Genetic Variation of *Renibacterium salmoninarum* Genes in Infected Salmonids

by
Jeffrey R. Burnett

A THESIS

submitted to

Oregon State University

Bioresource Research

in partial fulfillment of
the requirements for the
degree of

Baccalaureate of Science in Bioresource Research,
Biotechnology

Presented May 27, 2008
Commencement June 2008

Baccalaureate of Science in Bioresource Research, Biotechnology
Thesis of Jeffrey R. Burnett
Presented on May 27, 2008

APPROVED:

Dr. Daniel D. Rockey, Mentor

Dr. Ling Jin, Committee Member

Katharine G. Field, Director of Bioresource Research

I understand that my project will become part of the permanent collection of Oregon State University, Bioresource Research. My signature below authorizes release of my project to any reader upon request.

Jeffrey R. Burnett, Author

Abstract

Renibacterium salmoninarum is the causative agent of bacterial kidney disease in both wild and farmed salmonid species worldwide. The genome of this pathogen has significant synteny to the ubiquitous, soil-dwelling *Arthrobacter spp*. though it is 1.9 Mb smaller, suggesting that reductive evolution has occurred. Recently, our group finished sequencing and annotation of the *R. salmoninarum* American Tissue Culture Collection 33209 strain, which has been cultured in vitro since 1974. The genome contained a significant amount of inactivated genes caused by frameshift and extrachromosomal elements. This led us to hypothesize that reduced evolutionary pressure due to lengthy laboratory cultivation on the genome may account for the mutations found. To test this hypothesis we analyzed four genes with different mutations typicalof mutations observed throughout the genome using samples acquired from around the world. We used basic molecular biology to elucidate the sequences and compared. We found that although the ATCC 33209 strain had been cultured extensively in vitro, its sequence is representative of what is found in nature. The mutations leading to inactivation of regions of the genome suggests the pathogen *R. salmoninarum* has experienced reductive evolution.

Introduction

Renibacterium salmoninarum causes bacterial kidney disease (BKD) in both wild and farmed fish populations worldwide. This Gram-positive, rod shaped, facultatively intracellular pathogen with a genome size of 3.2 megabase pairs utilizes both horizontal and vertical *in ovo* transmission methods within salmonid species (1). Much of the BKD research conducted to date has used the American Tissue Culture Collection (ATCC)

strain 33209, isolated in 1974 from a yearling Chinook salmon at Leaburg Hatchery, Oregon, USA (2).

Over the past three years, our laboratory has worked to complete the genome sequence of this pathogen. Analysis of the completed genome suggests that *R. salmoninarum* evolved from an ancestral *Arthrobacter* species, a ubiquitous, soil-dwelling group of organisms, based on 16S ribosomal RNA data (3). The transition from environmental bacterium to fish pathogen was associated with significant genomic reduction to shed many genes that were not needed in the fish host. A preliminary analysis of the genome demonstrated that not only is there evidence of significant gene elimination, but the pathogen also seems to be in the process of purging genes that are apparently unnecessary to the organism. Upon further examination, we noted that many of the inactive genes were affected by either frameshifts or expansion of a single family of insertion sequence (IS) elements, which led us to hypothesize that the genome of the ATCC strain may have experienced deleterious effects caused by extensive culture in vitro. It appeared to us that many of the genes that were inactivated were being eliminated due to lack of evolutionary pressure in culture. The purpose of this work was to investigate the integrity of the lab culture strain and the possible role of reduced evolutionary pressure on the strains used in sequencing, verification of concurrence of genomic data with clinical isolates collected worldwide, and analyze the reductive evolution of *R. salmoninarum* from its nearest ancestral genus, *Arthobacter spp*.

Materials and Methods

R. salmoninarum strains and kidney samples.

ATCC33209 was originally isolated from a yearling Chinook salmon at Leaburg Hatchery, Oregon, USA in 1974. It was acquired through the American Tissue Culture Collection (Manassas, Virginia) (4). mt239 is an attenuated strain that was isolated from an atlantic salmon (*Salmo salar)* in Scotland. mt239 is unable to agglutinate red blood cells (5). Isolate 684 was provided to us by O.B. Dale of the National Veterinary Institute, in Oslo, Norway. 684 was originally isolated from a clinically diseased brown trout (*Salmo trutta*) obtained from a hatchery in Aurland, Sognefjord, Norway. 684 has an increase in p57 production that causes increased virulence within its host (6). Clinical samples A and B consist of purified DNA from kidney tissue collected from two naturally infected brood stock from Upper Columbia River spring Chinook salmon (*Onchorhyncus tschawytscha*) being reared at the Bonneville Fish Hatchery, Bonneville, OR. All DNA isolation of fish kidneys was performed by Gina Capri of the Rockey Laboratory using a QIAgen total DNA extraction kit (Netherlands). According to manufacturer's instructions.

Gene selection.

Identification of genes of interest occurred using ERGO integrated genomics (Chicago, IL) software. Analysis of open reading frame (ORF) shifts was performed and four ORFs were selected because of representation of general genome mutation pattern, and relevance to normal bacterial function. These genes encoded for: 1. Fibronectin Binding Protein (FBP), a cell adhesion molecule involved in pathogen function within the host; 2. Tetracycline Resistance Protein P (TetP), a tetracycline efflux pump; 3. Dipeptide

Permease Protein (dpp), a protein that functions in peptide import; 4. Citrate Synthase, an enzyme involved in the first step of the Citric Acid Cycle.

Primer Design.

DNA amplification primers (Table 1) were designed using Macvector™ (Cary, NC) computer software. The primers were designed to generate a product of approximately 500 base pairs, flanking the apparent frameshift. The ATCC33209 genome sequence was used to select appropriate primers, and all of the primers were experimentally tested for fragment size. PCR fragments were created using platinum *pfx* polymerase (Invitrogen) because of its high replication fidelity. Standard PCR conditions were optimized for our particular genes. A standard 5 minute first step at 94°C was employed, followed by 30 cycles at 94°C for 45 seconds, 52°C for 45 seconds, and 72°C for 60 seconds. The final elongation step was performed at 72°C for five minutes. PCR reactions were separated by electrophoresis through a 1% agarose gel at 100 volts. Correctly sized fragments were isolated from the electrophoretic gel using a Qiagen gel extraction kit.

Plasmid construction.

PCR fragments were inserted into the Invitrogen Zero Blunt® TOPO® expression vector via blunt end ligation in a five-minute ligation reaction (Invitrogen, Carlsbad, CA). This plasmid was used specifically for its ease of ligation, as well as its ampicillin resistance gene and M13 primer sites within its polylinker region. M13 primer sites were used for subsequent DNA sequencing.

Chemically Competent Cell Preparation.

Escherichia coli (E.coli) DH5-α cells were grown overnight in an incubator/ shaker at 37°C in Luria Bertoni (LB) broth. 100μL aliquots were taken from overnight culture and

added to 10 mL fresh LB medium. The culture was grown to a concentration of 2.5 x 10^8 cells. Cells were pelleted, washed with transformation buffer, and then repelleted using a Beckman Coulter J2-21 refrigerated centrifuge at 6000 rpm. The pellet was then incubated on ice for 2+ hours.

Transformation.

Using 100μL of chemically competent *Escherichia coli* DH5-α cells per reaction, 1 – 2μL of plasmid ligation containing the PCR fragment was added and allowed to incubate on ice for 30 minutes. The solution was placed in the 42°C water bath for exactly 30 seconds, returned to the ice, inoculated with 400μL Super Optimal Catabolite Repression (SOC) media, and gently stirred with a pipette tip. Tubes were placed in 37°C shaker for one hour. After one hour, the solution was centrifuged at 2000 rpm for two minutes and, subsequently, 400μL of supernatant was removed. The resuspended pellet solution was plated on LB^{+amp} agar plates, incubated at 37°C overnight and then stored at 4°C until needed.

PCR screens and nucleotide sequence analysis.

Using the plates made via transformation, colonies were selected, streaked onto LB^{+amp} agar plates and incubated at 37°C overnight. The next day, we collected biomass from each streaked colony with a sterile loop and extracted the plasmid using the QIAprep Spin Miniprep kit (Qiagen). Preliminary restriction with EcoRI was performed to ensure fragment was in fact present in the plasmid, and extracted DNA was sent for sequencing. We sequenced bidirectionally using both forward and reverse M13 primers to confirm the resulting sequence information. All sequencing was accomplished at the Center for

Genome Research and Biocomputing (CGRB) at Oregon State University. Sequence alignments for analysis were performed using Macvector® software.

Generation of Disrupted ORF table.

Using both Macvector™ and ERGO software from Integrated Genomics (Chicago, IL), we tabulated data gathered from annotations by the Insitute for Genomic Research (TIGR) and ERGO. ORFs were called based on sequence similarity with closest homologs. Once genes were identified, they were analyzed for mutations. Nonsense mutations, such as frameshifts, were categorized, numbered and then divided by the total number of ORFs of that particular category.

Results

Renibacterium salmoninarum ATCC 33209 had a single circular chromosome with 3,155,250 base pairs and a G+C content of 56.3%. There were 3,507 open reading frames (ORFs). Of the 3,507 open reading frames, there were 730 inactivated by frameshift, or approximately 20.8% of the total number of ORFs. Notable concentrations of inactivated ORFs were found within the categories of membrane transporters (43%) and general metabolic pathways, including fatty acid/phospholipids (31%), central intermediary (28.2%), and energy metabolism (28.3%). In contrast, those gene categories that seemed to be the most intact were those that participate in nucleotide synthesis, protein synthesis, and mobile and extrachromosomal elements ORFs. Each of these had below 10% of total ORFs inactivated (see table 2).

