	59
1325252	1325499	247	['MTGSEESVECAPRLRRKALNQTLFNKSNQAIRVGACEYGLIVKKIISITSDHARNHIVI']	59
5168115	5168299	184	['MSIKGMTPAQPIAPPSSRPSLPKARSRAGENPLVHIILAFWALWHCRHARPPDTSLTPL']	59
5312701	5313062	361	['MTGSEESVECAPRLRRKALNQTLFNKSNQAIRVGACEYGLIVKKIISITSDHARNHIVI']	59
744517	744907	390	['MESKSKLLVFSPMKRRVEELVCLGFQRGCIWRLVWLAVRNDLFIWWIKLLGVICCCDL']	58
2674749	2674973	224	['MSTKPHEASTMSEFFRHGSVEERRSVYRMAATAAIDEQKDVIRSAKSGEFSMAKCKM']	57
5047173	5047407	234	['MTTIPSKGNDKQASDACDRKIGRKRPAFLNQDTLDRIAHEHADELKQLGSRQEKR']	55
6149843	6150193	350	['MDLAPDGHRDRVGHDCPPTPIAAWCHTELGGVDSGHSRDPMGGLAVFRAWGRIDS']	55
1425702	1425944	242	['MLAGGKGVAGDRESEGSPRQNAGLTNRKRIEAAQRGEVAQIAKAQYLHGTL']	51
1819837	1820072	235	['MLAGGKGVAGDRESEGSPRQNAGLTNRKRIEAAQRGEVAQIAKAQYLHGTL']	51
2069226	2069525	299	['MLAGGKGVAGDRESEGSPRQNAGLTNRKRIEAAQRGEVAQIAKAQYLHGTL']	51
3595543	3595772	229	['MLAGGKGVAGDRESEGSPRQNAGLTNRKRIEAAQRGEVAQIAKAQYLHGTL']	51
4393115	4393344	229	['MLAGGKGVAGDRESEGSPRQNAGLTNRKRIEAAQRGEVAQIAKAQYLHGTL']	51
5140366	5140644	278	['MRLTRQGGIMNKPLSNWLHDLAVALGLIPPPLQPVPIPTDEEQRKRQPRRR']	51
3649300	3649887	587	['MPAFPIRWRPCAMAPSAGPWRGSNWLMVRTAALPRSPRLPAPTPPARPGR']	50
3465819	3466034	215	['MNYATYYYANTYAWRFSQSRSGQPAASDRSSTGGNAAANTNSTICRTPR']	49
3030117	3030414	297	['MLRRDLPPPVLAGWRQAHEPHFIRFTSPACVPCAAAGPRLAAASARLP']	48
3023073	3023312	239	['MPTCPHPTTPFPLRASSPRQARALPLPRNPRNSRSSDRGALSRQVG']	46
2460805	2460990	185	['MWRGAYGGFTLGPPPLPWVRHDDIRRRRSPRILPHRRSDRTSNKPP']	46
4741282	4741504	222	['MHCCVVWVFCCRLWPGLKTGKWPRTRTGSRCPSVRWPGPSTRPIAV']	46
25576	25817	241	['MTLADRLLYKPIDFLRSRCHCRSKNSPHASGRSQTCTAHQGVSCF']	45
445027	445175	148	['MRHPLFQRLPCRSGLARECVRSDNNFVWSGLFAGKPAPTRVCGMA']	45
869588	869762	174	['MPHLWQTAFNARTRRLIGGVRGSCKNCEKSDMATLANPGPMLSIR']	45
5400811	5401042	231	['MIDSLWPSVAPIVGCPSWQGGAWSNLSALNAKLGYSQNPGVVCV']	44
3519581	3520109	528	['MTMKERLLGRLGSYLRIESQIYHAAWRTMVNIGAMNIGVKQHG']	43
6060398	6060538	140	['MMTLYLWERACPRNKRRGAWHRLRRCSRASPLPQSKIAPAVRN']	43
5152251	5152474	223	['MPSNPRLKGEEGGRSMTWISCATEHTDNIPKHQSGRLRRV']	40
6052440	6052574	134	['MMRGPCRIGFTREEDTAVDGTGFARVRGQARSHRDRASF']	39
948987	949131	144	['MPPGVEPTSLRPVSATLDSPAQCFRDAPSACSQAMPSV']	38
1327024	1327303	279	['MSSHTNCLIHCRRRFSWSEPGNEHSHRMLISGFCQIVL']	38
3342180	3342428	248	['MFAVGIALLAVGLGTSNPGLWIPGAALMAARIVKRARS']	38
2603491	2603620	129	['MHQPWSQRRQRRLRRPGLPGLLPGTTQRLRLRFSESS']	37
867906	868063	157	['MPASSRARPLPQESRRTCGHYRTCGSGHAREEGGTG']	36
2153952	2154088	136	['MWCLADQAPPARRIARGLAHRSGCPAVCCRRGFTRE']	36
5047068	5047220	152	['MMERLSCRRFLTPSVSEMTRWKLMDTPKGKRSAPA']	35
5078755	5078924	169	['MYDCARQPRRGLWKTTMDLPSSYSLTRFIQKDLTD']	35
5041158	5041344	186	['MDSAIGLRSRLGGRETAIRRLGVLLGVSGSSRRA']	34
1360740	1360933	193	['MPWVSAGATQVSGTVRLLWERACPRKGRYRQYT']	33
4656028	4656184	156	['MIRLRHLVTACATPSRRSTLISDCRAGSLYKDF']	33
6159324	6159445	121	['MPSRLMEECTYERAMIAALKCVEIRHYSPAQQQ']	33

30149	30363	214	['MTAGALPLLPPSKICRAAGSAGRPTIRQHAFV']	32
509143	509289	146	['MPGRMSETYPVLIVSKLLMRVIKAHARWRWRA']	32
1246600	1246771	171	['MHVMGSHQGRQGFVTTFSAPVRSSAECFEADT']	32
6149607	6149886	279	['MRPQCSRRFSRRLSCRMAASESTSKRLQSLSH']	32
324585	324706	121	['MKLPVGPPGWIRIASNLRRGVGSWRHIAARF']	31
2446501	2446611	110	['MLTSACMALPVANGACLPVSAAPRWASLCLQ']	31
6060375	6060544	169	['MDCGSGLAREHRRSRCHAPRRLFRGHARSHR']	31
6148575	6148744	169	['MDFLNAELFRDRTCRSQAVAGGHDDLHIFCL']	31
1311691	1311797	106	['MGAGTPANTGKAGGIHRSGCFAGSPAPTGRH']	31
2065047	2065175	128	['MFRQSKIRQAGLILFATTLLLILPNLTRLFG']	31
5421102	5421216	114	['MRTSRGLWCPCGGGRARERAGSVYAQPHSLL']	31
524278	524391	113	['MRSSAARVRLETFTAYETERRAGGARSHRR']	30
589725	589849	124	['MAVLSGETVESRPVASNGSGVGLPGQETAK']	30
1168489	1168702	213	['MCALRLVGRLVGAPFFCLGYLQWLCWPFRG']	30
377053	377155	102	['MSNQAIFCNRAVDKRLKALIFKNRVGSERN']	30
752286	752453	167	['MMLDDAALKTAAECSVTASKLRFLGCFRLS']	30
3865694	3865804	110	['MQRVFLRSRRRGAWHWLRRCSRVNPLLQAL']	30
1745857	1746077	220	['MMLRRALRMLTSGSSNPGRRRNGQVFITA']	29
4372516	4372722	206	['MSFDRNLAYSYTLNDSYAEKTNIHCKWRG']	29
869768	869906	138	['MNSLACAKAADVSRTALNRRQMRSSRMSV']	29
6069632	6069738	106	['MGRSSTRLLLPKPELPHPTNCNTYRKTRC']	29
3497906	3498032	126	['MRRPPVRCGMGARVRMMAMRRVSSRPAG', 'MRDGGACQDDGNATSFVAPGGVTVLFKT']	28
6023115	6023224	109	['MFPRLETCTDPVGAGTPAPTGGASAFRF']	28
58416	58522	106	['MAAEEAGSSEVVTAGIASRILNPRSPD']	27
171209	171368	159	['MGACEYGLIVKKIISITSDHARNHIVI']	27
176638	176810	172	['MGACEYGLIVKKIISITSGHARNHIVI']	27
194339	194493	154	['MAVRQFQKVIDGILTSKTSRDHRHSSA']	27
524792	524928	136	['MGACEYGLIVKKIISITSGHARNHIVI']	27
557333	557464	131	['MFRFCLVSRMTGSPEPGFLPKRFGWLS']	27
1429562	1429781	219	['MRNPLGCGGQVGDLEVKVLYTPGKGKC']	27
1627043	1627177	134	['MPRRRRAAGLSRLTPVYLKPDYRRSRP']	27
1828706	1828846	140	['MSNAPGCLHSIHRSGFTRECGSGCDGE']	27
2233054	2233189	135	['MPASSRVNPLLHRIGGGHTSVVEHRPS']	27
2548534	2548669	135	['MGACEYGLIVKKIISITSDHARNHIVI']	27
5549065	5549161	96	['MGYIDRNSGIGVLKPEMSALDSPRCSA']	27
16195	16301	106	['MLSLFLNGHFVIAGAIWIRSEALASNS']	27
475059	475183	124	['MRTASWATCLKRQSWSSLLYDQTCTSD']	27
1995379	1995553	174	['MRRYPRASSDVAAAQGPSVDCDDDVHV']	27
3525192	3525470	278	['MTVPSPTLTPCTCRPFANCCANTTAAS']	27
4440343	4440461	118	['MGKLALFLGGFLLLTILIGILGTIPPS']	27
5222617	5222784	167	['MRGKEAVDRTGVLLCSRMREWPKWHRR']	27

5966020	5966123	103	['MDLSGPARSRACWVPPPIRSADPRAVA']	27
172966	173201	235	['MSDRSLKIRICDRYRLNTSFTAGGSG']	26
178437	178617	180	['MSDRSLKIRICDRYRLNTSFTAGGSG']	26
699608	699840	232	['MSDRSLKIRICDRYRLNTSFTAGGSG']	26
2816896	2817002	106	['MSGKFFEKKLSFIRGLGYILQRSRQD']	26
4075450	4075633	183	['MRITLLRGCADTHQCWLRSLSRAHEP']	26
5973917	5974054	137	['MAMRCLLVALWNIMLVSFLIFVCASG']	26
733923	734019	96	['MPCRRGFTCECDGGCTAAFAGEPAPT']	26
1440091	1440303	212	['MRITLLRGCADTHQCWLRSLSRAHEP']	26
3498215	3498296	81	['MKPAAPVFAGRLPANAVAAVTVNDQG']	26
4411188	4411337	149	['MQATLHPWPFHTRKPSTNPPHFMPRE']	26
4515012	4515169	157	['MLKLQLYEFAPGQTPIRRAAMAAGGQ']	26
4909822	4909933	111	['MDGACSAYSREQDRSEGTPERSAGAG']	26
5310604	5310808	204	['MSDRSLKIRICDRYRLNTSFTAGGSG']	26
5874546	5874649	103	['MALSRACPLPQVLRRPQGLRYLGRSG']	26
27155	27279	124	['MADSLYQAVIDFQLLKAGKPPSTIQ']	25
1441840	1441990	150	['MLPNERVAQMASALIPVGTACCDHV']	25
2925572	2925711	139	['MLPNERVAQMASALIPVGTACCDHV']	25
1689552	1689682	130	['MPVPAFSRASPLPQVLHRLQGWRAL']	25
3826306	3826453	147	['MLPNERVAQMASALIPVGTACCDHV']	25
3969163	3969263	100	['MYQAYSRKLFASGQHRGELAVCRLH']	25
4073761	4073897	136	['MLPNERVAQMASALIPVGTACCDHV']	25
4488719	4488843	124	['MDSDGRRVGTMIADRQRRPQPATRR']	25
5110815	5110923	108	['MTPARLQGYTVRRVLPVTLLRTAHS']	25
5338014	5338105	91	['MFVVTRSAPVRPSAARAALGLTDAA']	25
5545373	5545520	147	['MLPNERVAQMASALIPVGTACCDHV']	25
5912342	5912448	106	['MPAQCHRVHPSFLQASSMAWSPWPQ']	25
1017243	1017352	109	['MRSILLWMIGVPIPVIILIWFFMH']	24
1427685	1427899	214	['MINKKPALKNTLYRGQPQALTLFS']	24
6048731	6048807	76	['MAGEFLYYRVRLPINGLHAHRNCY']	24
796859	796986	127	['MTKKATASIASHLLQQKPGGCSAA']	24
1275675	1275832	157	['MPWLRRSGEGRQPSNQLKAAPATI']	24
1963611	1963738	127	['MDGTGFARVRGQARSHRVAPAFRN']	24
4488946	4489048	102	['MDCRAMKYAHMGGRFPPVCKQGLR']	24
5122749	5122849	100	['MVKEACHERKNNVSQLGIRDREGT']	24
6148826	6148923	97	['MSRALGSVYVARWTANGCSLATRP']	24
2065190	2065275	85	['MASLSSRFGSIAAALHGAGLLAS']	23
2896544	2896649	105	['MRPVRGQARSRRFFTMLNRVRTL']	23
3070633	3070764	131	['MWTKPAYTDLRIGFEVTMYFANR']	23
4564404	4564522	118	['MGHLFIFFSPTLRFRPAICSMEE']	23
5229134	5229230	96	['MPGNIFAVAASSRVNPPLRGVYP']	23