The four genes selected for analysis are shown in Figure 2. These four genes were selected to represent the apparent trends within the overall genome. dppB (Figure 2A) is

severely degenerate, possibly as a result of the insertion sequence IS994, found within the cluster of genes (7). Citrate synthase is affected by a number of mutations that led to generation of three fragments and inactivation of the gene (Figure 2B), whereas inactivation of tetP occurred by a single point mutation or deletion effectively splitting the gene into two ORFs (Figure 2C). The fibronectin binding protein gene (Figure 2D) is curious, as the gene is intact up until the sortase signal, LPxTG, which commonly enables the resultant translated product to be anchored into the Gram positive cell wall. The signal motif is now separate from the gene and thus, even if the mRNA is transcribed and the protein translated, the pathogen would be unable to utilize it on the cell wall surface.

We amplified the regions shown with the bars (figure 2) from the sample strains of *R. salmoninarum.* We obtained correctly sized PCR products, as well as correctly sized fragment size from transformation and the resultant plasmid + inserts were sequenced. Sequences were compared against the ATCC 33209 sequence already published.

Eighteen of the twenty samples sequenced were identical to the ATCC 33209 strain (Table 3). All twenty of the samples sequenced represented the ATCC sequenced data, with two samples, 684 in the fibronectin binding protein gene and clinical isolate B in the dppD/F, reflecting further mutation at single base positions, neither of these changes would create a native intact open reading frame for the gene in question.

Discussion

Recently, our lab group finished sequencing and annotating the complete *R. salmoninarum* ATCC33209 genome (Appendix A). The genome was found to include extensive missense and nonsense mutations. We hypothesize that the accumulation of so many mutations leading to frameshifts (Table 2) may have resulted from extensive culture in vitro. The suggestion was made that reduced evolutionary pressure during cultivation caused the mutations in the genome because there was no need for the pathogen to retain genes unnecessary for survival in the laboratory. However, sequence comparisons between four unique strains and the cultivated strain revealed that for the four genes, sequences were conserved throughout all of the isolates. The two mutated genes observed were identical to ATCC 33209, with the addition of further point mutations. This supports the conclusion that the mutations within the ATCC strain are representative of what is found in nature. It also confirms that the mutations found within the ATCC strain are not a result of reduced evolutionary pressure due to in vitro cell culture.

As is evident in Table 2, there is reductive evolution occurring in *R. salmoninarum,* away from its presumed ancestor *Arthrobacter spp.* One significant trend is that the genes coding for proteins that the pathogen can aquire from its host are inactivated in much larger quantity than those that are involved in DNA synthesis or other essential cellular functions. Also, as a free-living, soil-dwelling species, *Arthrobacter spp.* probably requires multiple copies of the same gene to be able to synthesize needed molecules, in contrast to the situation of a pathogen in a host / pathogen relationship. The gene categories with less frameshifts are also genes that would be theoretically activated more,

and those with many frameshifts. One example of highly active genes is Mobile and extrachromosonal elements, especially IS994. It is possible that accumulation of IS994 was a recent and explosive event, as this element is nearly completely conserved in a high percentage of the eighty accumulated copies. IS 994 lends to the idea that activated gene products have a lower percentage of inactivation due to frameshift.

Acknowledgements

This project was funded by grant USDA-CREES 2004-35600-14173, entitled Whole Genome Sequencing of *Renibacterium salmoninarum*. Research funding was provided in part by the Howard Hughes Medical Institute Summer Undergraduate Research Fellowship at Oregon State University. Special thanks to Dr. Kevin Ahern, HHMI program director at OSU for assistance in my preliminary project through the summer of 2006. Special thanks to Gina Capri for DNA preparation and to Sara Weeks for guidance and mentoring throughout my training. Thanks to Brendan Jeffrey and Kelsey Sandoz for intriguing discussion and feedback on project objectives. Thanks most of all to Dr. Dan Rockey for excellent tutelage, as well as infusing in me a passion for science that has shaped who I have become and what I want to do as a profession.

References

1 **Balfry, S. K., L. J. Albright, and T. P. T. Evelyn.** 1996. Horizontal transfer of *Renibacterium salmoninarum* among farmed salmonids via the fecal-oral route. Diseases of Aquatic Organisms **25:**63-69.

2 **Sanders, J. E., and J. L. Fryer.** 1980. *Renibacterium salmoninarum* gen. nov., sp. nov., the causative agent of bacterial kidney disease in salmonid fishes. International Journal of Systematic Bacteriology **30:**496-502.

3 **Grayson, T.H., Alexander, S.M., Cooper, L.F., Gilpin, M.L.** 2000. *Renibacterium salmoninarum* isolates from different sources possess two highly conserved copies of the rRNA operon. Antonie van Leeuwenhoek **78:**51-56

4 **Sanders JE, Fryer JL.** 1980. Renibacterium salmoninarum gen. nov., sp. nov., the causative agent of bacterial kidney disease in salmonid fishes. International Journal of Systematic Bacteriology **30**:496-502.

5 **Senson,P.R., Stevenson,R.M.** 1999. Production of the 57 kDa major surface antigen by a non-agglutinating strain of the fish pathogen *Renibacterium salmoninarum*. Diseases of Aquatic Organisms **38**:23–31.

6 **Wiens, G.D., Pascho, R., Winton, J.R.** 2002. A Single Ala^{139}-to-Glu Substitution in the *Renibacterium salmoninarum* Virulence-Associated Protein p57 Results in Antigenic Variation and is Associated with Enhanced p57 Binding to Chinook Salmon Leukocytes. Applied and Environmental Microbiology **68**:3969-3977.

7 **Rhodes, L. D., T. H. Grayson, S. M. Alexander, and M. S. Strom.** 2000. Description and characterization of IS*994*, a putative IS*3* family insertion

sequence from the salmon pathogen, *Renibacterium salmoninarum*. Gene **244:**97-107.

dppD/F (F)	5’ GGC.CGC.GAT.GTG.TCG.ATG.GTT.TT
dppD/F (R)	5’ GCC.CAA.GTG.CGG.CAC.TGC.AGC
Citrate Synthase (F)	5’ GCG.CGG.AGA.GAA.GTT.CTT.GTG
Citrate Synthase (R)	5’ CGA.TGC.GGT.GCG.ACG.TTT.T
tetP (F)	5’ GCC.TAG.CGA.CGC.AAA.AG
tetP (R)	5’ ATA.GTG.ACT.AAG.CAA.TCG.GTG
FBP (F)	5’ CTG.ACG.CCA.ACG.GTA.AAT.ACA.CC
FBP (R)	5’ GGC.GGA.TTC.TCA.ACA.CTC.ACG

Table 1. Primers used for amplification of four fragmented ORFs from clinically infected salmon.

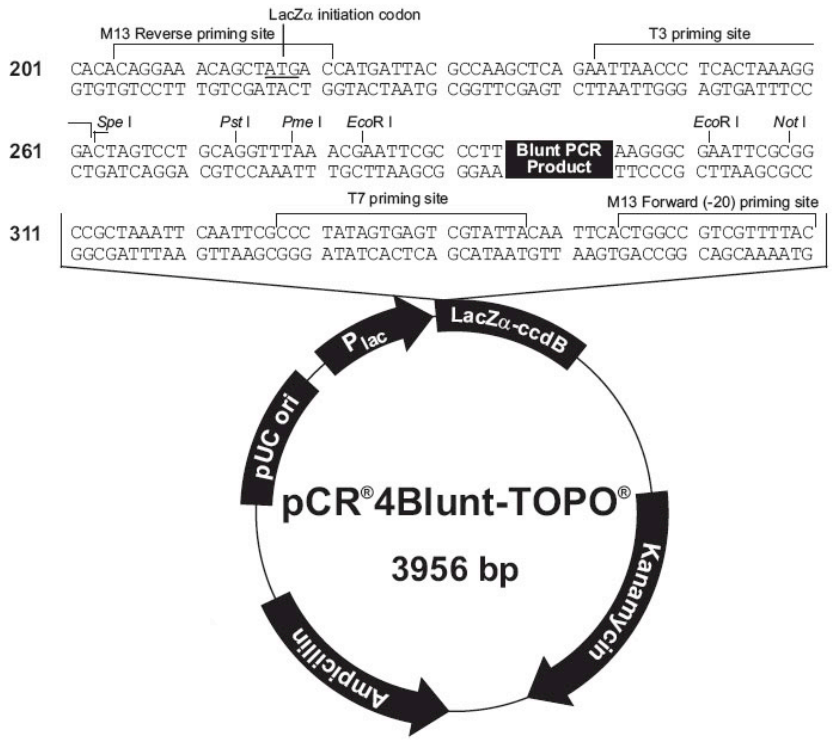

Figure 1. Map of plasmid used in cloning of PCR products (Invitrogen, Carlsbad, CA).

Category	Total	Partial	Percent
Transport/Binding proteins	346	152	43.9
Fatty acid/phospholipids metabolism	100	31	31.0
Central intermediary metabolism	170	48	28.2
Energy metabolism	304	86	28.3
Regulatory Functions	240	53	22.1
DNA metabolism	108	23	21.3
Cellular processes	90	18	20.0
Biosynthesis of cofactors/prosthetic gps.	109	24	22.0
Protein fate	152	28	18.4
Amino Acid Biosynthesis	111	18	16.2
Cell envelope	215	27	12.6
Signal transduction	6	1	16.7
Hypothetical/Unclassified/Unknown Fx	1182	198	16.8
Nucleotides	61	6	9.8
Protein Synthesis	128	9	7.0
Transcription	34	2	5.9
Mobile and extrachromosomal elements	151	6	4.0
Total ORFs	3507	730	20.8

Table 2. Disrupted *R. salmoninarum* ORFs: category, number, and percentage.

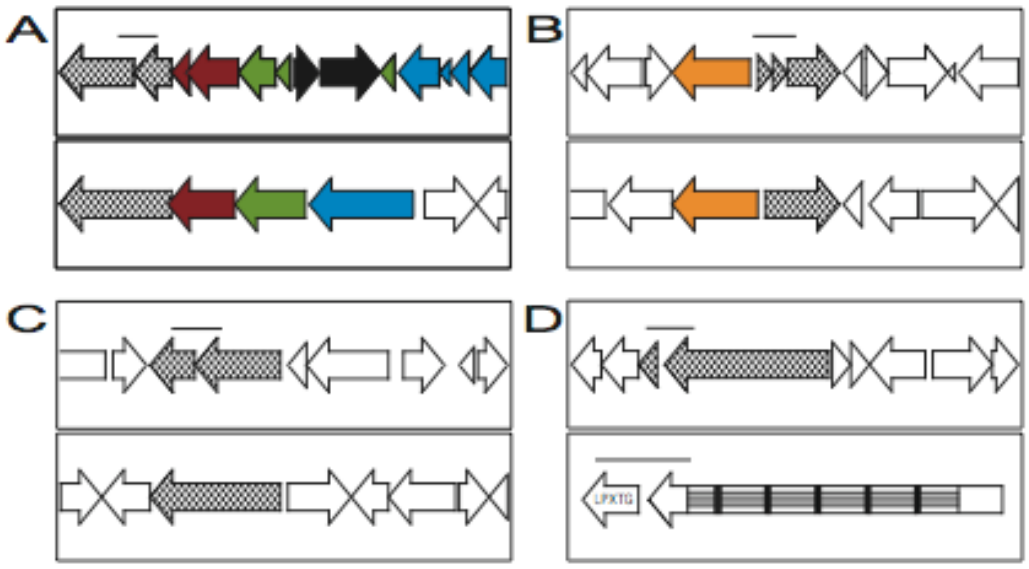

Figure 2. Four mutated gene regions from *R. salmoninarum* selected for analysis (upper panels), compared to intact homologs (lower panels), except for panel D which is an amplification of the gene of interest in *R. salmoninarum.* Panel A is dipeptide permease D/F, panel B is citrate synthase, panel C is tetracycline resistance protein P gene, and panel D is fibronectin binding protein gene, with emphasis on the sortase signal motif LPxTG separated from the upstream region of the gene by frameshift. Homologs are from *Arthrobacter spp.* (panels A,C) and *Streptomyces coelicolor* (panel B). Checkered arrows represent genes of interest. Black arrows represent IS994. The bar shows the 500 base pair region that we amplified and sequenced in this study.

	ATCC	mt239	684	A	B
dppD/F	Y	Y	Y	**N**	Y
Citrate synthase	Y	Y	Y	Y	Y
Tetracycline Resistance P	Y	**N**	Y	Y	Y
Fibronectin Binding Protein	Y	Y	Y	Y	Y

Table 3. Sequencing results of cloned PCR fragments. N stands for different than, and Y stands for identical to, *R. salmoninarum* ATCC33209 strain.

Appendix A

Below includes the paper to which my research is a part. It is included for reference to publish data only. It was submitted to the Journal of Bacteriology (March 2008).

This page intentionally left blank.

The genome sequence of the fish pathogen *Renibacterium salmoninarum* suggests reductive evolution away from an environmental *Arthrobacter* ancestor

Gregory D. Wiens[1¶], Daniel D. Rockey[2¶], Zaining Wu[3], Jean Chang[3], Ruth Levy[3], Samuel Crane[4], Donald S. Chen[4], Gina R. Capri[2], **Jeffrey R. Burnett**[2], Ponnerassery S. Sudheesh[4], Matthew J. Schipma[5], Henry Burd[5], Anamitra Bhattacharyya[5], Linda D. Rhodes[4], Rajinder Kaul[3], and Mark S. Strom[4]*

[1]USDA-ARS-National Center for Cool and Coldwater Aquaculture, Kearneysville, WV 25430

[2]Oregon State University, Corvallis,OR, 97331

[3]Division of Medical Genetics, Department of Medicine, and University of Washington Genome Center, University of Washington, Seattle, WA, 98195

[4]Northwest Fisheries Science Center, NOAA Fisheries Service, Seattle, WA, 98112

[5]Integrated Genomics, Inc., Chicago, IL, 60612

Short Title: Complete *R. salmoninarum* genome sequence

Key words: fish pathogen, reductive evolution, virulence

[¶]GDW and DDR contributed equally to the research and manuscript and are considered co-first authors

*Corresponding author

2725 Montlake Blvd. E., Seattle, WA 98112

mark.strom@noaa.gov

ABSTRACT

Renibacterium salmoninarum (*Rs*) is the causative agent of bacterial kidney disease and a significant threat to the healthy and sustainable production of salmonid fish worldwide. The pathogen is difficult to culture *in vitro*, genetic manipulation is challenging, and current therapies and preventative strategies are only marginally effective in preventing disease. The complete genome of *R. salmoninarum* ATCC 33209 was sequenced and shown to be a circular chromosome of 3,155,250 bp that is predicted to encode 3,507 open-reading frames (ORFs). There are a total of 80 copies of three different insertion sequence (IS) elements interspersed throughout the genome. We estimate that 21% of the predicted ORFs have been inactivated via frameshifts, point mutations, insertion sequences and putative deletions. The genome has extended regions of synteny to *Arthrobacter sp*. FB24 and *A. aurescens* TC1, but is approximately 1.9 Mb smaller and has a lower G+C content, suggesting significant genome reduction has occurred since divergence from the last common ancestor. A limited set of candidate virulence factors appear to have been acquired via horizontal transmission after divergence of the species; these include capsular polysaccharides, heme sequestration molecules, and the major secreted cell surface antigen p57 (also known as major soluble antigen). Examination of the genome reveals a number of ORFs homologous to antibiotic resistance genes, including β-lactamases, efflux proteins, macrolide glycosyltransferases, and ribosomal RNA methyltransferases. The genome sequence has provided new insights into *Rs* evolution and will facilitate identification of chemotherapeutic targets and vaccine candidates for future utility in prevention and treatment of infections in cultured salmonids.

INTRODUCTION

Infection of salmon and trout by *Renibacterium salmoninarum* (*Rs*) causes Bacterial Kidney Disease (BKD), a progressive, multifocal granulomatous disease that involves all internal organs, and in particular, the kidney (22, 27, 69). This Gram-positive bacterium causes morbidity and mortality in both farmed and wild fish in nearly all regions of the world where salmonids are found (68). Despite being recognized as a major fish pathogen for over 70 years, mechanisms of transmission, pathogenesis and immune evasion remain poorly understood. The disease is chronic in nature and mortality most often occurs in 6–12 month old juvenile salmonids and prespawning adults (18). The spread of BKD has followed the rapid expansion of salmonid culture (22), and to date, most recorded outbreaks of BKD have occurred in fish culture facilities with losses as high as 80% in stocks of Pacific salmon and 40% in stocks of Atlantic salmon (*Salmo salar*) (22). The chronic nature of the disease has hindered accurate estimates of fish losses in feral fish populations. In the Pacific Northwest of the United States, the prevalence of *Rs* in juvenile Chinook salmon can vary between 60 and 100% in different populations although pathogen burden is usually low in the majority of these fish (17, 54). BKD is also the greatest cause of infectious disease-related mortality in several endangered species restoration and conservation programs (26, 36).

Rs is one of the few vertebrate bacterial pathogens known to be vertically transmitted by intra-ovum infection during egg maturation (7, 20, 21). The pathogen can also be transmitted horizontally from infected fish sharing a water supply (1, 41). Prophylaxis or prevention of *Rs* infection is challenging, in part due to the ability of the bacterium to survive phagocytosis and possibly replicate within macrophages (2, 34, 64,

75). For Pacific salmon, there are no efficacious vaccines, and antibiotic therapy never fully eliminates the pathogen (16, 57); however, for Atlantic salmon, which have greater innate resistance, live *Arthrobacter davidanieli* (commercially known as Renogen) immunization can provide significant cross-protection (33, 59). The ability of *A. davidanieli* to protect against BKD is presumably a function of the overall genetic relationship between the genus *Arthrobacter* and *Rs*. Both are high G+C Gram- positive members of the Micrcoccaceae (27).

There are many challenges in studying the biology and pathogenesis of *Rs*. The bacterium grows extremely slowly *in vitro*, with up to six weeks required for primary isolation (27). While genetic tools for study are available, the process to introduce or inactivate specific genes is arduous and time consuming (53). The availability and analysis of the *Rs* genomic sequence is expected to significantly contribute to our understanding of this pathogen and provide opportunities to create more directed genetic tools for understanding its biological and pathogenic mechanisms. We have therefore recently completed and analyzed the *Rs* genome sequence in order to pave the way for better understanding and management of this fish pathogen and its devastating effects on salmonids in culture.

MATERIALS AND METHODS

Clone isolation. *Renibacterium salmoninarum* ATCC 33209 was originally isolated from a yearling chinook salmon (*Oncorhynchus tshawytscha*) at a salmon hatchery in Western Oregon (60), and although others have reported this strain to be modestly attenuated in comparison to other *Rs* isolates (40, 52, 62), challenge experiments carried out prior to

extraction of DNA for sequencing showed that it had not lost virulence due to laboratory passage (Supplementary Figure S1). A single colony of the pathogen was isolated and cultured in KDM-II medium (19) with 10% bovine serum supplementation for production of DNA. Genomic DNA was prepared as described previously and the DNA used for construction of the small and large insert libraries for sequencing (72).