1149885	1149970	85	['MARASPVFAGKPAPPGIAPAFRN', 'MVDGTGFARVRGQARSPRDRASF']	23
1388489	1388590	101	['MHRLACHVFALYISCAPCCQLFL']	23
1828850	1828931	81	['MLLHLIGLLYADRRRSGFTRERA']	23
2655309	2655455	146	['MARASPVFAGKPAPTGIAPAFRH']	23
4189966	4190046	80	['MHSTISRLHAARVVQRAVQGHNL']	23
5719376	5719469	93	['MRERMQAHGWHHISAHRQRSCRT']	23
410702	410784	82	['MRLALWADVVNSWITGVSSGAL']	22
733188	733374	186	['MSGTKRNTSNIGDRKTTYKSVS']	22
1276476	1276556	80	['MKRALLNHEIRRQIERSKRNQK']	22
2598563	2598661	98	['MKARHYLDVCIESYEELRESTF']	22
2857227	2857330	103	['MSELHFLPDCKRYFTLPYYMGG']	22
3521264	3521429	165	['MNLEQRRSCLLVCLRLYPEREN']	22
5248800	5248887	87	['MKLLDLTSRLAAAACSGLFFGV']	22
5955779	5955903	124	['MPAPTGPAQASRPRGTCGATAR']	22
5988712	5988807	95	['MARALRVVVPVRLGAARAALRG']	22
6150250	6150368	118	['MLGCGAAEWRWPACSWSWRGLR']	22
875891	875974	83	['MTRTPYPGRGGCTCHSRVDPPP']	22
1246695	1246793	98	['MRSSAPVRSSAARCPPQDILRL']	22
4675992	4676103	111	['MKQTWKFSPCGKKTQAQENGHL']	22
5163875	5163965	90	['MANFGAFGAALRPLADKPAARG']	22
5971761	5971914	153	['MFEGLNFWFIEGFLFIEKYRGV']	22
786579	786788	209	['MPALSRARPLPQESRILQALW']	21
796858	796942	84	['MLWECGLVSRSGCGAAPGLLL']	21
890747	890836	89	['MPMRSKAGIEPLGICRIYEFQ']	21
915034	915179	145	['MGAALCRERAAKRPRQSKLRS']	21
1627168	1627267	99	['MLGGENFCAKPESMVASLNDG']	21
2643817	2643930	113	['MRGAGKPHLRAVRNTAKALVR']	21
3401711	3401795	84	['MRGWAGAVSAGCIRALERELL']	21
3518860	3518967	107	['MERAFVMGALGCWCWGGERFQ']	21
5000560	5000672	112	['MGMTRMASQAGIEPLSGVPPA']	21
6035361	6035470	109	['MANSLAKPPSKRSVCACPESL']	21
668064	668187	123	['MGAALRRDRAAQQPQAFSFAA']	21
1274877	1274970	93	['MPIRTGGRERILLAHHRRCGP']	21
4223967	4224077	110	['MTPYVPLLQRFQRTRAGRAEP']	21
4630275	4630368	93	['MIYMSGHFFDLSSRLKCLFCS']	21
5873961	5874068	107	['MPAYSRVSPLLQVQRSPQGLW']	21
5947085	5947210	125	['MGPPCGPSRASPLLHRSLRPL']	21
6148645	6148791	146	['MVCSLACWRCPIRSNPPPRRR']	21
1420337	1420471	134	['MDSLDALRRTGLPLRPSLDG']	20
1567346	1567438	92	['MNTVTSLLRPHPHAVSLYPG']	20
2548361	2548533	172	['MLSQVFEVKRVSRKNFKINA']	20

2840588	2840663	75	['MRIKVVMAVWLCAASVGVRR']	20
5456190	5456268	78	['MPDAARFRCLEQGLFSLSQK']	20
5845576	5845674	98	['MHCGRECGTIVPIALNMPGL']	20
5914126	5914216	90	['MRPVSPSSRARPLPQGNFEG']	20
288195	288283	88	['MPNRSIGSCKHLWLRHGSWL']	20
465174	465252	78	['MNISLIITTKLTTFRQHSAP']	20
1747519	1747605	86	['MLEDLPPKKSQEASASARAA']	20
2050566	2050678	112	['MRASSRVNPLLQAPRRLRAL']	20
4197697	4197772	75	['MLAMEHKRSVALLTQFSKML']	20
4292640	4292718	78	['MPCQTRQSWLDSHIRTGSGR']	20
4321010	4321088	78	['MDSISKLRILCCWHQGLRVI']	20
4551909	4551986	77	['MRPDRDTRPLPQGPRQLFRR']	20
4731445	4731534	89	['MPGLASQEARPRGSRLVACS']	20
5326431	5326533	102	['MEATLSCDRPAPGAGHLKCG']	20
5758169	5758260	91	['MRGFVPSRGAAGTGTVHSDD']	20
5828851	5828938	87	['MELVPSFQVNVGSAHKVRVL']	20
6150200	6150553	353	['MTTRNRMSNRMVVSAIWYAV']	20
1298378	1298491	113	['MAQMASALIPVGTACCDHV']	19
1626874	1626942	68	['MVTRGWHCETIVKRVGMPL']	19
1817596	1817671	75	['MCRWRILVASPCLSGGAKR']	19
1993562	1993633	71	['MAIHEPCSAAQPHCRAFAI']	19
2087362	2087448	86	['MPPLGLLKSQSASPGLPRA']	19
4970247	4970371	124	['MDTAASGPGDSRVWPREPS']	19
5525710	5525785	75	['MAEKNNFCHNACSPPGPVH']	19
5545472	5545582	110	['MRSTASFPRIGNHSKVFGG']	19
6166434	6166496	62	['MPFLTYPQGPKVALGPLSI']	19
475506	475584	78	['MKTDVLQHTFVLPGARLPR']	19
4491886	4491972	86	['MRTENFVAVGLLSNAPRQS']	19
4967677	4967759	82	['MSLDSDCVSLIHRQAIGIR']	19
5084549	5084646	97	['MGDPAPTRYVQGLNVVPCP']	19
6148948	6149640	692	['MACSSETPVLSRTSARSTR']	19
466642	466725	83	['MDWNVQWCGLLRLAQAPR']	18
697363	697553	190	['MKFNELDNSEGAVRATIK']	18
2099937	2100066	129	['MAPLRGLARSHRYSLAPG']	18
2137793	2137874	81	['MTVLVQPGLAPWLHSLRQ']	18
3515969	3516070	101	['MNLLILYFAGVRSSLITC']	18
3982867	3982984	117	['MAWCAYNSSPSDRALARQ']	18
4196872	4196948	76	['MISAWNENSAATTASSIG']	18
4629121	4629247	126	['MYRNNDTGIIIAGFWQFS']	18
4655819	4656003	184	['MILIIITGFAIRYPLQET']	18
4716752	4716871	119	['MHTCTSFARCAVTFTTTQ']	18

5075462	5075548	86	['MKNTEKAANLVLQMVVMD']	18
5342556	5342637	81	['MGCLAAHSRASPLLHVYL']	18
5477094	5477182	88	['MPAYSRVNPLLRTGSRLS']	18
5883907	5883976	69	['MFATCAGLFAGKPDPTAA']	18
6067185	6067254	69	['MRLRRSPSRACRLPLGPQ']	18
967826	967965	139	['MRLMASSRVNPLQGLHRI']	18
1274386	1274528	142	['MQLLPYSSGWPLTLPRCN']	18
1457541	1457629	88	['MAWRNLEWMYLFQHSQHC']	18
1555934	1556078	144	['MIVTAMAYASPKRHTGAA']	18
2074129	2074187	58	['MTTGFFAGTPAPTDVSRP']	18
3754336	3754403	67	['MERVRLNSARLGLPADCF']	18
4428855	4428955	100	['MRLTGCQRKNIQAFEKVV']	18
4484573	4484686	113	['MTDMPPTSSAWPLRPARV']	18
4936756	4936836	80	['MAPPSLWLSKSLSLWPLP']	18
5182333	5182427	94	['MGGIRLPFYRSASGREIR']	18
5728788	5728943	155	['MRCHATCWIIARCMDLRS']	18
170996	171127	131	['MLYDSHPATRDRSEPSV']	17
524611	524716	105	['MLYDSPPATRDRSEPSV']	17
668059	668165	106	['MRRDRAAKRPRQFKRRS']	17
741580	741726	146	['MSRVTANLKEARGKMLA']	17
1330528	1330597	69	['MDLLESGDGTGFAGVRG']	17
1582669	1582727	58	['MVVCKYCWVVLRYSLSA', 'MFVNTAGWYSGTASALD']	17
2158408	2158566	158	['MLEKRTSFYKISHRPAA']	17
3376502	3376570	68	['MVSARLNRPGFRGGSTL']	17
3518935	3519005	70	['MSMLTRWALIWIRQGNT']	17
4120928	4120997	69	['MAHVPTRGCPQQRPRST']	17
4564135	4564207	72	['MPDCRPCNSDNYSSQTF']	17
4748152	4748269	117	['MDPPSGVDIMRGQDKTV']	17
4852428	4852538	110	['MCRERAAKQPRHSVLRS']	17
5425322	5425406	84	['MDNLRERRCSPLLARPA']	17
5692538	5692639	101	['MPASSRARPLPRVAHRR']	17
105075	105148	73	['MTRTGHRSQSVWTCYLR']	17
278708	278802	94	['MRHVCRSGLPAKSPAPL']	17
400293	400371	78	['MVSPSLPLASSNTTVAL']	17
1818684	1818744	60	['MLSSNSDSSGTFLPGSA']	17
1883800	1883958	158	['MIPSPALRCASTNHMTD']	17
2035461	2035547	86	['MQRSPGGALSILWKLQL']	17
2672346	2672572	226	['MRPWLGGPQSRGSAWVH']	17
2817095	2817184	89	['MAQKVAQNSSETENHQA']	17
3275532	3275614	82	['MKVRLAGHLRTDRTRTL']	17
4207874	4207936	62	['MAGRARKKSFAISTTKG']	17

4379625	4379725	100	['MNERSPTKPAKQAGLIA']	17
4957517	4957577	60	['MNPLLQGTVSAEHGLRT']	17
5750035	5750106	71	['MAAVFAIALPTAGPAEP']	17
5908053	5908117	64	['MRQLWAGFFGPEPLKGN']	17
413987	414062	75	['MAYSRASPLPQVLHRP']	16
414065	414182	117	['MPANRPCEWYTVRHIA']	16
1276573	1276644	71	['MPLGERIRGLQAKTSQ']	16
1395326	1395391	65	['MWSYRWHAPGALAIRH']	16
1666922	1666998	76	['MCVSPLAPDKPRLYWA']	16
1831845	1831993	148	['MTELAARVGETAGRLA']	16
1999700	1999887	187	['MQLTETNLGADKSRGP']	16
2488148	2488311	163	['MAAHCRGCRLAGPAMD']	16
2573170	2573240	70	['MIRFRWPSMNWVPACH']	16
3519509	3519578	69	['MVAQLSWLFRMVCLMC']	16
3561622	3561704	82	['MERPCVSMGRSFTRLR']	16
4467647	4467701	54	['MGKRIATPGAATTCVG']	16
5075655	5075757	102	['MGIPPLQISRPTRWDC']	16
5248708	5248812	104	['MKGSPPRLSSLPGSPT']	16
58238	58323	85	['MHDDGRPHGGETGDSE', 'MTMAGLTEVKPEIQND']	16
753532	753598	66	['MRRLVETQKIYAPARK']	16
1627193	1627266	73	['MNSKERCISHRSGLQP']	16
1690928	1691013	85	['MQGFASGCCVSGQKAN']	16
1691748	1691813	65	['MRRSAARAALDLTGAQ']	16
1944533	1944595	62	['MGRAPTCLEAIADCMK']	16
2633357	2633416	59	['MRPAITVLGPGQAGSG']	16
3648890	3649081	191	['MAAACWRVSPRATCRR']	16
4231743	4231810	67	['MSISVSLAEASSDYME']	16
4295385	4295439	54	['MFGPAAVPWPTRATGC']	16
4302344	4302459	115	['MASELIPVGTACCDHV']	16
5072928	5072990	62	['MRRPGFLPLAPLLSFV']	16
5366556	5366666	110	['MNIMRGTAASCAGLRG']	16
5401393	5401462	69	['MQRISVLVVNRLGLGC']	16
5955781	5955973	192	['MPHRLSRSCPTGTAWP']	16
130306	130414	108	['MKLSQIVLWASELNR']	15
144220	144308	88	['MSQGYLARPDFRGAM']	15
1503013	1503163	150	['MFGCALALGKLCTIQ']	15
1582731	1582811	80	['MPPERSPHALRVRAR']	15
1883668	1883727	59	['MLEVPALNTGVMIRA']	15
2624235	2624386	151	['MGILPRNIFAFGNVP']	15
2857736	2857830	94	['MQRTKDSPIKSAQAI']	15
3030417	3030477	60	['MLARLHARLSPRTSH']	15

3866926	3867036	110	['MVAATTRSPPGVHTA']	15
4471488	4471550	62	['MASPRTGGLAACREG']	15
4501202	4501261	59	['MSDAETIRELRKRPG']	15
4547575	4547643	68	['MSGLTGLTALPALHP']	15
4588024	4588108	84	['MPAPTGGGAGSQAMV']	15
5697290	5697366	76	['MKGLFHADSPAQGPY']	15
780524	780591	67	['MPKVKEPGPVPGSFV']	15
869953	870041	88	['MHRLSCLLFLALHTA']	15
1274304	1274374	70	['MLGVAMATVLLNWFF']	15
4156792	4156873	81	['MDRRVCFARQPSVSS']	15
4747850	4747926	76	['MRICRKSPNKTVWSL']	15
5041430	5041497	67	['MKAEVMGQFCQGDYI']	15
5560550	5560682	132	['MAQRRLTVWSARKAI']	15
5604447	5604513	66	['MYIERLGAYCGSVVQ']	15
5655546	5655613	67	['MPANATPQIPQADDN']	15
6148445	6148644	199	['MRSRNNSALRKSTEK']	15