Genome Sequencing. Whole genome shotgun sequencing was used for assembling the *Rs* ATCC 33209 strain genome, as described (35, 73). Randomly picked small insert plasmids (average 2.6 Kbp insert size), cloned into the blunt-ended pUC19 vector, were sequenced at both ends using universal forward and reverse sequencing primers according to standard protocols established at UWGC. In all, 52,183 reads were performed, providing 9.7 X sequence coverage with an average Q20 of 590 bases/read. A *Sau*3A-partial fosmid library was also constructed in the pCCFOS vector digested with *Bam*HI according to the standard protocol (50). A total of 768 independent fosmid clones were picked and paired-end-sequenced, with 91.8 % of the clones giving successful mate-pair end-sequences. BigDye terminator chemistry and capillary DNA sequencers (Model 3700, Applied Biosystems) were used for generating the sequence data (35, 73). The sequence data was assembled and visualized using Phred/Phrap/Consed software tools (23, 24, 29). The sequence quality and contiguity was improved by carrying out six rounds of experiments designed by the Autofinish tool in Consed (30). The Autofinish designed experiments primarily utilized small insert clones to improve the sequence quality and/or close some gaps. Following six rounds of Autofinish, a manual finishing utilized both small insert as well fosmid clones. This

involved 1) use of specialized sequencing chemistries to sequence difficult regions, 2) PCR amplification and sequencing of specific targeted regions, 3) transposon mutagenesis of the small insert clones followed by sequencing to fix misassembled or difficult to assemble regions, and 4) shotgun sequencing of the targeted fosmid clones to fix long-range misassemblies in the assembled genome. In all, 8,968 finishing reads were attempted during the course of this project. In addition, 52 small insert clones were mutagenized by transposon mutagenesis and independent clones picked to generate consensus sequence for these clones. We also identified 29 fosmid clones that spanned misassembled or large gap regions, which were sequenced and assembled independently. The consensus sequences from transposon mutagenized, and the fosmid clones were used as backbones in the main *Rs* genome assembly to resolve misassembled regions. The final *Rs* 33209 genome assembly was oriented so that the genome coordinates start at the origin of replication (ori) as position 1, as defined by the GC skew, characteristic nucleotide composition, and presence of *dnaA* immediately following the ori.

Genome assembly validation. The final genome assembly was validated by two independent methods. The gross-scale long range validity of the genome assembly was established by complete agreement between the virtual fingerprint pattern of the genome in *Pme*I, *Pac*I, or *Swa*I restriction enzyme domains and the experimentally determined restriction fragment sizes determined by pulse-field gel electrophoresis (data not shown). For 1 Kb and larger scale validation, fingerprint data was generated from the paired-end-sequences of 768 fosmid clones by digesting with three independent restriction enzymes, *Bgl*II, *Hind*III and *Pst*I; generating 2,256 or 97.9% of the theoretical 2,304 fingerprints

from these clones. The fosmid paired-end-sequence and experimentally derived fingerprint data was used for assembly validation by comparison with the virtual fingerprint patterns from the assembled genome using the SeqTile software tools developed for this purpose in our Genome Center (35, 73). The fosmid paired-end reads anchor the clone to unique position in the genome, while the fingerprint data is required to compare experimentally derived fingerprints with the sequence derived virtual patterns. The assembly validation at 1 Kb resolution scale was thus achieved by comparing fosmid-based fingerprints in *Bgl*II, *Hind*III and *Pst*I restriction enzyme domains with the virtual patterns from the finished genome assembly. A complete correspondence between the virtual and experimentally derived fingerprint pattern of the genome in the three restriction enzyme domains of *Bgl*II, *Hind*III and *Pst*I was observed; thus validating the genome assembly. The genome sequence was deposited into GenBank as accession number NC_010168.

Phylogenetic trees. 16S rRNA sequences were obtained from the NCBI taxonomy browser (http://www.ncbi.nlm.nih.gov) search for Actinobacteria (high G+C Gram-positive bacteria) with known genome sequences. 16S rRNA sequences were compared using a predetermined alignment from the NAST Alignment Tool (12). A manually inspected, multiple alignment was used as input for the generation of phylogenetic trees, using the program package MEGA (v 3.1) (38). Both Neighbor Joining and Maximum Parsimony methods, run using default settings in the MEGA program, were employed to obtain character-based trees. Both methods yielded similar clusters and arrangements of taxa. The genome sequences of *Arthrobacter* sp. BF A20 and *Arthrobacter davidanieli*

have not been determined but the 16s RNA sequences were included for comparative purposes.

Open reading frame identification and annotation. ORF calling was performed as described (48). Two rounds of automated annotation were initially performed utilizing the ERGO Genome Analysis Suite (Integrated Genomics, Chicago, IL) during genome assembly and closure (48). The final finished genome sequence was also subjected to automated ORF calling and annotation on 8/23/06 using The Institute for Genomics Research (TIGR) automated annotation software (67). Subsequently, each ORF was manually examined by three teams of annotators with final annotation responsibility for ORFs as follows: locus tags RSal33209_0001 through RSal33209_1207, the Northwest Fisheries Science Center; locus tags RSal33209_1208 through RSal33209_2411, the NCCCWA; and locus tags RSal33209_2412 through RSal33209_3558, Oregon State University. During the annotation process, a large number of partial genes were noted. Most of these ORFs were also flagged by ERGO and TIGR ORF-calling software as possible frameshifts and point mutations. The sequence data for each candidate frameshift was manually inspected to rule out sequencing or assembly errors. These ORFs were also compared against putative orthologues in *Arthrobacter sp*. FB24 and *A. aurescens* TC1 genomes. A final examination and the designation of candidate partial ORFs was completed by one annotator (D.D.R.) and the protein coding regions were designated as either "partial", "partial, N-terminus" or "partial C-terminus". The partial ORFs were then categorized using the TIGR Cellular Main Role list, as shown in column T of Supplemental Data I, tab 1.

Sequencing of selected *Rs* genes from different sources. DNA from laboratory strains MT239 and 684 were prepared as previously described (72). DNA was also prepared from kidney tissue of naturally infected brood stock from Upper Columbia River Spring Chinook Salmon (*Onchorhyncus tschawytscha*) being reared at the Bonneville Fish Hatchery, Bonneville, OR. These fish had recently died of *Rs* infection and their kidneys were heavily infected with the pathogen. Each of these DNAs was used as template for PCR amplifications of selected genes, using *Pfx* DNA polymerase and oligonucleotide primers shown in Table 1. Products of these amplifications were cloned into the pCR-Blunt II-TOPO vector according to the manufacturer's directions (Invitrogen, CA). Plasmids were purified and the nucleotide sequence of the cloned insert was determined in both orientations. Each of the sequences was compared to the corresponding gene in the completed ATCC 33209 genome sequence using ClustalW in the MacVector DNA analysis program (MacVector, Inc., Cary, NC). Any gene sequences that differed from those in the 33209 genome were reamplified, recloned, and resequenced for confirmation.

Comparative genomics of *Rs* with *A. aurescens* TC1 and *A. sp.* FB24. The ERGO Workbench was used to determine common and unique Role/IDs (RID) between *Rs*, *A. aurescens* TC1 and *A. sp.* FB24. Clusters of similar RIDs were generated using Build Clusters function using a protein similarity threshold of $1.0e^{-10}$ and subset clusters were extracted by organism sets containing either individual, paired combinations, or all three genomes.

RESULTS

General genome features. The final *Rs* ATCC 33209 genome assembled as a single circular chromosome of 3,155,250 bases (Table 2) and included 51,799 shotgun reads in the final genome assembly (Genbank accession number NC_010168). Cumulative GC skew analysis identified a transition in the G+C content in proximity to *dnaA* (Figure 1). There are 46 tRNAs and two rRNA operons located at 364,813-370362 bp and 674,875-6804 bp in relationship to the putative origin of replication. The G+C content is 56.3% in concordance with previous measurements (3). This value is 6% to 9% lower than either sequenced `species (43) (Table 2). The ATCC 33209 strain lacks any associated plasmids or apparent integrated phage. The genome was predicted to code for 2777 CDS and manual annotation identified 730 "partial" ORFs that are putative pseudogenes. Of the partial gene sequences, 360 were disrupted by frameshifts, 208 by point mutations, and 162 by uncharacterized changes including insertion sequences and putative deletions (Figure 1 and Supplementary Data Table 1, tab 2). A determination of the precise number of deleted genes is complicated by the large number of hypothetical genes that appear to be disrupted, and therefore, our number should serve as an estimate of total disrupted genes. In our analysis, the categories with the highest percentage of disrupted ORFs were in membrane transporters, where 43% of ORFs are identified as partials, and fatty acid/phospholipids metabolism, where 31% of ORFs are partials (Table 3). In contrast, the lowest number of partial ORFs are in protein synthesis, transcription, and mobile genetic elements, with each class being below 8%.