S-3: Putative ORFs obtained in Imipenem stress condition

Grouped reads			Putative ORFs	
Start	**End**	**Length**	**Protein sequences**	**Length**
3519431	3520260	829	['MVVIPDKFKGLLGILKSKLDMGLSESTGMQRITSEEVKWSKGPSLSILPYGLELAGV TPSPMLSAEPKNKKNCFKFYCSEQHGFYKSDIYGVNSTAIETEIYQVEGQSTYSVRFD DEGKAIRASGVIFENRVPDISCRLEDDGEYWCYEYRCEAARIVSMLAYASNSAPGTEI FIERDGESVVGLYFYDKNSKVHLYKV']	199
4199182	4199712	530	['MFESKSVEQTKPHCKEQKLNKPVYAYEEIPLIIPGLYDPNRNSYSVLIGDNGAGKSR LLSNISSMLAQKAKHRPDSAYTWRFNTNVRLDGMPKVITVSTSPFDTFKLPKSDSPY DPENSNYRYIGMRGAGSMSSGSITLISSAAVGLLDKRLRMSAYTR']	159
4624409	4624880	471	['MSPEKKIMATIRKGYRRPHDIEMLASEIYTWLCNDKLASREILMEFVSSVNNSKFP DVIQLTFEYLKRLSTHESELLYEESEKIGHLFDSINIMTTLGLHHDDNIIKESDELIINALK SKRFTNPPKQINTEKPWWSRISDKLLKEHIPNKSL']	153
744445	744947	502	['MRKILINENGFIENPFASNMPGFMNKRSWGGIVYRRGKQIQIASFLPDEEKGRRIS VLGVSKGMYLAFGLVGCKERLVYLVDQVTRGDLLLRPVEVSSVPSIYDIELTMPLRIYI LEARQAMLIDQMNSLQIDMAKFKEAVAANMA']	146
6148946	6150561	1615	['MDADRFGYGCGVPIQRCGHSVPAGFPDDFHAGWPHRSLLRSGCSHYLINVARPN ARAQGALTNLRRHQVAPGPCPQDGPSHRGRWSGGRYSAHPRAYGRPPARSPWR EGAGRRGGPGGRKCYR']	123
3971664	3971975	311	['MKGTPWRRLASIVACSRGIPYRRLGVERRRGDFSYGLYCANSQKRVMALLLPSSPT IQAGLPAKEPGLLANVHLPFTIVAATVFAGKPAPMYGLAHTV']	98
5203740	5204086	346	['MLMFHAILAALLMAGLPLAEAASTPPRLNTPVPGAPGTPTPTPYPQITPSTPPKAY DSQPGAPLLPPMPVPGPPKDQPLPGLRPEPPKPVVEDD']	94
3649300	3650211	911	['MTVRGSPQSPRSVYWPWATSMLPCLMAALPAGALRAASCFVMSTCQAKLSAN WWRACATRLRWLPNRSRRCLMDRPMWWSSMPAVSTNTRP']	91
4740810	4741162	352	['MLHITTIKEADGSLTRNLKEEPTYLPEEKGFVRVAKVDGFWKLVPKGDSTEVTYQ VHTEPGGSVPAMVANKFVVDAPFNTLKGLRERAEKN']	91
3525195	3525546	351	['MRSANAWHCSTLAPLNSSANSSPLARPLTWASNCAWLSSRPSCSRSSWRKACR CRVSVSVRVPSRSNSRAVSMGYPDEKTQARGS']	85
4372454	4372783	329	['MDKSSMSHAPDTTAPKLSSLTIPSIVNLSSGRAGLTIDGQATDDLSGIKNIVVSFD RNLAYSYTLNDSYAEKTNIHCKWRG']	81
2064997	2065278	281	['MQCRCDRPEARRQGSQRVFRNNNKKAVRHVQTIENSPSWAHFIRHHVAIDLA EPDTFVWLTAAVAPTGNAQVTCWP']	76
2528919	2529191	272	['MRDSSGDADCDFRRPSVGLAEGGDGHGCPSSAEAPWMGPAARTPGSKTVARE PRSVAQGPDDGSQRFLVTFCRV']	74
5973806	5974105	299	['MIGNKNLYSRAGLIGRIMRICTVSILLTMPKVFVYRGLVDLSEVDRFPVAMRCLLV	74

			ALWNIMLVSFLIFVCASG']	
1744945	1745214	269	['MDREPTNQEIAAALGIDEDQVDKYRQEAVLLGDGSWLVHFSYDMPRELRHSFT GSFTAIVACVESCVDGRAVD']	73
6158975	6159258	283	['MQWSRSSSLAKCALAWPLFLAMWRPDDWRGCQQAEQSQVVECTHDNLVLTV IKFGDTGSTTVPIRIRGKKKPP']	73
4224276	4224680	404	['MSLSLAAGNNLAPVTKTARYVISGAMRENPAENKPRWPSSAPVHCWLPRVKAYE LKTGLGIAPICSGDDTEP']	72
4593694	4593991	297	['MAATVNGGWKLASRAGPFAGKPAPTRGLRLITNMCPPPIGCRSGFTREYGSGGN GERRVEIGQQARPFRG']	70
321881	322135	254	['MDMNAASESALRPSAVDRQSLRLLAKRLKHHGSIRVRTTDPRRLLAGRYPQGLIS EAEMQALMAVWH']	67
869589	869906	317	['MWRANEQSGMCKSGRCVPNSPEQAADALIPYECLNRSITHARHLDHKSNCRTC GKRRSMHEPED']	64
1994998	1995320	322	['MHVGACATDVGDGRGAPYPGCSDFGKLNNVYGRKPNSLSAKPVAPSSASKPAS HARALGNRSQS']	64
5544896	5545108	212	['MMRAGSPAKGPAACKYVSGCSNIHQSPCLCAPASLQLRPLPSWQALMCLPSVQ DIQCHSLLV']	62
2672318	2672622	304	['MARAMHPSHARASPHQSPFCRRKTLKAASVRIPCDPGWAAHKAAAAHGCIEC WISVPDPL']	60
171118	171379	261	['MTGSEESVECAPRLRRKALNQTLFNKSNQAIRVGACEYGLIVKKIISITSDHARNHIVI']	59
176567	176815	248	['MTGSEESVECAPRLRRKALNQTLFNKSNQAIRVGACEYGLIVKKIISITSGHARNHIVI']	59
524682	524942	260	['MTGPEESVECAPRLRRKALNQTLFNKSNQAIRVGACEYGLIVKKIISITSGHARNHIVI']	59
697575	697821	246	['MTGSEESVECAPRLRRKALNQTLFNKSNQAIRVGACEYGLIVKKIISITSGHARNHIVI']	59
1325160	1325502	342	['MTGSEESVECAPRLRRKALNQTLFNKSNQAIRVGACEYGLIVKKIISITSDHARNHIVI']	59
2548442	2548686	244	['MTGSEESVECAPRLRRKALNQTLFNKSNQAIRVGACEYGLIVKKIISITSDHARNHIVI']	59
5168108	5168382	274	['MSIKGMTPAQPIAPPSSRPSLPKARSRAGENPLVHIILAFWALWHCRHARPPDTSLT PL']	59
2674749	2675023	274	['MSTKPHEASTMSEFFRHGSVEERRSVYRMAATAAIDEQKDVIRSAKSGEFSMAKCK M']	57
610857	611083	226	['MLLLSLAREAFLIGAWVLKGGGVISPTVCTARTVRSELWPSSSLQVLPYKRVYPRIR']	57
1156298	1156512	214	['MPVRASSRVNPLLQRPSQVCVVGAGSPAKRPAQAPKMSLPGDNPVVLRDIRHDPA G']	56
1503008	1503197	189	['MAEVMHHHAGRTPSNIGVPDPSPQKRLCSAARLPLASCALSSKIRRLFFAGRPLRL']	56
2867065	2867288	223	['MAGCPSSAPPPSRGCSRSAAQWAQSWSAGSWTDATLTGSSPSPTPSAACASSHW AP']	56
27261	27559	298	['MYTDAVDAMKQYHEAQASGSSAEEVERLRQIAESQFRAVSEYQLNALGYQSRRPH' , 'MSACVKSQNHNFGLSASISSTLWDISLVVRTEGITDPEVASESIKPSGRLTTRSD']	55
5047165	5047460	295	['MTTIPSKGNDKQASDACDRKIGRKRPAFLNQDTLDRIAHEHADELKQLGSRQEKR']	55
4488962	4489215	253	['MPQLIQSTISRGMRTGRSPTGGVVSGGETRYLGIDVAGVRHHLWRVLGPVQPSEM ']	55
949894	950194	300	['MRVAGLKEASALLFADFVKFRPTAIPGGHPSAFSRTKWTWSEVQACCLRNHKHR']	54
1425598	1425944	346	['MRRDPRPPVMHMAGGFCVRQAWRVARMRNPLGCGGQVGDLEVKVLYTPGKGK C']	53
2857019	2857197	178	['MTVASIGPHPGGLVPSFGPLNFNITAVVIRRHRLYSVSRLRQSSLKLYRMGCR']	53
2857543	2857711	168	['MHGTGVDSTFKKISSSVDKYFEKYEGYISSHDSVRADDKILSELAAHIKSCN']	52
1995327	1995600	273	['MHQAMLLQPKAPASIVMMTSMYEQETLLPLLWAAAAAVTYALLAKNVSLRLN']	52
22669	22867	198	['MRNALVSISVNTIASLLTALVLHWITRPLSDLSIQPITASNSGSKSSSCSS']	51
741574	741738	164	['MLAGGKGVAGDRESEGSPRQNAGLTNRKRIEAAQRGEVAQIAKAQYLHGTL']	51
2069201	2069467	266	['MLAGGKGVAGDRESEGSPRQNAGLTNRKRIEAAQRGEVAQIAKAQYLHGTL']	51
4564384	4564587	203	['MRLFRGITMVTSGSPFYLFLSYTQVSARHLLNGRVTKKSRTLRKAGRSGTK']	51
3595543	3595766	223	['MLAGGKGVAGDRESEGSPRQNAGLTNRKRIEAAQRGEVAQIAKAQYLHGTL']	51
5140356	5140656	300	['MRLTRQGGIMNKPLSNWLHDLAVALGLIPPPLQPVPIPTDEEQRKRQPRRR']	51
4075665	4075852	187	['MRGAGRASVACAGLFAGKPAPTTCGTVFKGCDIPVGDHPALPGCAVAQQT']	50
3465794	3466023	229	['MNYATYYYANTYAWRFSQSRSGQPAASDRSSTGGNAAANTNSTICRTPR']	49

3865672	3865856	184	['MPAKAFSRDSHRNDWPETKVQRVFLRSRRRGAWHWLRRCSRVNPLLQAL']	49
4348636	4348914	278	['MRWHKSPKPNTCTERCDVDPTGISRKVARITLGDLYACHCATARRKATG']	49
1113332	1113523	191	['MVVLRGLAPLVCALLAACCRRVEPNADFRIYIRNWRGPPFQGAGPITS']	48
2626690	2626912	222	['MTKYFGCRPPPAPRSPCWKQAWRCAVRALKVNRPALFSVVPCGVTSAA']	48
1712024	1712170	146	['MSGDGHTGPAQGRGGRLRQAQPPARSAASGAFLQLWDACLAAGRCVS']	47
2488144	2488316	172	['MLDGSPLPGLPSGRPGNGLGSVSRAWCMQRWFQDPPDRQAVADDAGA']	47
3054614	3054804	190	['MAGRLKAWRRPREQLSCSISESWRDPCGSGLASRKGRAAAPGSRHHA']	47
4551906	4552086	180	['MGAALCRDRAAKRPLDVCIAAKTAGAAVRPDRDTRPLPQGPRQLFRR']	47
818678	818846	168	['MMCAIVNSSQFYAGKIARSNGMPKISCARTGVCPMNGCAKLDLQST']	46
1296375	1296561	186	['MPDRALSRVNPLLQGTIVPCRRSPSPSGLRCRAANMRCTASVPQRR']	46
2573090	2573237	147	['MTGQSSSPSLYRSASCRNGRKRFCPLERILMIRFRWPSMNWVPACH']	46
3023072	3023265	193	['MPTCPHPTTPFPLRASSPRQARALPLPRNPRNSRSSDRGALSRQVG']	46
3390569	3390730	161	['MHLALGLIRARAFHPVRPRLRLNTPTRLLGAHRVTHLPGAGVLKID']	46
5974100	5974318	218	['MVMGHRNCDILFEIYRTVSVGCSEHAGKKLCPLSVYLRHFTTPPFG']	46
6048579	6048815	236	['MSVPALSRVNPPLQATRPLWERVHPRRRHRRVCRYRPAKQRFPIGW']	46
6150380	6150560	180	['MSFPRIARSSDYPVAPSYRIPNRADYHPVAHAVASGHGRAGTIRCA']	46
2460804	2461000	196	['MWRGAYGGFTLGPPPLPWVRHDDIRRRRSPRILPHRRSDRTSNKPP']	46
4741276	4741507	231	['MHCCVVWVFCCRLWPGLKTGKWPRTRTGSRCPSVRWPGPSTRPIAV']	46
25576	25817	241	['MTLADRLLYKPIDFLRSRCHCRSKNSPHASGRSQTCTAHQGVSCF']	45
445027	445313	286	['MRHPLFQRLPCRSGLARECVRSDNNFVWSGLFAGKPAPTRVCGMA']	45
4169967	4170207	240	['MRSCHGEGARSCTQVLTRCPVAEFAGPRERLAELGQRSNRVAVGA']	45
4221194	4221381	187	['MNSATELRPAVGRGNVRACSELEMQYCADSYGSGCMRLWLRRARP']	45
4472836	4473058	222	['MEWRGVLNVHFDMRLACPARCWMMASGARRPKDGPEREDGSLKGP']	45
707354	707512	158	['MYRSRSRKESFPLVGAPLGGTLQVALTSARLIGMIRGLFCLLFT']	44
1712014	1712235	221	['MDSATRAKSGKSGMTINRSEGCSRYTAPGSQTSVPKLQECATCG']	44
2085456	2085621	165	['MVVAKVTKAVGSPHPAPTLRSGVPSLRACSREDRAAGPILGPSA']	44
5152206	5152478	272	['MDTRMPSNPRLKGEEGGRSMTWISCATEHTDNIPKHQSGRLRRV']	44
5400810	5401036	226	['MIDSLWPSVAPIVGCPSWQGGAWSNLSALNAKLGYSQNPGVVCV']	44
194089	194316	227	['MLIWQSVSGRLTVTWILTPVTLTLSSKSVAPRDLRLLRCMKKC']	43
1414468	1414611	143	['MIGAARPAGPSTLLPASVQDLACEPRALPTGGICPSCADRSSL']	43
5552374	5552536	162	['MADISDLEGFPRSMKRVLLPSWVGLVVSTTLFFAFNWLVRVLK']	43
6168799	6169060	261	['MVLVLSLLPVAGAASTVVSLRVENRFKARRPTFSATIVWRLNG']	43
1441829	1442025	196	['MVPNARERGRRPHWCAALLPNERVAQMASALIPVGTACCDHV']	42
2603493	2603625	132	['MPPLGLHQPWSQRRQRRLRRPGLPGLLPGTTQRLRLRFSESS']	42
4022681	4022876	195	['MRREVTGMDARQALGPRMGPAARSSREQARSEGTPERSAGAG']	42
5260639	5260775	136	['MLPVPALSRVNPLLRRRAPNHRCVSATDQRRSGFTREEAGPA']	42
509115	509374	259	['MALARLTLSPAGPPGPSPFAFSFTARCGKAAPVAGAGRFES']	41
752288	752439	151	['MRRRSGKAKTAQEAEFTGCNRAFCGCFQRSIIEHQGFSHRP']	41
4449516	4449771	255	['MDTHRWLDTGRRRSGFTREWAGPAGQSPPADHRCRHHRIRG']	41
27328	27554	226	['MSLPLGLMDSEATSGSVIPSVRTTRLISQSVELILADSPKL']	41
2154010	2154145	135	['MGRPADRAPPGRVNDQGCWTAGPAYSRVNPRLQHTAGHPDR']	41

5007972	5008201	229	['MDARRAETPLGGSVHDSRPARARPTRLPRSSHLPRWPDTRC']	41
32680	32891	211	['MTRRMDRGSDYHMPYPQPNFRTACNHNLGGLGQLLPHTLR']	40
2857201	2857551	350	['MKVICHFCQFFLIRQALLWLFLNCTMVVMPLSSPTRLFGT']	40
4963683	4963881	198	['MGWVFALGVAFGRGVRLEVFVWAALDLRGTTKLTANPSRS']	40
5966014	5966146	132	['MLGLPIACGQQGLMDLSGPARSRACWVPPPIRSADPRAVA']	40
949683	949816	133	['MHRMRRGKSCAPLDPDQSAGLAGCGFEAYQAGCDKLSRS']	39
6052440	6052570	130	['MMRGPCRIGFTREEDTAVDGTGFARVRGQARSHRDRASF']	39
1388489	1388625	136	['MPPVRTVIAGCIGWPATCSRSISHAHRAASSSCDRPYQQ']	39
2303074	2303205	131	['MPVRASSRACPLPLTAYAENLAPYLWERACPRTPAQLPP']	39
3343007	3343214	207	['MDPSGHLWSVATGMTILIQLREHGESIRSRRAGRQASLS']	39
6052439	6052602	163	['MWELLSCLISKSWRDPCGSGLAREHGRSPCHPPRCLLRG']	39
6062887	6063018	131	['MRAGPVRGQARQACAVTVGAACPRRGHQQQHKPQPEPRK']	39
526470	526727	257	['MSSHTNCLIHCRRRFSWSESGNEHSHRMLISGFCQIVL']	38
948986	949157	171	['MPPGVEPTSLRPVSATLDSPAQCFRDAPSACSQAMPSV']	38
1327024	1327297	273	['MSSHTNCLIHCRRRFSWSEPGNEHSHRMLISGFCQIVL']	38
2034781	2034983	202	['MPGEGWHPDRKAGADTRCDLGRNFVTECQMASKQRPFA']	38
2488379	2488522	143	['MPKGQCAQTCGRGRGGHPYHATVTWTVRRSRTRRKPAP']	38
3342197	3342429	232	['MFAVGIALLAVGLGTSNPGLWIPGAALMAARIVKRARS']	38
2436529	2436714	185	['MNNPACLAFSAQQNAKRRLGTPPLQGYAEPPLCRYRTD']	38
3703695	3703944	249	['MGSMHVPGCSEEFRHLKPLVSDCQCSVDLWRAPVALNS']	38
5560497	5560797	300	['MQRHPAWQLMHQAGCILNEKGCMMRQPGNFPAFLRLQG']	38
9346	9502	156	['MHSPHGRLSTGLIHRGEKPFWSINGCFVVVPNVSTCG']	37
157729	157930	201	['MRGACGGGLRKGGHRNTSESPLLLLIGPGAVLTGCPG']	37
796800	796941	141	['MRPIATQGRSHRYSIGLLWECGLVSRSGCGAAPGLLL']	37
2158392	2158557	165	['MTGFVKKESYCEPKIRYKAATRFQAAGVVTSLKVRFA']	37
4221262	4221390	128	['MQPLPYESAQYCISSSEHALTFPRPTAGRNSVAEFTS']	37
4414891	4415056	165	['MKPPSSDPTVRLGGLHHAFTLFSRPPLHSEGSSASLR']	37
2657668	2657818	150	['MQPIRSTRLLLQALQLIQALRSTCRNCLVLRMGQTQC']	37
4960913	4961068	155	['MSRMGREAAPGLQLRRKDRRGRSAALSRRKAAPAGIV']	37
867906	868066	160	['MPASSRARPLPQESRRTCGHYRTCGSGHAREEGGTG']	36
1184557	1184682	125	['MRSSAARAALDLTGAECVEPDTCGARSRLKPDTSEN']	36
1999911	2000058	147	['MMAVLAFMLNNPTKIIIASGLGRLDVILHVSAVWRS']	36
2065145	2065271	126	['MADRLLVVIPEYPLASLSSRFGSIAAALHGAGLLAS']	36
2153947	2154153	206	['MWCLADQAPPARRIARGLAHRSGCPAVCCRRGFTRE']	36
3030115	3030320	205	['MDTVSRRCLTSAEWPGQALEPEHRSRHPGRASRARA']	36
3030313	3030458	145	['MRRGRPAPGRGVCTATLTTTSTPQVVAGCWRVFMHA']	36
4192583	4192717	134	['MRRAWAMRAQRELVPEGSATAKALDYSLKSWIALTR']	36
4861064	4861284	220	['MQSKRAAWGGAWGSRCLLTVPASSRVNPLPQGRHMP']	36
5266622	5266745	123	['MERPPTESRASQAALPHLLRSRASAARLQYLMHMQL']	36
474984	475288	304	['MTRRARPTDEPLGAGDEHTFPRTSPSSTRPLRRVSD']	36
1149874	1150005	131	['MGAALPAKPTPWWMARASPVFAGKPAPPGIAPAFRN']	36

1273212	1273376	164	['MGAGAPANTGEARAIHRVACFAGLPAPTVSVPASGS']	36
1349029	1349153	124	['MRKRKRPRPRRQEGYGTNEDAGTGHASLSGRGRGQI']	36
1747510	1747792	282	['MTFALPLVREVDGSERHDQCSPHRDQKWTRSLSAHG']	36
2589425	2589556	131	['MLLVHPRRCGTSTVEALWLLLRMNGELGVVSGSAHK']	36
4393259	4393441	182	['MRQAWRVARKRNPLGCGGQVGDLEVKVLYTPGKGKC']	36
4856064	4856185	121	['MGAGMPAKNIARWMARASPVFAGKPAPTGIAPAFRN']	36
2095064	2095205	141	['MAWGLWDRALQDVRNIRLGGCVKAAGCLPQVVWCL']	35
2719459	2719651	192	['MKAAPLATKCYGFNPCDNRPQPVFKGLQAIPAMVG']	35
3352110	3352240	130	['MLMSPSLRYLKRLVLFRGSDSVASSKQGVHRGWAE']	35
1350207	1350363	156	['MDNLNTPTLFIRHNRPSTLSGPNPKPGSAPTNSAG']	35
2150899	2151285	386	['MRCFTCLPDSFRCTSARESLPRIARRMLRKYLPWT']	35
2245445	2245645	200	['MMASSRVNPFLHGGLCPTTDLCPSSIRRRSGFTRE']	35
4224520	4224705	185	['MEGTALSRLGNSQPSRISLLHRTFLGAFPGLHGQS']	35
4488710	4488882	172	['MSSLGHLNGWVDSDGRRVGTMIADRQRRPQPATRR']	35
5047028	5047225	197	['MMERLSCRRFLTPSVSEMTRWKLMDTPKGKRSAPA']	35
5078755	5078930	175	['MYDCARQPRRGLWKTTMDLPSSYSLTRFIQKDLTD']	35
5573963	5574113	150	['MGAGRRRSGFTRERAGPAGQFPPAVHRCSRHRIRG']	35
5756495	5756707	212	['MCPRRGRSGFTREYGGGGNAERQVEIGQQARPLRG']	35
2280944	2281062	118	['MSGLREPDTCTRGSGHARETGDAVDGTGIAGVRG']	34
2295457	2295577	120	['MIIGTAAMNFCVRPRRRKFDPRPKRATVKPPPSR']	34
4379538	4379757	219	['MLGGLRWRPFIHCSGLWRRIASASSRRKSPHTTG']	34
401528	401653	125	['MLLANIVVSSQCAVYIEATIRPVGAEASVCDGLT']	34
1274287	1274489	202	['MYYVAIAIATAALFIWMATDASEVQLIWAIICHS']	34
1883670	1883797	127	['MRPLWLPCSSQVRSSLNKGVTGTFLSCTDHNPRV']	34
2406832	2406988	156	['MAPALPVFAGEPAPTTAVPNHRCCAGPSMPVSNH']	34
3414626	3414770	144	['MSHGPPGVYASRYRKISGRDKATWWCTLAHPASQ']	34
3648884	3649182	298	['MPITSSAISRARTGRSVPSWPTPWNGRPRPAVMC']	34
4306868	4307050	182	['MRRCVMCTGVISNDACFIRIGYALTVFFARDYVL']	34
4349949	4350119	170	['MYDSNIYDPPELSTLGYLYDGNGRYRGHCKVTAA']	34
4675985	4676103	118	['MYEGIDRETDVEVFALRKKNSGTGERSSITTTSI']	34
5041138	5041426	288	['MDSAIGLRSRLGGRETAIRRLGVLLGVSGSSRRA']	34
5044720	5044917	197	['MKKAGSRAWYQGMRDGGALATLRKPYLPKNASGF']	34
5373356	5373489	133	['MPAPTGAALLLGWCDPCGGAPARAEANTDVSLSH']	34
194307	194494	187	['MYKVNSLAVRQFQKVIDGILTSKTSRDHRHSSA']	33
1360740	1360931	191	['MPWVSAGATQVSGTVRLLWERACPRKGRYRQYT']	33
4372233	4372366	133	['MHSDQLEGTSEKRRTYLHRKWQVRSDYLRELPS']	33
6159279	6159484	205	['MHVRTSNDSRLEMCRNSSLLTRATAVKGPYLLR', 'MPSRLMEECTYERAMIAALKCVEIRHYSPAQQQ']	33
1728612	1728728	116	['MLVSVGMGLPANTGYARGSHRGGCFAGKPAPTA']	33
1828849	1828976	127	['MIPAGATLGCLLLHLIGLLYADRRRSGFTRERA']	33
2660877	2661053	176	['MPRLRWGMDSLSVMHSCLLTDKRGSLRSNCQSP']	33
5971868	5972007	139	['MFLTKFLGWQLGAAWSRTLLGMDIKLISRCSKV']	33