Analysis of selected partial genes in clinical *Rs*. The high rate of partial ORFs within the sequenced *Rs* genome suggests the possibility that the strain used for sequence analysis had undergone mutagenesis during its extended culture in the laboratory. To examine this, genomic DNA from the reference ATCC 33209 strain, from two different clinical isolates, and from infected kidney tissue of naturally infected fish in Oregon, USA, were used as templates to amplify DNA surrounding four selected regions that appear to contain frameshifts in the *Rs*-33209 genome sequence (Table 1 and Figure 2). These genomic regions represent the types of gene inactivations that are spread thoughout the chromosome. The targeted chromosomal regions included: 1) a dipeptide permease operon (RSal33209_1659 through RSal33209_1671) that appears severely degenerate, perhaps as a function of integration of an IS*994* element into the center of the operon (RSal33209_0105 and 0106); 2) a citrate synthase gene (RSal33209_2899, 2900, and 2901) that contains 2 apparent frameshifts relative to an homolog in *Streptomyces spp*.; 3) *tetP* (RSal33209_1873 and 1874), a gene encoding an efflux pump with an apparent frameshift in its coding region; and 4) a putative fibronectin binding protein (RSal33209_1552) that is separated via a mutation from a short ORF containing a Gram positive anchor domain (RSal33209_1551, Figure 2). A 500 bp fragment of each target gene, with sequences that might be mutated near the center, was amplified and sequenced for each of the strains. Of the 20 total examined sequences, in all cases, the frameshifts found in the completed genome were also found in the amplified DNA samples, and only 2 sequences had unique single base changes relative to the ATCC 33209 genome. These data provide support for the conclusion that the observed frameshifts are highly

conserved in clinical (wild type) strains of the pathogen and did not accumulate during laboratory culture of the ATCC strain.

Genome evolution. In order to determine the origin of *Rs* and better understand the evolutionary process shaping the genome, phylogenetic and whole genome alignments were performed. Analyses of the 16S rRNA sequence with reference to the high G+C bacteria as well as phylogenetic analysis of common genes identifies *Arthrobacter spp*. as the closest known relative of *Rs* (Figure 3a and Supplementary Figure S2). Two *Arthrobacter* genomes, TC1 and FB24, have been recently published or released (43). The *Rs* genome is 1.44 Mb smaller than the chromosome of TC1 and 1.55 Mb smaller than the main chromosome of FB24. In addition, these two strains of *Arthrobacter* have several large plasmids that are not present in the 33209 strain nor are sequences with high similarity to *Arthrobacter* plasmids represented in the *Rs* chromosome. A whole genome, nucleotide alignment using Mauve (Multiple Genome Alignment Tool, http://gel.ahabs.wisc.edu/mauve/) identified large blocks of synteny between the two *Arthrobacter* genomes, and in comparison, rearrangements within the *Rs* genome (Supplementary Figure S3).

The *Rs* chromosome and the two genome sequences from the *Arthrobacter spp.* share 1562 protein clusters (defined has having $>1x10e^{-10}$ threshold similarity) that include a combined total of 7336 ORFs consistent with the phylogenetic grouping of *Rs* within the *Arthrobacter* clade (Figure 3B; Supplementary Data Table 2). Approximately equal numbers of ORFs are part of the core set of similar genes (*Rs* ORFs=2273; *A. sp* FB24 ORFs=2507; and *A. aurescens* ORFs=2556). The two *Arthrobacter* species share

740 protein clusters (1917 ORFs) not found in *Rs* that may have been lost in the course of genome reduction. Similar numbers of unique ORF clusters were identified in each of the three microorganisms ranging from 818 to 933 clusters suggesting similar levels of genomic divergence. In comparison to all sequenced microorganisms, a total of 240 *Rs* ORFs were without function or similarity and are thus unique to *Rs* (Table 2). These may represent novel potential diagnostic targets (Supplementary Data Table 1). In comparison to the chromosome of the *A. aurescens* TC1 genome, *Rs* possesses a relatively lower percentage of ORFs in energy metabolism (8.86% vs 17.87%), cellular processes (2.65% vs 4.5%), transcription (1.11% vs 2.15%) while containing a high percentage of ORFs in central intermediary metabolism (5.13% vs 2.83%) and mobile and extrachromosomal element function (4.48% vs 0.85%) (Table 4).

The genome of *Rs* contains three recognized insertion sequences: IS*994* (69 total copies), 10 IS*Rs2* (10 copies) and IS*Rs3* (one copy). The genome of the related *A. aurescens* TC1 contains a total of 46 ORFs encoding functions consistent with transposons or insertion sequence elements, 23 of which are on the main chromosome. None of the IS elements from either sequenced *Arthrobacter* genomes are represented in the *Rs* genome. As originally described by Rhodes et al. (55), the IS*994* element consists of two open reading frames, labeled *orfA* and *orfB*, which encode candidate transposases of approximately 120 and 360 amino acids, respectively. The genome sequence revealed that there are 69 and 67 copies of IS*994 orfA* and *orfB*, respectively. This insertion sequence is distributed relatively randomly through the genome in both orientations. There is a single stand-alone *orfB* and there are 2 copies of *orfA* that are separated from *orfB*. There are three examples where *orfB* is interrupted by a common frameshift, but the

open reading frames in most copies of IS*994* are generally highly conserved and intact. Rhodes et al (55) previously demonstrated that each of the seven copies of IS*994* sequenced contained 24 nucleotide inverted repeats of at each end of the element. The same 24 nucleotide inverted repeat is common to all intact copies of IS*994* in the genome.

IS*Rs2* is a 1343 base pair IS element with a conserved 17 base pair inverted repeat at each end, and with a single open reading frame of 1071 base pairs encoding a protein product of 357 amino acids. The predicted protein is most similar to transposases from *Rhodococcus* and *Gordonia spp*. The sequence of IS*Rs2* is highly conserved with variation at only six nucleotide positions resulting in only two amino acid substitutions among the nine full-size copies. A tenth copy of IS*Rs*2 is truncated, encoding only the carboxy terminal 139 amino acids of the predicted protein.

The third *Rs* IS element, IS*Rs3*, is represented in the chromosome as a single candidate transposase inactivated by a central frameshift. This interrupted sequence encodes a short IS*200*-like transposase with identity to transposases in *Streptomyces spp*. and *Mycobacterium spp*., among others. IS*Rs*3 is located between *msa*1 and a copy of IS*994*(RSal33209_0135 and RSal33209_0136).

Information processing. Despite the large numbers of open-reading frames that are truncated in this organism, house-keeping genes such as large and small ribosomal protein subunits are all present and intact. *Rs* also possesses most of the aminoacyl tRNA synthetases with the exception of those for arginine and glutamine. Likewise, *Rs* bears a full complement of factors for translation initiation (*e.g.* IF-1, -2 and -3), translation elongation (*e.g.* EF-G, EF-TU, EF-P and EF-Ts) and peptide release factors (*e.g.* RF-1

and RF-2). The chromosomal DNA replication machinery is also highly homologous to that seen in the sequenced *Arthrobacter spp*. genomes, including a DNA polymerase with αβδδ'γ/τε subunits.

DNA modification and repair. The *Rs* genome appears to have a very poor representation of DNA restriction and modification enzymes. No classical type I, II or III restriction-modification systems exist in the genome. *Rs* lacks any ORFs encoding a DNA polymerase V function, while the *Arthrobacter* spp. genomes encode both subunits of this protein. DNA polymerase V (homologous to *E. coli* UmuDC) is involved in the bacterial SOS response. DNA polymerase V and RecA are involved in translesion synthesis (TLS) to avoid DNA replication blocking lesions. Since TLS is highly inaccurate, SOS induction leads to mutagenic replication by-pass by DNA polymerase V. Absence of DNA polymerase V in the *Rs* genome indicates the bacterium does not perform TLS in a RecA/Pol V dependent manner.

Central carbohydrate metabolism. The *Rs* genome encodes most of the enzymes participating in core central metabolic pathways, including: glycolysis, pentose phosphate, tricarboxylic acid (or Kreb's) cycle, pyruvate cycle and anaplerotic reactions. *Rs* is likely able to utilize several sugars and polyols. For instance, the organism is predicted to import and utilize glucose, fructose, arabinose, gluconate, glycerol and citrate.

Amino acid metabolism. *Rs* is able to synthesize de novo the following amino acids: histidine, aspartate, lysine, threonine, valine, leucine, isoleucine, D and L-alanine, glutamate, glutamine, arginine, tyrosine, phenylanalanine and tryptophan. Conversely, *Rs* is unable to synthesize serine, glycine, cysteine, asparagine and methionine *de novo*. The inability to synthesize these amino acids is manifested frequently by the presence of truncated open-reading frames (ORFs) encoding putative enzymatic functions. For instance, in the 3-step phosphorylated serine biosynthetic pathway that converts 3-phosphoglycerate to L-serine, the phosphoserine transaminase function is likely absent since the ORFs encoding this function are truncated (e.g. Rsal33209_0497, Rsal33209_0496). Since the biosynthesis of cysteine is dependent upon serine availability, this might arguably limit intracellular de novo cysteine synthesis, despite *Rs* having an intact pathway from serine to cysteine. Other routes for cysteine production either from sulfate or homocysteine are similarly hindered. However, the Rs genome appears to possess an ABC-type alkylphosphonate transport system (e.g. RSal33209_1533, RSal33209_1531, RSal33209_1530, RSal33209_1529) that is likely able to transport external O-phosphoserine, that is present in fish tissue (37) into the bacterial cell. The latter can then be transformed into serine using the canonical serine biosynthetic pathway using phosphoserine phosphatase (e.g. RSal33209_0497, RSal33209_0496). In the case of tryptophan, the caveat is that although phosphoribosylanthranilate synthase is absent, it is likely that another function substitutes for this missing enzyme (4). Interestingly, inspection of the *Rs* genome identifies a tryptophan-2,3-dioxygenase (RSal33209_2215) an enzyme that transforms tryptophan into N-formyl-L-kynurenine. Some intracellular pathogens (e.g. *Lawsonia*) are able to

utilize tryptophan breakdown products via N-formyl kynurenine to produce anthranilate that can be utilized in the classic tryptophan biosynthetic pathway. However, the absence of an arylformamidase encoded in the *Rs* genome suggests the bacterium is unable to degrade this compound to L-kynurenine, potentially leading to an accumulation of this compound.