30145	30286	141	['MTAGALPLLPPSKICRAAGSAGRPTIRQHAFV']	32
130295	130438	143	['MHRSVPGVLKFHETIANSFMGFGTKPVKRNSE']	32
131277	131468	191	['MSFPFGWRKYIMPQFLCAHSPSEGERAWGVEQ']	32
1013175	1013302	127	['MQPFAGTPAPTGITRPGGHVIPVGAGVPAKGR']	32
1246600	1246827	227	['MHVMGSHQGRQGFVTTFSAPVRSSAECFEADT']	32
2054666	2054814	148	['MREGSMQGRSGPFAGEPAPTGAAHHWNLVVTV']	32
2821335	2821479	144	['MLCMRDLGRGGGSNRRSKFYLSLSDQSTGMAS']	32
3520963	3521461	498	['MLTIRSIVVTSTTRRIIATSTTCNTAAGICTS']	32
4302458	4302590	132	['MDHSLIREQSSTPVRSTASFPRIGNHSKVFGG']	32
1351962	1352095	133	['MNGRTPNRSKKLPSPARHAPRSKVNSGILCAL']	32
5971727	5971866	139	['MIDENLPLAISCSAVTSVEDLVEQADPKVTGR']	32
324595	324702	107	['MKLPVGPPGWIRIASNLRRGVGSWRHIAARF']	31
1120261	1120403	142	['MALCVPALWTGLAASSRTSGKAASLSLWQAC']	31
1151145	1151270	125	['MSCLHRCSGLLTMSLGRNRLEGKCSVNRPYN']	31
1298152	1298303	151	['MGCILFDERSPSPSGLRCRTANRMRMTSDLH']	31
1307629	1307805	176	['MRRGHANALAHLPPEHVPELYRYSVERCSSA']	31
2395393	2395508	115	['MRGRSVLIIASEYPAAREKGFGAVRDDGPTT']	31
2446487	2446611	124	['MLTSACMALPVANGACLPVSAAPRWASLCLQ']	31
2856848	2857040	192	['MKQKWVSWASCGATWCFCCCASKCWQQYHRI']	31
4748066	4748293	227	['MILLACGKFKQTFVLDPPSGVDIMRGQDKTV']	31
6060375	6060544	169	['MDCGSGLAREHRRSRCHAPRRLFRGHARSHR']	31
1311691	1311801	110	['MGAGTPANTGKAGGIHRSGCFAGSPAPTGRH']	31
1777333	1777545	212	['MGLLATPMYTMTYTLYGIKAWDAKLNRLKPS']	31
2302360	2302504	144	['MRSAPRVFASQPNRRAMLRPSATGPARQLPQ']	31
5222848	5222956	108	['MGCILFDERSPSPSGLRCRTANRMRMTSDLH']	31
5466642	5466836	194	['MPVAAMYNDRFCPQRHKAARSRARPLAIKLD']	31
5909224	5909342	118	['MIQGLVTCQTSDCFSRRALCTVGSGCWTVAK']	31
6133852	6133997	145	['MTCAFEQVPKTCGIPKEAGPAIERPGYQRII']	31
589723	589876	153	['MAVLSGETVESRPVASNGSGVGLPGQETAK']	30
1168483	1168703	220	['MCALRLVGRLVGAPFFCLGYLQWLCWPFRG']	30
1700324	1700459	135	['MPYRHTTARRVCRHAADSVILAGFSPNSHE']	30
475601	475696	95	['MGAEFEAAHDSPRVGCRPAVFDQRLCIPAC']	30
752288	752453	165	['MMLDDAALKTAAECSVTASKLRFLGCFRLS']	30
1191447	1191552	105	['MLAVPASSRASPLLRGVHWFQNQRRGGFAR']	30
3948695	3948842	147	['MWGTGWVCLFSVCQLLAGTLVKSVIYSGAT']	30
4302283	4302527	244	['MTTYRRVNTSPPLTPRACFFARKNSGLSFV']	30
5728637	5728921	284	['MSGSVWGIITPSRGPNSARQPGICRWPLVS']	30
697157	697272	115	['MLAVRCVRHNPCSACEIERRPRGARSHRR']	29
1105279	1105448	169	['MQERKQARNGRGRARLCWLQMSIPSLFLT', 'MRKVAERRSCKNENRRGTAGVAPDFAGCK']	29
1745844	1746077	233	['MMLRRALRMLTSGSSNPGRRRNGQVFITA']	29
2233046	2233198	152	['MAMPASSRVNPLLHRIGGGHTSVVEHRPS']	29

3226945	3227046	101	['MSMAMPDMEQYVTATGHRRGGFFSMAWRQ']	29
4196377	4196525	148	['MLRLGMTQNETAKKLGLPTTTVHLYWHRQ']	29
975151	975303	152	['MLFCAWWFCVACTDLFAGKPAPTWIALPL']	29
1565400	1565531	131	['MDFSQDFGLFSPYFVIFSRYLPPVPPPSA']	29
1747804	1748038	234	['MCDLNAQTAHNWKWSARLAYAGVAPAVTR']	29
1822539	1822671	132	['MLVLQVISLACSHMPVNRAALRGRAAAVA']	29
2035231	2035358	127	['MRLIIWPIITFLTAPSVGTAYDEGLLISL']	29
4936691	4936840	149	['MAFQELKPLAIAVEAIGREVSFFLVIACS']	29
5075352	5075524	172	['MGWKFLLSALLLWVSGYSASLLRLLTCFG']	29
5338005	5338111	106	['MLPGVFVVTRSAPVRPSAARAALGLTDAA']	29
5411662	5411758	96	['MRPCSRVNPCPQVRKILKLAMNPPYLQRD']	29
5574118	5574245	127	['MGVGNCATHHLNAGAITVGAGLPAKPPAR']	29
6166669	6166823	154	['MADQGSMVLGWTGILCTLRIGGNLLGGSM']	29
236918	237118	200	['MFHGPSTVGQRTGSESAGPSESLAIAPV']	28
1349573	1349779	206	['MSSPGTSLCRTPLVGVGGASSFAERLTG']	28
3497906	3498039	133	['MRRPPVRCGMGARVRMMAMRRVSSRPAG', 'MRDGGACQDDGNATSFVAPGGVTVLFKT']	28
3971739	3971967	228	['MASRPGSFAGKPACMVGLEGRRRAITRF']	28
5516025	5516182	157	['MCMLQLKIGVKHSLSRPTWVCLKADDWS']	28
5699304	5699522	218	['MGVASSLHDFAGRQNPRSTAVPYHWSPE']	28
5912275	5912448	173	['MKRKAASRCEDRSGNAKRQLLRPGRPRH']	28
484438	484594	156	['MRVIGSVTGCSGGYTAIGQPGCMPASAD']	28
529669	529779	110	['MDSPANSDMRYLWPYEFSGLSSSPPQSA']	28
1501714	1501896	182	['MTTLPYKNNDNDVYPTLSVAAARQHPGH']	28
1784902	1785013	111	['MAAFIPTDRRPALLESGLSRLFAPCTAP']	28
2087801	2088020	219	['MGFWAFGLLGFWAFGLCAFASSRHSFTM']	28
3536925	3537038	113	['MGNGRIRQKNDMQIMQGRAVAMNLAFHA']	28
4228256	4228404	148	['MSHNATCRHRRRQLGSTRKLVGQITVPF']	28
5075636	5075748	112	['MRLAVPSGWSRDLQGWDAHGDWNEGDLF']	28
5401219	5401356	137	['MKPWGFTCRGRVKSNLMLVRFVRGSACR']	28
5549055	5549168	113	['MHCSGGCPRQTFPVSIRLCQNSYLYTPR']	28
6023086	6023224	138	['MFPRLETCTDPVGAGTPAPTGGASAFRF']	28
6158664	6158870	206	['MVRLIDASLWSSENGLRFDEAVNNACFR']	28
41915	42056	141	['MIEPDSTSSFASNGPTSAATPTSLIRG']	27
58372	58582	210	['MAAEEAGSSEVVTAGIASRILNPRSPD']	27
466622	466726	104	['MPMVDGHRPMDWNVQWCGLLRLAQAPR']	27
709141	709233	92	['MAGQASVFRIPRRPWALLPVTSARVWL']	27
753028	753185	157	['MLRHQCAYSPTTMSSRPGSIPVTVSAQ']	27
1298318	1298507	189	['MRGKEAVDRTGVLLCSRMREWPKWHRR']	27
1367434	1367568	134	['MAFSWRRPGDGLCKSTCCSRRLYSRLL']	27
1429562	1429783	221	['MRNPLGCGGQVGDLEVKVLYTPGKGKC']	27
1828713	1828856	143	['MSNAPGCLHSIHRSGFTRECGSGCDGE']	27

1983628	1983805	177	['MPRWARIFIAAAPDRSGCQILRYMQVS']	27
2470263	2470426	163	['MGVPSAPYKVPLTVAGQVFLSYNSSRV']	27
2472099	2472226	127	['MPASSRVNPLLRRPAPKHQCVSTPNQV']	27
2672454	2672582	128	['MRAGAGMRGMHGPGQRRSTGSTVLVTL']	27
2925563	2925755	192	['MRGKEAVDRTGVLLCSRMREWPKWHRR']	27
3392047	3392211	164	['MHRTTPWFMKMSSASSSPNAATLLPVP']	27
3517631	3517826	195	['MLGRLMAGGLHKIASIVSGINIRLCIW']	27
3518858	3519132	274	['MGNVFNRYIQVSMLTRWALIWIRQGNT']	27
5000531	5000672	141	['MLSTAAWRQRRRRRQDGNDENGVPGRN']	27
5549065	5549161	96	['MGYIDRNSGIGVLKPEMSALDSPRCSA']	27
5973693	5973820	127	['MRNSLYWAVLFWRHYRSWCWCIYRIVI']	27
16182	16320	138	['MLSLFLNGHFVIAGAIWIRSEALASNS']	27
85022	85144	122	['MQPIATQGRSHRGSAYDVLCGNGLASR']	27
1431422	1431535	113	['MQASSRVNPLLPGMAISLRVALGLAIE']	27
1457537	1457629	92	['MAESRMDVPFPTLAALLGLASGTVTHR']	27
2016425	2016526	101	['MVDNISLGRHLRHSQGAITGRLVHARV']	27
2034187	2034402	215	['MDDKICELPTLLTVPPLRRNAACYSPP']	27
2034993	2035176	183	['MSRNRLIHRSTSDFLDRKGEVHWLSIG']	27
2057786	2057907	121	['MPELASSRVNPLLRPTVAPPLGIAQSM']	27
2710888	2711007	119	['MPASSRVNPRLQRPVSNPRPVSTLVQA']	27
2867347	2867440	93	['MSQPRSTRHRPGSCSASATGLAHWPCG']	27
3525196	3525470	274	['MTVPSPTLTPCTCRPFANCCANTTAAS']	27
3826281	3826465	184	['MRGKEAVDRTGVLLCSRMREWPKWHRR']	27
4073719	4073902	183	['MRGKEAVDRTGVLLCSRMREWPKWHRR']	27
4073974	4074061	87	['MFDERSPSPSGLRCRTANRMRMTSDLH']	27
4440342	4440463	121	['MGKLALFLGGFLLLTILIGILGTIPPS']	27
4491871	4492052	181	['MAGIRLTIHNRLSIQAPERYSSGALLS']	27
4636211	4636315	104	['MSRPQSMGIVDPHLYGCLQAGSPYLRR']	27
4861055	4861284	229	['MLAGSGIVVACAGLFAGKPAPTGTSPP']	27
5008210	5008331	121	['MLGASGSPQRTLSVLVKMGSISARGRK']	27
5080139	5080298	159	['MRRVTGAMVVVRDPGPISKTVTPECFA']	27
5222600	5222793	193	['MRGKEAVDRTGVLLCSRMREWPKWHRR']	27
5312685	5312845	160	['MGACEYGLIVKKIISITSDHARNHIVI']	27
5390526	5390630	104	['MRAVYRCGLSKPVVDEQRPTETKTLKK']	27
5545354	5545528	174	['MRGKEAVDRTGVLLCSRMREWPKWHRR']	27
172966	173196	230	['MSDRSLKIRICDRYRLNTSFTAGGSG']	26
178402	178629	227	['MSDRSLKIRICDRYRLNTSFTAGGSG']	26
475066	475164	98	['MCARRQPLRARQSDVHVWSYSKEDQL']	26
699607	699845	238	['MSDRSLKIRICDRYRLNTSFTAGGSG']	26
1820032	1820232	200	['MPLCYRASKGDGMSVQKSAEAVVSGG']	26
2550533	2550712	179	['MSDRSLKIRICDRYRLNTSFTAGGSG']	26

2848792	2848901	109	['MGIYKISAWGARGRVFESLRPDHIFQ']	26
2939076	2939257	181	['MRITLLRGCADTHQCWLRSLSRAHEP']	26
3827998	3828212	214	['MRITLLRGCADTHQCWLRSLSRAHEP']	26
4075434	4075656	222	['MRITLLRGCADTHQCWLRSLSRAHEP']	26
4444085	4444188	103	['MLLRGDHPDSRRSWTGFPHRSLIVIG']	26
5224309	5224549	240	['MRITLLRGCADTHQCWLRSLSRAHEP']	26
5280875	5281001	126	['MPGLLCSPFAGKPAPTLIGVGVRRSL']	26
5541032	5541178	146	['MGGASGGWFGGLGEGARCTTQQFQRL']	26
5934018	5934116	98	['MYPSTGRRAKTIGTGRCRPGRGIIAR']	26
157734	157877	143	['MVILMYSCARLFVVPLRMLPAQRPTR']	26
435267	435474	207	['MPRLLLHVVSRGADYKTSFGCLPRRL']	26
733920	734019	99	['MPCRRGFTCECDGGCTAAFAGEPAPT']	26
1058887	1058984	97	['MRRKEYRLADAYRAARDRPCANRKQT']	26
1440085	1440307	222	['MRITLLRGCADTHQCWLRSLSRAHEP']	26
1811394	1811516	122	['MLLVLASSRVNPLLRRSVSNHRCNPL']	26
2479053	2479183	130	['MPAQAHGQATGPGIRYVPWMPRHALA']	26
2821562	2821662	100	['MMTLRQYTDQPTHSELMAAQYGEAPT']	26
2900629	2900724	95	['MDHRRRHTDRCLLPWHNLWRNLRKQQ']	26
3754307	3754433	126	['MDDKRTYQMERVRLNSARLGLPADCF']	26
3763733	3763851	118	['MSWATHWRLWWWRSGRTNTGCLLPRW']	26
4382845	4382937	92	['MDDVLRNGKQLAADDAGDPIAVERDA']	26
4514994	4515170	176	['MLKLQLYEFAPGQTPIRRAAMAAGGQ']	26
5081539	5081639	100	['MEGLVATQATPVHPFMLDCFLGPCMG']	26
5310592	5310809	217	['MSDRSLKIRICDRYRLNTSFTAGGSG']	26
5453099	5453321	222	['MRITLLRGCADTHQCWLRSLSRAHEP']	26
5751669	5751807	138	['MRGVSGVPGSRASSRVNPLLRALHRL']	26
5793347	5793452	105	['MASLELQVMAVRGLPGLSDRLAACSV']	26
5849139	5849237	98	['MRRRPVAIGTDGAKTARDRRVLDQSS']	26
5874546	5874649	103	['MALSRACPLPQVLRRPQGLRYLGRSG']	26
9111	9223	112	['MKKKKRHIKSFFEELITLKWITFCG']	25
27155	27265	110	['MADSLYQAVIDFQLLKAGKPPSTIQ']	25
81482	81572	90	['MGTSAPRLKPSPNNLAIGCGNPGSG']	25
176363	176564	201	['MQVGVLSLSRSSCFVLVSRKVSQLQ']	25
499165	499282	117	['MPETCRSLCGSRHARERAWPADLYG']	25
741457	741575	118	['MCARHGALRVSATRLVVVAKWVTWR']	25
977834	978013	179	['MRGVCSRPSSATIAPASLMQSRLKQ']	25
1685135	1685227	92	['MLVCRMGGALSVGVVLRARGAGRRR']	25
1817558	1817669	111	['MRRGVFLCRWRILVASPCLSGGAKR']	25
1818649	1818754	105	['MYPACSCDVEGTGSSAAREKGAARV']	25
2271618	2271710	92	['MGAWHRSEATRHMPWADPSFNRAQA']	25
2271792	2271933	141	['MKQTQEKRSRYRFFCACPEAIAPEG']	25

2796816	2796927	111	['MIALFAFAPEAACNVWSRSGHTPAC']	25
3578352	3578478	126	['MFMSESTFRRAATGIFASCSECMTA']	25
3673844	3673937	93	['MRLTTARTTNGWRRVVSGSAPPGAR']	25
4174389	4174492	103	['MSRWRRRACGFWLSGRRDCGLERDR']	25
4521893	4522151	258	['MVWYTPPRPKGHRGESQENADRPRP']	25
6156238	6156560	322	['MPKYRGRAGFLSQLYGGYAVATNLG']	25
194402	194491	89	['MTRWGTSLPERRLLATSGGTMSMIS']	25
410802	410890	88	['MNVFPHKPRCDQTSHKCSFTTTRVV']	25
1443071	1443225	154	['MLTIDFACSGTFAGKPAPTRHRAPP']	25
1595520	1595617	97	['MQPFAVRTAPAKGGATPTAMPARAI']	25
1617758	1617888	130	['MCRDRARSGPWHSRYDASYRSHRQV']	25
1691748	1691857	109	['MIVVRNFAPVRRSAARAALDLTGAQ']	25
2034399	2034759	360	['MNACLTSHRLHFRGRSASGKGNCSD']	25
2807779	2807942	163	['MLQEVERIPPPSNTRVRCSSFVPGP']	25
3595785	3595872	87	['MCARHGALRVSATRLVVVAKWVTWR']	25
4003865	4003948	83	['MTGCVNRKQAGCQFRQASGQSANPA']	25
4905413	4905524	111	['MSVRRHSVGECLRTILRELAKPVHS']	25
5676054	5676142	88	['MPACSPVRLRAGTATLTLRLSLMFS']	25
874324	874421	97	['MPAKRPVQATKIFPAAMAATGDSA']	24
1427671	1427899	228	['MINKKPALKNTLYRGQPQALTLFS']	24
1879816	1879950	134	['MLRRCFKARGSSGDADCDFRRPFS']	24
1883897	1884017	120	['MGRSLFQQGRSWGERACKDACGQN']	24
1915650	1915764	114	['MFLTQIVCAAGSHGVKRLCKKRPI']	24
2050508	2050587	79	['MPPRRRRSLRDCGAVCSLCGPLRG']	24
2069525	2069621	96	['MCYRASKGDGMSVQKSAEAVVSGG']	24
2715181	2715285	104	['MHENCSCWAGGGSAKHPGRPGNKS']	24
3288132	3288246	114	['MLADIDLHHQHGGQASLLLTCCWG']	24
3515975	3516129	154	['MLGSISGRKPRCPYKSSIQKLRLS']	24
3579414	3579519	105	['MRMVGTITRHEKCKRVCPMVCTWC']	24
3784547	3784724	177	['MLETVAGAVALAMGTLSLHAHNAC']	24
4192463	4192566	103	['MRQQLPDVTDLLGRYQVAENSLAA']	24
5000796	5000883	87	['MRRSLGSVRNPEKLNRRPSGAEIL']	24
5382382	5382511	129	['MAPDSSPAPILWERACPRRGRYCP']	24
5404015	5404165	150	['MAGATWVTMGSAAVAIRPVAKPLR']	24
324315	324406	91	['MIERTISWEQAERGRRGCYKRTHM']	24
997708	997810	102	['MKGCACAGPFAAEAAPALPVQERL']	24
1017205	1017313	108	['MITGIGTPIIQSSIERMAFLLDRW']	24
1275675	1275835	160	['MPWLRRSGEGRQPSNQLKAAPATI']	24
3595443	3595548	105	['MCYRASKGDGMSVQKSAEAVVSGG']	24
3749353	3749484	131	['MAWRKPGIRWARAGVSFAIAPARP']	24
4393015	4393120	105	['MCYRASKGDGMSVQKSAEAVVSGG']	24

4933047	4933137	90	['MCKVAMPRGTQNCPTCGLSPSETR']	24
5122749	5122854	105	['MVKEACHERKNNVSQLGIRDREGT']	24
5377564	5377650	86	['MSSLQVLPQLAKVVAATKFKFRES']	24
5873961	5874081	120	['MLIVPAYSRVSPLLQVQRSPQGLW']	24
5975698	5975837	139	['MRFGTGAQPHQSQASFCYGTRAQV']	24
6148762	6148927	165	['MSRALGSVYVARWTANGCSLATRP']	24
492422	492549	127	['MPSARVWGRTPLDGLKKMLDNKA']	23
588961	589065	104	['MQPSRRIGWSIPLPCALVHSALD']	23
889278	889395	117	['MASQGDVGHFLLEVAGLLFGQGR']	23
901464	901559	95	['MRSTFRFAGLRVLAFPKCCVGPF']	23
1854683	1854820	137	['MALMRFIQDRRAALRPFRDTRVL']	23
2896547	2896670	123	['MRPVRGQARSRRFFTMLNRVRTL']	23
3070625	3070770	145	['MWTKPAYTDLRIGFEVTMYFANR']	23
3803949	3804055	106	['MRRRNGHPAYGRAQQALCRFRPA']	23
4441076	4441163	87	['MSCSTAQAYPSPCTSNLALRRAR']	23
5340743	5340829	86	['MWHQSAAYPYLGPDGPVPEHHAR']	23
5474928	5475061	133	['MCRSGFTRERAGTANITVQFSLD']	23
5669407	5669523	116	['MKLQVAGKGIPCRCHLQLRKHRS']	23
5759393	5759543	150	['MPVRASSRVNPLLRRPRERRNPL']	23
638038	638141	103	['MLASSRARPLPQGLHNLPVTTHQ']	23
691683	691767	84	['MGPPVDKLGALSCRLVISRPPEL']	23
1380279	1380452	173	['MRCRAQDELSLTLPSRCKRLRPE']	23
1848437	1848554	117	['MPIIPANPSRPYHAICSAYSYKL']	23
2035359	2035565	206	['METPAMMPPGPDRCAKRRLIQGE']	23
2655291	2655454	163	['MARASPVFAGKPAPTGIAPAFRH']	23
2856749	2856852	103	['MTHPDFSRSMPTSQLYEPNLALE']	23
3351479	3351579	100	['MKTASECSNTTRKLRFLGCFRLA']	23
3914808	3914970	162	['MKNDPVEWLGSIFCMGLCWPLRG']	23
4026269	4026372	103	['MCRDGPQGAPAGAVGPAENTNIS']	23
4492942	4493034	92	['MEGSEQPRSVLECDMDAQRLARP']	23
5073017	5073160	143	['MGREYTGANLNWSTLEHCRRNIW']	23
5541099	5541197	98	['MQQMWERPCVAMRRAGGARFHRR']	23
5545601	5545686	85	['MRSPVGDHPALPGCAVAQQTGCA']	23
5807877	5807966	89	['MSGLQGVGQGTCGAYQIERPRKT']	23
5926144	5926269	125	['MSVCDLAATMRGIRAAGVSNKKS']	23
5971306	5971419	113	['MRRGFSGQGFSLRSHAVRQKWIQ']	23
6048644	6048719	75	['MPGGTCKPCDAFFAGEPAPTGAA']	23
6174638	6174756	118	['MRLSKRVASATLFLFLFSPPVFH']	23
144592	144684	92	['MPPFSLARAPCRSPAVSLHDVA']	22
410702	410789	87	['MRLALWADVVNSWITGVSSGAL']	22
890727	890836	109	['MEFFVADAFQGWHRTPWDLPHL']	22

1276472	1276556	84	['MKRALLNHEIRRQIERSKRNQK']	22
2044528	2044608	80	['MWIGINECFTCVDLIWAGAVAG']	22
2064100	2064196	96	['MCWPVRGLTRAYTLMDGHHLPL']	22
2088010	2088086	76	['MAAMPPLGLLKSQSALPELPRT']	22
2230678	2230780	102	['MQPEAFGYRLEIRHCALGLAKT']	22
2598562	2598661	99	['MKARHYLDVCIESYEELRESTF']	22
4425218	4425335	117	['MDALDGSGGPCFCAPRKGRLSV']	22
4473194	4473291	97	['MQAEGRRLIADIRQRLPLLEVS']	22
5044696	5044872	176	['MEGRASVMLPELLRRAFPDTKL']	22
5248715	5248895	180	['MKLLDLTSRLAAAACSGLFFGV']	22
5955776	5955937	161	['MPAPTGPAQASRPRGTCGATAR']	22
5988712	5988867	155	['MARALRVVVPVRLGAARAALRG']	22
6150246	6150378	132	['MLGCGAAEWRWPACSWSWRGLR']	22
298009	298140	131	['MRWHCWNRVNQATTPVPFFLPL']	22
875893	875977	84	['MTRTPYPGRGGCTCHSRVDPPP']	22
2486793	2486898	105	['MRKVPDVKCTMAAPIEHRAGGA']	22
3664033	3664109	76	['MFLGVLAGSIVRPGSSDSSEIF']	22
3672865	3672941	76	['MGRPDHSGRSGCIERGLSAGSD']	22
3764724	3764854	130	['MSYPSYSRAWSARSQAMVIARR']	22
4320808	4320930	122	['MSSVSDRVWPITVLHEGQQWVG']	22
4533609	4533694	85	['MRIGGILAGIRKLCCLCGPLRR']	22
5163875	5163965	90	['MANFGAFGAALRPLADKPAARG']	22
5266621	5266741	120	['MCIRYCNLAAEALERSKCGNAA']	22
5331640	5331740	100	['MRRMLHLRRPPALRLRAAELTR']	22
5401081	5401199	118	['MLMAAVPENAKRSGYVGRVFDA']	22
5809288	5809376	88	['MTGQSCGSPALCVSSAVRQNSK']	22
668036	668186	150	['MGAALRRDRAAKRPRQFKRRS']	21
786579	786701	122	['MPALSRARPLPQESRILQALW']	21
861828	861924	96	['MLAAGWPMFIRERRVKKVPCC']	21
894533	894669	136	['MNGLDERLFSGRNSPARQPGT']	21
899614	899700	86	['MIICNGHGIEFRLRGARSKER']	21
915032	915177	145	['MGAALCRERAAKRPRQSKLRS']	21
958399	958507	108	['MQGSWPVVNLPATGCKAARDA']	21
1477167	1477262	95	['MRYEETRQPIPAAQVASSRPT']	21
1627145	1627272	127	['MLGGENFCAKPESMVASLNDG']	21
2137819	2137897	78	['MAALTAPVTLPGVSCTFFFGK']	21
2256257	2256406	149	['MPRQKGTRRKGGKVTMRHLRE']	21
2429871	2429971	100	['MLPVPASSRVNPLLQRPLKPL']	21
2643815	2643942	127	['MRGAGKPHLRAVRNTAKALVR']	21
2809393	2809517	124	['MQSRCCRLQLSFFFPVTLVRP']	21
3401709	3401801	92	['MRGWAGAVSAGCIRALERELL']	21