Cofactor biosynthesis. *Rs* appears to have a complete set of genes encoding enzymes for biosynthesis of riboflavin, FAD, FMN, pyridoxine, pantothenate (including 4'-phosphopantetheine), coenzyme A, lipoate and menoquinone. In addition, genes for the biosynthesis and re-cycling of folate, essential for purine and thymidylate metabolism, are present in the genome. Moreover, the machinery to synthesize hemes a, d and o are also encoded in the genome.

Fatty acid and lipid metabolism. The absence of discernible homologs of the *fabA* and *fabZ* genes in the *Rs* genome suggests that the organism is incapable of synthesizing both unsaturated and saturated fatty acids and, therefore, has to scavenge these compounds from the host. Moreover, the loss of the phospholipid biosynthetic genes suggests that *Rs* is dependent upon the host for acquiring lipid building blocks and thus resembles the situation in rickettsial pathogens. Although *Rs* lacks any ORFs encoding phospholipid degradative capabilities, *Rs* does encode a triacylglycerol lipase, which hydrolyzes the carboxyl ester bonds in mono-, di- and tri-glycerides to liberate fatty acids and alcohols. This lipase is likely essential for maintaining the lipolytic activity and is responsible for management of fatty acid availability of the bacterium. In addition, *Rs* possesses genes

encoding enzymes capable of performing *beta*-oxidation, suggesting that the organism can metabolize fatty acids and use them as carbon and nitrogen sources.

Bioenergetics. *Rs* is capable of micro-aerophilic growth and respiration. The organism appears to be able to utilize a variety of carbon substrates for energy including: pyruvate, lactate, succinate, malate, glycerol-3-phosphate, proline, butanoyl-CoA and fatty acids. The latter is particularly noteworthy since the organism likely scavenges host-derived fatty acids to complement its fatty acid biosynthetic deficiencies. *Rs* possesses genes responsible for menaquinone biosynthesis (*e.g.* RSal33209_2700-RSal33209_2702; RSal33209_2638), which is the electron carrier for the *Rs* electron transport chain, with the electron acceptor prior to oxygen being a cytochrome c oxidase. The presence of the full complement of enzymes in the tricarboxylic acid (TCA) cycle and glycolytic components suggests that ATP generation in *Rs* takes place via substrate-level phosphorylation. The *Rs* genome also bears a full complement of ORFs encoding the proton-transporting F_1F_0 ATP synthase complex.

Rs is not capable of anaerobic respiration. The organism does possess related but truncated enzymatic systems for anaerobiosis suggesting the existence of an anaerobic respiratory chain in its evolutionary past.

Absence of urease. The genomes of both *Arthrobacter* strains have ORFs clusters encoding the nickel metalloenzyme urease (α [*ureC*], β [*ureA*] and γ [*ureD*] subunits) as well as urease accessory protein genes (*ureE*, *ureF*, *ureG* and *ureD* and high affinity nickel transporter). These ORFs are involved in gastric colonization by *Helicobacter* spp. and utilization of urea as a nitrogen source in a variety of environmental

microorganisms (10, 42). *Rs* lacks all of these ORFs and thus appears to be unable to utilize urea during colonization of the salmonid kidney.

***Rs* resuscitation promoting factors.** Resuscitation promoting factors are family of proteins that are common in high G+C bacteria and involved in bacterial re-activation of growth from a dormant state. *Rs* contains two genes (RSal33209_2532 and RSal33209_2995) were identified that each contains an RPF domain. These genes encode predicted proteins of 220 amino acids and 384 amino acids, respectively. Both proteins have predicted signal peptide sequences suggesting that they are exported. The predicted protein product of *Rs* ORF RSal33209_2532 has high sequence similarity to the original Rpf protein discovered from *Micrococcus luteus* (44) while the predicted protein of *Rs* ORF RSal33209_2995 is most similar to *Mycobacterium tuberculosis* RpfB (45). In addition to the RPF domains, *Rs* Rpf contains a peptidoglycan binding motif, LysM, while *Rs* RpfB contains three domains of unknown function (DUF 348) as well as a single G5 domain identical to *M. tuberculosis* RpfB.

Secondary metabolite synthesis or sequestration. The *Rs* genome carries one fragmented cluster of genes that may be responsible for siderophore synthesis (RSal33209_0770 - RSal33209_0772). This set of genes is highly conserved between the *Arthobacter spp*. and *Rs*, and the genes reside in a highly syntenous region of the chromosome. *Rs* also encodes multiple operons associated with sequestration of siderophores, including ferric siderophore binding proteins and permeases associated with transport of siderophores across the cytosolic membrane (51). There are three

clusters of these genes. One cluster (ORFs RSal33209_2862 through 2859) is present in both *Arthrobacter* species and is degenerate in *Rs*. A second cluster (RSal33209_1684 through 1861) is present in only *A. aurescens* TC1 and contains intact reading frames in *Rs*. The final gene cluster is largely degenerate but contains an intact *fepG* homolog (RSal33209_3347). Therefore, *Rs* encodes enzymes comprising an entire ferric siderophore import machinery, but no single operon encodes all components, and the process may be complicated by inactivations in certain genes.

Two very large ORFs (RSal33209_2539, RSal33209_2540) appear to contain modules for polyketide synthesis machines, and products of adjacent ORFs may also participate in the process. These ORFs are not found in either *Arthrobacter spp*. While it is not possible to predict an end product of this pathway, this set of genes may be involved in synthesis of unique secondary metabolites that protect the pathogen from competing microorganisms.

St**ress Response Genes.** *Rs* possesses several ORFs encoding enzymes that confer resistance to oxygen radicals. The encoded enzymes comprise superoxide dismutase, catalase, peroxidases and thioredoxin peroxidase. These enzymes constitute a bacterial defense mechanism against bacteriocidal radical agents generated by host neutrophils and macrophages. Some of these enzymes may also afford resistance to nitrogen compounds and radicals produced by the host. One aspect of nitrogen intermediate resistance involves the reduction of reactive nitrogen intermediates to less toxic products. *Rs* encodes one enzyme (ferridoxin nitrite reductase, RSal33209_2621) possibly involved in reduction of nitrite produced by macrophages (*e.g.*). Interestingly, this protein is

homologous to an enzyme found in many plants including *A. thaliana*. *Rs* also encodes a signal-transducing histidine kinase, nitrate/nitrite sensor protein (*narX*; RSal33209_0978), suggesting that it is capable of sensing and destroying nitric oxide produced by the host macrophages. The products of these genes may function in resistance to host defenses, but homologs are all present within the *Arthrobacter* genomes (Supplementary Data Table 1). While genes involved in the stress response within a host are conserved, ORFs putatively involved in the response to environmental stressors are not found. Overall, of the 78 stress-related ORFs identified in *A. aurescens* TC1, only 36 (46%) are present in *Rs*. Interestingly, *Rs* appears to be particularly deficient in cupin-related genes with only 2 of 15 ORFs identified in *Arthrobacter* found. Cupin-related genes are involved in Mn accumulation and morphogenesis (43) but appears to be dispensable for *Rs* host survival.

Heme acquisition. *Rs* has acquired genes encoding several proteins involved in heme and hemin acquisition. There are two different gene clusters that share identity with heme acquisition loci of different bacterial species. ORFs RSal33209_2547 through 2552 encode proteins involved in these processes in several other systems. One of these, heme oxygenase (Rsal33209_26547), may also be important in stress response, as homologs in mammalian cells are upregulated under stress (8). ORFs RSal33209_2063 through 2061 also encode putative hemin binding and transport proteins, with highest similarity to similar genes in *Streptomyces coelicor*. Neither of these *Rs* gene clusters share identity with ORFs present in *Arthrobacter*, and therefore appear to have been acquired after species divergence. *Arthrobacter spp.* have ORFs with similarity to these

gene products, but they are found in gene clusters associated with ferric-siderophore acquisition as opposed to heme acquisition.

Capsular synthesis. Immunological and microscopic approaches were used to demonstrate that *Rs* produces a capsule and/ or soluble extracellular carbohydrate (15, 61). The nature of this product is not well characterized. The *Rs* genome sequence encodes a set of proteins that form the core of a capsular polysaccharide assembly pathway. These include homologs to at least seven proteins encoded in *Staphylococcus spp*. capsular synthesis genetic clusters (47). However, in contrast to *Staphylococcus*, the *Rs* capsular synthesis genes are not present at a single locus. For example, the ORFs RSal33209_1611 and RSal33209_1612 are homologous to cap5G and cap5H, but are physically separated from other genes in a putative capsular synthesis pathway. These genes share no significant identity with sequences from *Arthrobacter*, but are adjacent to genes involved with extracellular carbohydrate synthesis and transport (RSal33209_1613 through 1618), which are syntenous with homologs from both *Arthrobacter spp.* (Figure 4).