3647544	3647685	141	['MAWDIALAALRRSPGKGGASA']	21
4458752	4458874	122	['MLETTFTAASKPMELSERSTQ']	21
4851611	4851710	99	['MGLLCSPFAGKPAPTKAAYVL']	21
4852422	4852540	118	['MGAALCRERAAKQPRHSVLRS']	21
4965364	4965472	108	['MLDIIPTEWSVFGRLKFENSH']	21
5214265	5214373	108	['MLDSSRATSQDHLNLKRCGVP']	21
5773260	5773362	102	['MPSRPPEKLSMAQIVAELILY']	21
6000894	6000998	104	['MPASSRVNPLLQVVHRLRGLW']	21
194085	194207	122	['MSALLASISRSLSIARTLIAK']	21
403201	403332	131	['MKPVLYLWERACPRRPKPLPH']	21
668038	668187	149	['MGAALRRDRAAQQPQAFSFAA', 'MNCRGRFAALSRRKAAPTLSA']	21
726778	726880	102	['MFHISLPDPFTPFTTPPPRQH']	21
983542	983623	81	['MLRKIKPANYTRVPGDCLTCK']	21
1274868	1274977	109	['MPIRTGGRERILLAHHRRCGP']	21
1511570	1511653	83	['MKDRHSNIAPRPVLGYNCRPT']	21
1744921	1745153	232	['MPRAAAISWFVGSRSIGYSSR']	21
2059578	2059689	111	['MMFREPASSRARPLPQGYHRA']	21
2085361	2085455	94	['MPVTSLRKPCARPPEVAICIA']	21
2087613	2087700	87	['MPVTSLRKPCARPPEVAICVA']	21
2256117	2256197	80	['MPVTSLRKPCARPPEVAICVA']	21
2474114	2474218	104	['MPRLSCVLNSRACSRASPLLH']	21
2621636	2621738	102	['MTDTLRFPYSPNSVRHHETVM']	21
3070444	3070541	97	['MPALLRARPPPQVLHRFWNQR']	21
3329540	3329623	83	['MPCRTPSAFQVQDRHSPETST']	21
3647214	3647302	88	['MPGMAWGIALASFGAYIRADV']	21
3764434	3764587	153	['MQWRKESCCLSCSSRSCSALG']	21
4302202	4302282	80	['MPRPKNGHSRHARSPKTLFRG']	21
4382930	4383039	109	['MFRAMTHLMIFGRSRMSVMKH']	21
4408288	4408405	117	['MPSRSRPSRRNRLLLDDCGNC']	21
4747761	4747845	84	['MPLYPRASLGSAPGDYLFTGE']	21
5582459	5582586	127	['MYLPVFVGAGLRYWLDLAPGR']	21
5773098	5773195	97	['MALRQALGAGGVTCDECAHLR']	21
5947011	5947210	199	['MGPPCGPSRASPLLHRSLRPL']	21
6029464	6029546	82	['MRLAPGQNSDNACESSVRRCT']	21
6148653	6148776	123	['MVCSLACWRCPIRSNPPPRRR']	21
187111	187186	75	['MQGAIVFCGGKPHHRCFRLV']	20
586791	586890	99	['MWKDQRTAGAPLFVGAHGWL']	20
603611	603695	84	['MTPSARVQTPVEAACMRPRS']	20
743405	743488	83	['MCARYVKQPSVALHSGHTAS']	20
1420331	1420520	189	['MDSLDALRRTGLPLRPSLDG']	20
1440137	1440307	170	['MRGTEIVASTGVCLHTRVGE']	20

1441686	1441810	124	['MMRDHPALPGCAVAQQTGCA']	20
1504098	1504191	93	['MAACRIDVDQAKRGLSPLPG']	20
1567341	1567498	157	['MNTVTSLLRPHPHAVSLYPG']	20
1830109	1830260	151	['MFGAAACRPAAIGSTGLVNW']	20
2236951	2237031	80	['MAGTGCAGVRGHARSHRDRI']	20
2416469	2416580	111	['MTLQVGVVTCWGGQAHWLGQ']	20
2598379	2598515	136	['MTLSRLEVRTISSPFSRSTS']	20
2774049	2774147	98	['MIGCLYGPLRGQARSYTGVM']	20
2840587	2840663	76	['MRIKVVMAVWLCAASVGVRR']	20
3275886	3275970	84	['MIRLMVLPGSARRMARPVVF']	20
3351486	3351618	132	['MRRSSGKAKTAEEAEFTGCI']	20
3398858	3398934	76	['MTVGVVDVVLPTHPAWMESQ']	20
3450597	3450680	83	['MHEQYSEKAILSNVYGSRPA']	20
3571853	3571944	91	['MNFTAMTRHQHEHPFAGRIR']	20
4080857	4080936	79	['MCKVVYATLHDRDEGWLFRG']	20
4405155	4405297	142	['MTEFGGTGQKSEHLTSRRYP']	20
4632461	4632547	86	['MHSNEKRPRRAFFQRLTDFD']	20
4832027	4832176	149	['MQIRPITRSAKNAASYGRYL']	20
4866660	4866726	66	['MLVIEAVCAAGHGHVHFLDP']	20
5041534	5041663	129	['MIDTLRSTQATHETINVLGA']	20
5288372	5288459	87	['MGLFAGLPAPTSTAKSLKPA']	20
5453150	5453321	171	['MRGTEIVASTGVCLHTRVGE']	20
5477089	5477186	97	['MPVPAYSRVNPLLRTGSRLS']	20
5914126	5914240	114	['MRPVSPSSRARPLPQGNFEG']	20
589264	589352	88	['MMLLPGAVALSRSRCCLLPV']	20
1029013	1029115	102	['MGAREQDTAVDGTGFARVRG']	20
1582691	1582794	103	['MRGTLRGHLCRGLSDKLIKR']	20
2050535	2050681	146	['MRASSRVNPLLQAPRRLRAL']	20
2690491	2690613	122	['MATPCLQWPGAGKFAGILLG']	20
3477969	3478060	91	['MGRFGGLCSAARRVFRRAVH']	20
4292592	4292716	124	['MPCQTRQSWLDSHIRTGSGR']	20
4375990	4376078	88	['MCLVATQKTKPWMLVSRASS']	20
4731443	4731535	92	['MPGLASQEARPRGSRLVACS']	20
4896370	4896469	99	['MPASSRVNPLLRGTAHGPRL']	20
5190265	5190389	124	['MNAIPPGLFLPRRDRFMMAG']	20
5224317	5224488	171	['MRGTEIVASTGVCLHTRVGE']	20
5326431	5326533	102	['MEATLSCDRPAPGAGHLKCG']	20
5828851	5828962	111	['MELVPSFQVNVGSAHKVRVL']	20
306992	307097	105	['MRKSAVLTGGAYFFWAFLA']	19
917674	917750	76	['MDPCRSGFTREHRRSLLRG']	19
1626863	1626951	88	['MVTRGWHCETIVKRVGMPL']	19

1747782	1748004	222	['MCDLTGSAPGSFFRLSFTW']	19
1799832	1799938	106	['MGPLRGPSRARPLSQDSSV']	19
2026156	2026242	86	['MRCSMRVALRRETGTKRPV', 'MFYEGGPAPRNGHEAACVN']	19
2034189	2034633	444	['MILARREERPSGAGGRRKH']	19
2036286	2036387	101	['MYGSLLARKRMYRFLRSSL']	19
2087368	2087468	100	['MPPLGLLKSQSASPGLPRA']	19
2871119	2871204	85	['MKPTCRLSVLAQAVSGRRP']	19
2957055	2957195	140	['MGAGTPANTGEAGAIHRVA']	19
3199680	3199747	67	['MFGDDEFTTQVARPRWAEQ']	19
3403499	3403577	78	['MRCTDPAWAPSCSKTLVDW']	19
3805190	3805268	78	['MIVAAFKRFSTGVHPYSTL']	19
3969227	3969323	96	['MQPGSYGKRPANSPSELQH']	19
4073860	4073971	111	['MRSTASFPRIGNHSKVFGG']	19
4357682	4357907	225	['MPASTLADKPLTVCFGTCN']	19
4970243	4970372	129	['MDTAASGPGDSRVWPREPS']	19
5173036	5173107	71	['MQTGSSGSGPRVTALSGVA']	19
5222743	5222848	105	['MRSTASFPRIGNHSKVFGG']	19
5525710	5525804	94	['MAEKNNFCHNACSPPGPVH']	19
5679804	5679878	74	['MGAGEPANKGEALARHRGA']	19
5849213	5849291	78	['MATGLRRKGQEKLTPLFVT']	19
6150020	6150169	149	['MWHHAAIGVGGQSWPTRSR']	19
6166412	6166501	89	['MPFLTYPQGPKVALGPLSI']	19
837964	838043	79	['MIGLQCRAHCNPPTPRKSP']	19
1011475	1011543	68	['MGLDQNLPNEWRLCRPLRG']	19
1070222	1070336	114	['MEGERVRAARRMPLGFHRL']	19
1302553	1302672	119	['MRLSRCPTPQCSVADATKH']	19
1748667	1748905	238	['MHLHRAGSSGSHRSLKTSR']	19
1778166	1778261	95	['MIPSRCDRSLRVACPLLKT']	19
1983506	1983646	140	['MSPEERQRAHLQSAVQPPF']	19
3748760	3748851	91	['MHIPGELRRPAIIRPVLSP']	19
3899311	3899407	96	['MMGMVIAKVVLSIGLMRGI']	19
4876374	4876456	82	['MVGAGLPANAVVAATVNGG']	19
4967677	4967759	82	['MSLDSDCVSLIHRQAIGIR']	19
5161120	5161213	93	['MGEAQAPAPALACLRKGPE']	19
5324934	5325048	114	['MPSEGERPVPAHWRAGTGR']	19
5429078	5429215	137	['MNVGGRVAYVERTRIVPLN', 'MNPFCNVPSPRQAVKLRHP']	19
5614021	5614097	76	['MGSLRRPSAGAGLPAPASR']	19
5876176	5876273	97	['MPSITARTALLWSPDVALG']	19
5881269	5881371	102	['MLAYATQSELIPTGEECSG']	19
9225	9317	92	['MYRVLQLCSDPVLRLFQS']	18
29556	29633	77	['MAVAMELAGATRQTVDRC']	18

30254	30415	161	['MQEQDFAVSVAKRRSKAR']	18
335529	335597	68	['MLSGMASSRVNPLLHGGG']	18
557368	557443	75	['MTGSPEPGFLPKRFGWLS']	18
697363	697525	162	['MKFNELDNSEGAVRATIK']	18
850966	851043	77	['MTGIHARHLPAQGLRSAS']	18
1058090	1058214	124	['MNALYQSSRGLMLAWYAA']	18
1120416	1120553	137	['MVLDVLSHCLAPLHGSPP']	18
1569064	1569173	109	['MPRADHFRTHPNRAVRHA']	18
2004517	2004593	76	['MTCPENEIAYCRFAISYF']	18
2099932	2100063	131	['MAPLRGLARSHRYSLAPG']	18
2460894	2460974	80	['MSSWRTQGSGGGPSVKPP']	18
2676142	2676239	97	['MGIGAALCCEGPLQITHL']	18
3520744	3520879	135	['MVCVLLALLPLSGPSKRV']	18
3805275	3805338	63	['MDRHCPTRHPQGVQCLLG']	18
4121787	4121885	98	['MAKGCHLPWSPAAIICAL']	18
4372080	4372266	186	['MAFSAANCSNRSQSRTQA']	18
4467621	4467754	133	['MQLGKRIATPGAATTCVG']	18
4492959	4493046	87	['MSHSSTERGCSLPSTART']	18
4629133	4629266	133	['MYRNNDTGIIIAGFWQFS']	18
4716759	4716908	149	['MHTCTSFARCAVTFTTTQ']	18
5075432	5075548	116	['MKNTEKAANLVLQMVVMD']	18
5125068	5125207	139	['MPLPAPSRVNPLLQDAPA']	18
5342557	5342638	81	['MGCLAAHSRASPLLHVYL']	18
5371938	5372030	92	['MNNLGQALPGAAGAGPGP']	18
5883906	5884024	118	['MFATCAGLFAGKPDPTAA']	18
6158998	6159173	175	['MATTPVVRPPHGQEQRPC']	18
32748	32814	66	['MACGSRYLCPYAGSPEDA']	18
82024	82141	117	['MDGTGFAGVRGHARSHRY']	18
413988	414079	91	['MHTAQGLCSTCGSGLARE']	18
650158	650241	83	['MGRGSRETFQIRFLYFAP']	18
1011066	1011231	165	['MGTGFISIFGPDGVRLMA']	18
1099734	1099798	64	['MQPRPNGRRPPSGRHTHP']	18
1349674	1349810	136	['MGILMYWLKTRGVRRHSF']	18
2879005	2879151	146	['MHHWAAEPVLMCSDCSLI']	18
3354622	3354707	85	['MFCATKPGGRRRGLYPAV']	18
3370473	3370638	165	['MPPRCCPCSRKRCWSACC']	18
3572563	3572624	61	['MGPSMFCADRCIQTGHTK']	18
3654632	3654752	120	['MSHGSTLGAHLKRGIDHE']	18
4348883	4349050	167	['MRVSATRLVVVAKWVTWR']	18
4428851	4428959	108	['MRLTGCQRKNIQAFEKVV']	18
4484552	4484651	99	['MTDMPPTSSAWPLRPARV']	18