The regions encoding capsular synthesis proteins suggest multiple mechanisms of evolutionary divergence from the *Arthrobacter* group of bacteria. First, ORFs RSal33209_1436 to 1439 are located within a region of the chromosome that is highly syntenous with both *Arthrobacter* species (Figure 4A). Only one of the candidate capsular synthesis ORFs (Rsal33209_1436), is shared with one of the *Arthrobacter* species (TC1). Each of the other candidate capsular synthesis genes are not found in the sequenced *Arthrobacter spp*., even though each genome has genes for extracellular

polysaccharide synthesis at that locus. It is possible that this region has been a locus of divergence as *Rs* and *Arthrobacter* species have evolved to different niches. A different mechanism is found in ORFs RSal33209_1338 to 1342. This region may represent a genomic island acquired by *Rs* via a horizontal transmission event. These ORFs encode proteins that can be placed in capsular synthesis pathways, have a low G+C content (48-51%) relative to the general *R*s genome (56.3%) , and are flanked by a complete and partial IS*994* insertion element (Figure 4B). There is also evidence that genes encoding possible exopolysaccharide synthesis proteins were lost from the *Rs* chromosome via genomic reduction. Examples include *Arthrobacter* pathways involved in rhamnose and xylose metabolism, and genes of a Wzy-dependent pathway. None of these pathways are present in the *Rs* genome. Therefore, at some point in the evolution away from a progenitor species, *Rs* may have lost pathways associated with environmental exopolysaccharide synthesis, modified existing genes encoding exopolysaccharide synthesis proteins, and gained a pathway associated with protective exopolysaccharide capsular synthesis.

Gene products encoded by RSal33209_1750 and RSal33209_2984 are similar to enzymes involved in UDP-glucose metabolism, and homologs of these gene products participate in capsular synthesis in a diverse collection of bacteria (13, 47). These gene products also share sequence identity with proteins that participate in carbohydrate metabolism in many different organisms, and are represented in the *Arthrobacter* FB24 genome.

Protein secretion machinery and the sortase pathway. *Rs* contains a complete Sec-dependent secretion pathway with the exception of a YajC homolog (protein translocase). Four homologs of signal peptidase I (RSal33209_1086, RSal33209_1087, RSal33209_1255, RSal33209_2979) and one homolog of signal peptidase II (lipoprotein signal peptidase, RSal33209_2485) are present. A SignalP and TmPred search demonstrated the presence of at least 444 ORFs carrying a typical Sec secretion leader peptide (Supplementary Data Table 1).

The *Rs* genome also contains an ORF encoding a sortase, a member of a group of cysteine transpeptidases in Gram-positive bacteria that promote covalent anchoring of proteins to the peptidoglycan surface of the cell envelope (39). The precursor substrates of sortases have a distinct sequence motif, including an amino-terminal Sec-dependent signal peptide, the sortase cleavage site (see below), a carboxy-terminal transmembrane region, and a charged anchor domain at the C-terminus. Analysis of the *Rs* sortase (RSal33209_2896) shows that it is most closely related to the subfamily-5 or class D of bacterial sortases (SrtD), based on recent classification schemes (11, 14). Members of this subfamily recognize a LAxTG motif. While the *Rs* SrtD shares 65% and 62% identity with the sortases in *Arthrobacter aurescens* TC1 and *Arthrobacter* sp. FB24 respectively, most Gram-positive bacteria possess SrtA or subfamily-3 sortases that recognize proteins with motifs LPxTG and LPxTGG, respectively. Using the tripartite pattern search method of Boekhorst, et al. (5) as well as tools such as Pfam, SignalP, TMpred and ProDom, seven ORFs were identified in the *Rs*-33209 genome specifying potential sortase substrates containing the LAxTG motif (RSal33209_2105, RSal33209_3407, RSal33209_2525, RSal33209_3273, RSal33209_0619,

RSal33209_1326, and RSal33209_0402), with some predicted to be cell surface proteinases or adhesins (64). Interference of sortase activity with the anti-infective phenyl vinyl sulfone has demonstrated that bacterial cell wall localization of 8-10 proteins is inhibited (64), although at the present time it has not been confirmed that these proteins are the same as any of the putative sortase substrates. In fact, two of the seven potential SrtD substrates identified through bioinformatics analysis are annotated as pseudogenes (RSal33209_0619, RSal33209_1326), and the current search may have failed to identify additional complete sortase substrates. Of the seven ORFs with SrtD recognition signals, three are syntenous with *Arthrobacter spp*. homologs (RSal33209_3407, RSal33209_1326, RSal33209_0402), and three are unique to *Rs* (RSal33209_2105, RSal33209_2525 and RSal33209_3273). The remaining putative sortase substrate (RSal33209_0619) shares sequence identity with candidate acid phosphatases in Gram-negative bacteria.

In addition to the probable functional SrtD, there is a second ORF containing a sortase catalytic domain. This 123 amino acid ORF (RSal33209_1799) appears to be encoded by a remnant of a *srtA* gene. Two potential SrtA substrates are present in the genome, one complete and syntenous with a hypothetical protein in *Arthrobacter* FB24 (RSal33209_1929), and one unique ORF encoded by a pseudogene (RSal33209_1551). The latter ORF is adjacent to a large ORF encoding a putative fibronectin binding protein, but that lacks a sortase recognition motif (Figure 3). The presence of the degenerate *srtA* gene and at least one degenerate SrtA substrate may represent additional examples of specific genome decay in a pathway no longer needed for the intracellular lifestyle of *Rs*.

Antibiotic resistance genes. A search of the *Rs* genome yielded 68 ORFs homologous to genes encoding factors known to be involved in resistance to different antibiotics. These include a variety of multidrug transporters, several β-lactamases (e.g., RSal33209_0207, RSal33209_0436, RSal33209_0920), tetracycline (*tetA,* RSal33209_0457, and *tetB,* RSal33209_0459), and macrolide resistance factors. Macrolide resistance is an especially important concern since erythromycin (and to a lesser extent azithromycin) is commonly used to control *Rs* infections in both supplementation and conservation hatcheries. Several classes of macrolide resistance genes are represented in the genome, including two 23S rRNA methyltransferases (*rlmA^{II}*, RSal33209_2019 and *spoU,* RSal33209_3401), a macrolide-efflux factor (*mefA*, RSal33209_2224), a multidrug resistance efflux pump (*pvsC*, RSal33209_1691), and a 16S rRNA dimethylase (*ksgA,* RSal33209_2994). The genome also contains an ORF with similarity to macrolide glycosyltransferase (*mgtA*, RSal33209_2959), but the ORF is truncated and likely non-functional. The ORFs encoded by *rlmA^{II}, spoU*, *mefA*, and *ksgA* show high synteny with homologs in *Arthrobacter spp.,* while *pvsC and mgtA* have other ancestral origins.

Major Soluble Antigen. The *Rs* Major Surface Antigen (MSA) is a unique protein that makes up approximately 70% of the bacterial cell surface protein in *Rs* (74)*,* and which is secreted into fish tissues at up to mg per ml concentrations in tissue (66). MSA is found in culture supernatants, fish tissues, and on the cell surface as intact 57kDa proteins as well as defined fragments of the full length molecule (71). This protein has

immunosuppressive properties and is considered a major virulence factor (9). We have previously shown that this gene is duplicated in the *Rs* genome of many strains including 33209 (46, 72), and this was confirmed by genome sequencing. The entire *msa* coding sequence, and over 700 downstream base pairs, are exactly duplicated in each of the two copies. Both copies of *msa* are linked to nearby IS*994* sequences, and, one *msa* copy is adjacent to the single degenerate ISRs*3* sequence. Interestingly, a third copy is present in many clinical isolates, and these strains exhibit increased virulence compared to strains carrying two copies of *msa* (52). No homolog of MSA is found in any other sequenced bacterial genome.

Serine proteinases. As mentioned in the previous section, MSA is targeted by an unknown serine proteinase activity in bacterial culture supernatants. The enzymatic activity is associated with a >100 kDa enzymatically active nonreduced protein species (58). The *Rs* genome encodes several candidate serine proteinases, some of which are conserved between *Rs* and *Arthrobacter* spp. and several that are not found in *Arthrobacter* spp. Included in this group are a pair of subtilisin-like enzymes that are encoded on sequential ORFs (RSal33209_0104/105; RSal33209_0106). One of the pair is interrupted by a frameshift. The serine proteinases that do not have homologs in *Arthrobacter spp.* group into four families, with representatives at ORFs RSal33209_0898, RSal33209_1246, RSal33209_0571 and the subtilisin homolog. Several of these ORFs encode proteins that have secretory signal sequences, and therefore might have a role in the extracellular digestion of MSA. However, none of

these proteins has a predicted molecular weight of above 60 kDa, and are not directly implicated as the proteinase responsible for this activity.

Hemolysins. It was reported earlier that *Rs* encodes at least two specific hemolysins (22, 31, 32). These investigators expressed random *Rs* DNA fragments within *E. coli* and looked for hemolysis by plating on blood agar plates. One of these clones was sequenced and is identified in the genome as ORF RSal33209_3168. As noted by the original authors (31), the encoded product of this gene contains peptidase and proteinase motifs, and may function in digestion of proteins following secretion from *Rs*. There are three additional candidate hemolysins encoded by the genome, including ORFs RSal33209_0811, RSal33209_3195, and RSal33209_3047. Each candidate hemolysin has clear homologs within the *Arthrobacter*, indicating they were not horizontally acquired as an adaptation to the piscine host. It is possible that other reported hemolysins whose sequence is not present in the Genbank database may be encoded by one of these ORFs.

DISCUSSION

The genome sequence of *Rs* ATCC 33209, coupled with comparative genomic analysis of related microorganisms, supports the hypothesis that *Rs* evolved from an *Arthrobacter*-like precursor, largely via genome reduction. While *Rs* is a fastidious obligate pathogen of salmonid fish, members of the genus *Arthrobacter* are soil inhabitants that are noted for their ability to survive many types of environmental stress. However, there is no evidence that *Rs* grows for extended periods outside of the salmonid

host, and no member of the *Arthrobacter* family colonizes any host species. Therefore, this is an interesting model for the analysis of pathogen evolution.