4610413	4610490	77	['MAIGNDGLVSEAKTSRIR']	18
4709596	4709676	80	['MCRPMFRSRLHAGMGKDR']	18
4884337	4884456	119	['MNPLLPGTAFALNPARPL']	18
5182324	5182438	114	['MGGIRLPFYRSASGREIR']	18
5604370	5604514	144	['MRALYIERLGAYCGSVVQ']	18
5753624	5753762	138	['MQETNHSSEKKRDLPIKA']	18
108675	108808	133	['MTAVPSCRRPCLRDNKN']	17
170930	171110	180	['MLYDSHPATRDRSEPSV']	17
524612	524688	76	['MLYDSPPATRDRSEPSV']	17
656400	656477	77	['MSECPASTKRAERIFSC', 'MSALRQQSGQNEYSVAD']	17
697486	697546	60	['MLYDSPPATRDRSEPSV']	17
1099698	1099810	112	['MRVPSRWRPASIRTRLH']	17
1163585	1163693	108	['MGVGDRPRQRYLKQQEL']	17
1275678	1275772	94	['MVARLSAFTGAPQPGHA']	17
1278506	1278572	66	['MISDSRWLHEVDIVNLR']	17
1330527	1330593	66	['MDLLESGDGTGFAGVRG']	17
1387817	1387895	78	['MMLPSGAARPLWALVCR']	17
1400837	1400905	68	['MSLYRSKETRASRGFFR']	17
1635655	1635816	161	['MCRDRATQQPRRFMLRS']	17
1819907	1820001	94	['MSRVTANLKEARGKMLA']	17
2032419	2032512	93	['MRMGIAYNVADVEPDND']	17
2059570	2059677	107	['MFAGMPAPSGTPAHQAL']	17
2474097	2474197	100	['MAMAACSRASPPLHRSV']	17
2547151	2547229	78	['MSLRFFLNLPQRPCSAQ']	17
3376468	3376570	102	['MVSARLNRPGFRGGSTL']	17
4120928	4120997	69	['MAHVPTRGCPQQRPRST']	17
4564099	4564325	226	['MPDCRPCNSDNYSSQTF']	17
5344429	5344546	117	['MPASSRVNPPTRQRVTS']	17
5425322	5425436	114	['MDNLRERRCSPLLARPA']	17
5692531	5692644	113	['MPASSRARPLPRVAHRR']	17
5832169	5832334	165	['MGQNGDRHGRQATGSVS']	17
262965	263085	120	['MGAGLPAIGPVQVRRMN']	17
278699	278818	119	['MRHVCRSGLPAKSPAPL']	17
309867	309936	69	['MCTTRVTRRISVASHAS']	17
400290	400373	83	['MVSPSLPLASSNTTVAL']	17
445037	445102	65	['MSDRTHSRASPLLQGRR']	17
506998	507120	122	['MNPLLRRPRMPEACATL']	17
571585	571650	65	['MLPIKKPRATALGFWSC']	17
1387777	1387892	115	['MLHLKAASYLWGLACRR']	17
1818647	1818750	103	['MLSSNSDSSGTFLPGSA']	17
1859080	1859171	91	['MSNPKMLRHCRSIVGLP']	17

1883797	1884053	256	['MIPSPALRCASTNHMTD']	17
1981111	1981221	110	['MAPASRVNPLLQRASPA']	17
2817108	2817184	76	['MAQKVAQNSSETENHQA']	17
3595722	3595822	100	['MGDLEVKVLYTPGKGKC']	17
3663941	3664008	67	['MRFYLLESSRWHYPPVR']	17
3984043	3984165	122	['MAYSRVNPLLRGGRVNT']	17
4296833	4296928	95	['MKQLPSYTTARGREFPA']	17
4469117	4469201	84	['MLGAPSPSRQFRYLTTR']	17
4741784	4741862	78	['MSGYYPGTFVPFCSRES']	17
5312951	5313016	65	['MLYDSPPATRDRSEPSV']	17
5333192	5333280	88	['MPALSRACPLPQVAHRP']	17
5678078	5678147	69	['MSCQKDLMSLRWQGART']	17
5750020	5750106	86	['MAAVFAIALPTAGPAEP']	17
5908053	5908121	68	['MRQLWAGFFGPEPLKGN']	17
5975827	5975926	99	['MIYAVVPRPFHRVSAGG']	17
6070644	6070765	121	['MKKSRRDQVVASGHHWQ']	17
6147870	6147963	93	['MQQGPSPEELDEILGRR']	17
6159232	6159306	74	['MLHYRLSGNGLTVCADG']	17
6168654	6168736	82	['MEGGLCYPGTKAAVVSM']	17
400269	400412	143	['MPRWCSKKPTAARARP']	16
413985	414093	108	['MAYSRASPLPQVLHRP']	16
532227	532335	108	['MVELEAVSKQLSFRFD']	16
537472	537610	138	['MHDFLAAPAGRSQTSR']	16
748814	748867	53	['MSVTTRSVQSMEMFSY']	16
761841	761920	79	['MEYCQVKVVASSWVAI']	16
1336500	1336616	116	['METIIPPSEIMVSPAH']	16
1395314	1395381	67	['MWSYRWHAPGALAIRH']	16
1583447	1583555	108	['MLRIVRECFPFVLRGY']	16
1666920	1666997	77	['MCVSPLAPDKPRLYWA']	16
1831862	1831991	129	['MTELAARVGETAGRLA']	16
1968051	1968115	64	['MRPIRRQGCPGWRIRG']	16
1999696	1999894	198	['MQLTETNLGADKSRGP']	16
2252744	2252857	113	['MTVRASSLLQKVTRSL']	16
2263919	2263990	71	['MEWRAGSPGGRSTIVS']	16
2603669	2603748	79	['MPLSLTTDPRGLARSR']	16
2717609	2717762	153	['MVLIDCKGGCFAALRG']	16
2809508	2809585	77	['MYWLIPTHVTFGRGHH']	16
2855002	2855070	68	['MELNLGPRPARKEARG']	16
2897639	2897743	104	['MPRLGHCVDCAASSRL']	16
3262831	3262911	80	['MNPLLRRLVPNHRWTT']	16
3350907	3351027	120	['MHGPDEWRARLGRLAH']	16

3506567	3506682	115	['MGRHCAGRASRIFDKA']	16
3537004	3537069	65	['MRPLPIEFYYGVGSLT']	16
4166130	4166215	85	['MVIRGSCCCLGRPIRG']	16
4350015	4350155	140	['MPPITTVAIIEISQCR']	16
4543501	4543565	64	['MCAPYHLGRFWPETGA']	16
4987509	4987585	76	['MAKCRYLRVSLLPGEQ']	16
5075643	5075757	114	['MGIPPLQISRPTRWDC']	16
5117658	5117751	93	['MTVISLDCAHPVITGL']	16
5574135	5574210	75	['MPFTALEASRVNPLLQ']	16
5644551	5644644	93	['MLGAVIGAAARPNNRP', 'MLARAWRRYWGRCAPQ']	16
5665800	5665854	54	['MAPALPVFAGTPAPTG']	16
6104548	6104625	77	['MRDLRQENDVLATQGR']	16
86717	86784	67	['MAFSRACPLPQGLATP']	16
537476	537542	66	['MGVLQLVIGTGLTGHQ']	16
564417	564498	81	['MTSKITGTNSLHDLHQ']	16
589458	589578	120	['MQAKPLLIWVDRRLKE']	16
1123868	1124019	151	['MKVLGSIDIFVFEPAQ']	16
1567443	1567503	60	['MGHSQAIARGHPVQGR']	16
1627182	1627263	81	['MNSKERCISHRSGLQP']	16
1944533	1944605	72	['MGRAPTCLEAIADCMK']	16
1967989	1968135	146	['MRQPGQPWRLIGRIRG']	16
2067273	2067341	68	['MLWSRVQPKLYPCSIE']	16
2068059	2068171	112	['MPQPVHTRTLPRDRLV']	16
2624219	2624353	134	['MFRGRIPIRTMSKTRL']	16
3537399	3537464	65	['MAATPCKCWSYGLLGI']	16
4022598	4022666	68	['MTRRKGGRVTMRHLRE']	16
4121836	4121993	157	['MQALTVCPSADLSTGR']	16
4123442	4123510	68	['MYCLDAQACCVGCRAR']	16
4129228	4129282	54	['MKHVVNLWVYPKKGLH']	16
4189966	4190028	62	['MHAARVVQRAVQGHNL']	16
4295384	4295452	68	['MFGPAAVPWPTRATGC']	16
4909732	4909831	99	['MTRRKGGKVTMRPLRE']	16
5072934	5072999	65	['MRRPGFLPLAPLLSFV']	16
5095891	5095978	87	['MVQTAEALRFAPFFIL']	16
5097020	5097126	106	['MCIKARCTCRRTRLCQ']	16
5179434	5179513	79	['MPATGAENAVAGTCWP']	16
5267235	5267331	96	['MLKPNAEGVPTWQKLY']	16
5338682	5338770	88	['MHCVKECSAKNRRMAA']	16
5366556	5366663	107	['MNIMRGTAASCAGLRG']	16
5401392	5401496	104	['MQRISVLVVNRLGLGC']	16
5449038	5449122	84	['MELNSRITLEWQWRKH']	16

5481961	5482024	63	['MIPALDDWIKMDLDKI']	16
5496788	5496897	109	['MQHVQPAVILWERASP']	16
5571378	5571438	60	['MPASRARPLPQGWRML']	16
5591881	5591983	102	['MRQRLLERGETNRRAE']	16
5907967	5908041	74	['MRAGFACAAALSLPRH']	16
5933976	5934034	58	['MLGYKVFILDIAPSGA']	16
5955783	5955976	193	['MPHRLSRSCPTGTAWP']	16
6093342	6093408	66	['MCRTHPHLTTRIPCRS']	16
4560	4706	146	['MSLCREKLNPVPGVP']	15
101091	101153	62	['MPIGRLPAAVGVDCN']	15
136905	137011	106	['MRCEEASWVDAVVST']	15
144218	144306	88	['MSQGYLARPDFRGAM']	15
458779	458880	101	['MWERTCPRTPAKPVP']	15
484519	484615	96	['MTRTPVYRLSPFWMA']	15
532306	532370	64	['MKKFEKTLDTRTADK']	15
624953	625022	69	['MTGFVRKEAVCRMLV', 'MYEKRLFVGCWCRTV']	15
733252	733379	127	['MNEKKWLTSKTKNVE']	15
1159154	1159287	133	['MTPDLWERACPRSGA']	15
1278264	1278383	119	['MCRTDTKDCAMSYGQ']	15
1296627	1296791	164	['MSAHPRRRVIRTVSA']	15
1302554	1302645	91	['MLGSIGYRTLRCRAT']	15
1582720	1582811	91	['MPPERSPHALRVRAR']	15
1920458	1920551	93	['MPVPWSRANTAKGLR']	15
1959085	1959202	117	['MPLEHIATTAQNALL']	15
1976172	1976335	163	['MGLPCSPSRACPLPQ']	15
2074080	2074188	108	['MANGLFAGASQTSGP']	15
2531995	2532080	85	['MSSSRLTPLVAFSGE']	15
2624235	2624389	154	['MGILPRNIFAFGNVP']	15
2837164	2837230	66	['MVEWWHEKLRSFRTW']	15
2954568	2954726	158	['MMLPHPAGTRNHPLG']	15
3074258	3074336	78	['MSNRGSTVEPLCAIE']	15
3516157	3516275	118	['MVEHWTSDGEHYHFR']	15
3866921	3867083	162	['MVAATTRSPPGVHTA']	15
4041822	4041906	84	['MPCTTSGQKPRACVG']	15
4292769	4292833	64	['MPRGTQIKDTHLYRI']	15
4501185	4501293	108	['MSDAETIRELRKRPG']	15
4691987	4692125	138	['MKSPARYHANRIKQL']	15
4919517	4919585	68	['MRPTLSASHVQGVPS']	15
5176653	5176706	53	['MQPASSSRLLSWVTR']	15
5660209	5660304	95	['MLRLGLGDEHQRNIT']	15
5697290	5697366	76	['MKGLFHADSPAQGPY']	15

5773052	5773259	207	['MIPNRYVAKNRQVAC']	15
6100710	6100793	83	['MPQPACASRDLAPDR']	15
341394	341469	75	['MDHADCGYRARQGMT']	15
343549	343653	104	['MAGSRREATQGPRSE']	15
386670	386738	68	['MLFALALSRVNPLLQ']	15
401675	401775	100	['MDSPTLSMYGLAHTV']	15
753046	753203	157	['MTGIEPGREDMVVGE']	15
869923	870129	206	['MHRLSCLLFLALHTA']	15
891549	891625	76	['MGAGLPAPTEPQSGP']	15
1039394	1039505	111	['MLVGYVERLFECSRL']	15
1168554	1168724	170	['MMRAFFVGLDWPFRG']	15
1386459	1386569	110	['MLYIFADIKRKSCPI']	15
1480145	1480292	147	['MGAALCGPIATQAPL']	15
1552853	1552918	65	['MQVVQRNNDYSEIQE']	15
1567285	1567356	71	['MFFLFGRLADRCERH']	15
1633953	1634028	75	['MAVVRRLMCGESPDP']	15
1994930	1995035	105	['MAQALMDDAELGMMC']	15
2539051	2539139	88	['MSQPVCRILRQLACW']	15
2885878	2886010	132	['MRELGSHSGKRYAEQ']	15
2939075	2939239	164	['MSAHPRRRVIRTVSA']	15
3088553	3088628	75	['MSGSRSSEENPPLLR']	15
3210775	3210835	60	['MGLAQWQAGLFPDAR']	15
3342367	3342444	77	['MLCASRLRPWQRGGR']	15
3741932	3742012	80	['MTPGPSDGYTAMSES']	15
3756876	3756938	62	['MEKREMSAYRAGIQR']	15
3826070	3826173	103	['MRFNISCLRARPTAC']	15
3827997	3828161	164	['MSAHPRRRVIRTVSA']	15
4075435	4075598	163	['MSAHPRRRVIRTVSA']	15
4156790	4156893	103	['MDRRVCFARQPSVSS']	15
4302121	4302200	79	['MIECAQLRAEVAETG']	15
4303018	4303151	133	['MEIQPKSATSYGRYL']	15
4379676	4379741	65	['MRAFSSGGGRCDSSP']	15
4557482	4557576	94	['MVLPFVAEPSGGMLD']	15
5041436	5041500	64	['MKAEVMGQFCQGDYI']	15
5110807	5110893	86	['MRRVLPVTLLRTAHS']	15
5212103	5212184	81	['MRIAVFAQACDGGPG']	15
5382394	5382505	111	['MGSTGLFAGKPAPTG']	15
5628572	5628635	63	['MAAPALRVPPISSQA']	15
5655545	5655613	68	['MPANATPQIPQADDN']	15
5989810	5989911	101	['MSFSGSNWIRNPIDG']	15
6096698	6096799	101	['MNTRLASNVWSRADR']	15

6134321	6134430	109	['MAYRRPHGAVFLWAV']	15
6147418	6147563	145	['MDGLRSLRTESSRGR']	15
6148435	6148648	213	['MRSRNNSALRKSTEK']	15