Mechanisms for the evolution of pathogens from commensal or environmental organisms are just being elucidated (49). Stinear et al. (63) noted six genomic features that characterize rapid adaptation of pathogens toward the eukaryotic host environment. This includes: 1) expansion of insertion elements, 2) abundant pseudogenes and smaller genome size, 3) chromosomal rearrangements, 4) acquisition of foreign genes, and 6) a high degree of genetic relatedness or clonality within the species. We have identified each of these features in the *Rs* genome. There are 80 total copies of three insertion elements, including 69 highly conserved copies of the bicistonic element *IS994*. There are also 10 conserved copies of the Is*Rs*2, and a single degenerate copy of the IS200-like IS*Rs*3. There is a high abundance of apparent pseudogenes in *Rs* (n=730), and many pathways are apparently reduced via this strategy. The genome is approximately 1.9 Mbp smaller than the *Arthrobacter* genomes, with a concomitant reduction in functional capability and level of adaptability. Finally, strains of *Rs* are highly homogenous with respect to overall genomic structure biochemical properties and surface antigens (25, 28, 55, 70)

While the major force in *Rs* evolution was likely genome reduction, there are several examples of apparently horizontally acquired genes associated with pathogenesis. First and foremost, the pathogen acquired *msa*1 and *msa*2 via horizontal acquisition and subsequent duplication. In agreement with previous analyses, the *msa* genes are identical (46, 52, 72). MSA (also known as P57) binds leukocytes, mammalian erythrocytes and fish sperm cells and is the predominant protein produced in culture and in infected fish

tissues (65, 66, 69, 71). This protein is also the only conclusively proven virulence factor in the species and no orthologs are found in the two *Arthrobacter* genome sequences or in any other microorganism. One copy of this gene is adjacent to the degenerate IS*Rs*3 and both are adjacent or very near a copy of IS*994*. At present, the original source of *msa* is unknown.

Genes encoding enzymes responsible for capsule synthesis and heme/iron utilization are also present within the genome. While these are not completely encoded within apparent horizontally acquired islands in the genome, there are clear examples of horizontal acquisition of a group of capsular synthesis genes as a genomic island (Figure 4). None of these virulence-associated genes have homologs within *Arthrobacter*. Most of the virulence-associated genes in the pathogen are horizontally acquired; however, the precise mechanism for this in *Rs* is unclear. No plasmids or prophage elements were identified within the *Rs* genome although insertion sequences are abundant. The lack of restriction-modification systems in the *Rs* genome may contribute to the extensive disruption of the open-reading frames manifest in this organism.

The genome contains a number of ORFs encoding likely surface proteins, including eight ORFs that contain sortase cleavage motifs and has one intact sortase. These surface proteins provide bacterial envelopes with both species and strain-specific properties, and include factors involved in virulence and host immune system interactions, such as fibronectin and collagen-binding proteins, invasins, and proteins that make up Gram-positive pili or fimbriae. Three of the seven sortase-cleavage-motif containing ORFs are also present in *Arthrobacter* and half of these appear to be pseudogenes in *Rs*. Recent experiments using sortase inhibitors decrease adherence of *Rs*

to fibronectin and fish cells *in vitro* suggesting a role for these proteins in attachment (64). If these ORFs are confirmed to be directly involved in adherence, this would suggest that only minor modifications of sortase-motif proteins were required for adaptation to binding to salmonid cells.

As might be expected in a microorganism related to bacteria that survive in soil, the *Rs* genome contains many genes encoding antibiotic resistance factors, ranging from multidrug transporters to factors that specifically modify either an antibiotic or its substrate. The number of macrolide resistant factors is particularly interesting in that erythromycin is often the antibiotic of choice in treatment of BKD. While the MIC of erythomycin to *Rs*-33209 is low (0.008 μg/ml) (56), induction of expression of at least two genes (*rlmA*^{II} and *mefA*) upon exposure of *Rs* to erythromycin or azithromycin correlates with true resistance to these two macrolides (Sudheesh and Strom, unpublished). The identification and subsequent characterization of these genes through analysis of the *Rs* genome demonstrates a molecular mechanism for the generalized failure of macrolides to completely clear *Rs* from an infected fish population, despite the observed low MIC. Genome analysis has also led to the discovery of the *Rs* sortase, and a possible alternative for control of BKD through the use of sortase-inhibitory anti-infective drugs (64).

A total of 1562 ORF clusters shared similarity between *Rs* and *Arthrobacter spp*. This close genomic similarity may explain the efficacy of the commercial vaccine Renogen composed of a live, lyophilized *Arthrobacter spp.* that is injected into young fish. However, this efficacy only extends to Atlantic salmon (*Salmo salar*), and the vaccine does not provide significant protection to Pacific salmonids (*Oncorhynchus spp.*)

(57). At present, vaccine coverage against BKD remains below 5% for which statistics are available (6). Identification and characterization of additional cell surface proteins, such as through analysis of proteins processed and anchored to the cell surface via the *Rs* sortase, may lead to more efficacious vaccines that will stimulate long lasting cellular immunity against *Rs.*

The *Rs* genome sequence adds to our understanding of the pathogenesis of infection, but most of the the host-microbe interactions in this system remain poorly understood. The pathogen encodes selected horizontally-acquried virulence factors, including capsular synthesis genes, heme acquisition operons, possible hemolysins, and the poorly characterized *msa* genes. Most of these do not have homologs within the *Arthrobacter spp.*, and thus, may form the basis of the niche differences between these bacteria. While the classic BKD lesion is analogous to the granuloma found in piscine mycobacterial infections, *Rs* does not produce mycolic acids or other mycobacterial surface virulence factors. Approximately 7% of the ORFs in the genome are unique to *Rs*, and it is possible that many of these sequences encode proteins that are critical to the interaction between the bacterium and their host. We continue to use the genome sequence to identify and characterize potential *Rs* therapeutic targets, with a goal of assisting fisheries professionals in the prevention and control of BKD.

ACKNOWLEDGEMENTS

This project was funded by the NSF/USDA Microbial Genome Sequencing Program agreement number 2004-35600-14173 awarded to M.S.S., G.D.W., and D.D.R. G.D.W. was supported by Agricultural Research Service CRIS Project 1930-32000-002 “Host-

Pathogen and Environmental Interactions in Cool and Cold Water Aquaculture". M.S.S. was also supported by the NOAA FCRPS Biological Opinion Implementation Project and the NOAA Fisheries Service. We thank Eric Haugen, Donald Bovee, Karen Phelps, Regina Lim, Will Gillett, Hillary Hayden, Don Guenthener at the University of Washington Sequencing center. Mark Zabriskie of the Oregon State University College of Pharmacy provided valuable analysis of genes involved in secondary metabolism, and Theresa Walunis of Integrated Genomics was instrumental in generation of genome maps and preparation of the genome sequence for Genbank submission. We thank Walton W. Dickhoff for critical review of this manuscript. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture and the U.S. Department of Commerce.

REFERENCES

1. **Balfry, S. K., L. J. Albright, and T. P. T. Evelyn.** 1996. Horizontal transfer of *Renibacterium salmoninarum* among farmed salmonids via the fecal-oral route. Dis Aquatic Org **25:**63-69.

2. **Bandin, I., A. E. Ellis, J. L. Barja, and C. J. Secombes.** 1993. Interaction between the rainbow trout macrophages and *Renibacterium salmoninarum in vitro*. Fish Shellfish Immunol **3:**25-33.

3. **Banner, C. R., J. S. Rohovec, and J. L. Fryer.** 1991. A new value for mol percent guanine + cytosine of DNA for the salmonid fish pathogen *Renibacterium salmoninarum*. FEMS Microbiol Lett **63:**57-9.

4. **Barona-Gomez, F., and D. A. Hodgson.** 2003. Occurrence of a putative ancient-like isomerase involved in histidine and tryptophan biosynthesis. EMBO Rep **4:**296-300.

5. **Boekhorst, J., M. W. de Been, M. Kleerebezem, and R. J. Siezen.** 2005. Genome-wide detection and analysis of cell wall-bound proteins with LPxTG-like sorting motifs. J Bacteriol **187:**4928-34.

6. **Bravo, S., and P. J. Midtlyng.** 2007. The use of fish vaccines in the Chilean salmon industry 1999-2003. Aquaculture **270:**36-42.

7. **Bruno, D. W., and A. L. S. Munro.** 1986. Observations on *Renibacterium salmoninarum* and the salmonid egg. Dis Aquatic Org **1:**83-87.

8. **Choi, A. M., and J. Alam.** 1996. Heme oxygenase-1: function, regulation, and implication of a novel stress-inducible protein in oxidant-induced lung injury. Am J Respir Cell Mol Biol **15:**9-19.

9. **Coady, A. M., A. L. Murray, D. G. Elliott, and L. D. Rhodes.** 2006. Both *msa* genes in *Renibacterium salmoninarum* are needed for full virulence in bacterial kidney disease. Appl Environ Microbiol **72:**2672-8.

10. **Collins, C. M., and S. E. D'Orazio.** 1993. Bacterial ureases: structure, regulation of expression and role in pathogenesis. Mol Microbiol **9:**907-13.

11. **Comfort, D., and R. T. Clubb.** 2004. A comparative genome analysis identifies distinct sorting pathways in gram-positive bacteria. Infect Immun **72:**2710-22.

12. **DeSantis, T. Z., Jr., P. Hugenholtz, K. Keller, E. L. Brodie, N. Larsen, Y. M. Piceno, R. Phan, and G. L. Andersen.** 2006. NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucleic Acids Res **34:**W394-9.

13. **Dillard, J. P., and J. Yother.** 1994. Genetic and molecular characterization of capsular polysaccharide biosynthesis in *Streptococcus pneumoniae* type 3. Mol Microbiol **12:**959-72.

14. **Dramsi, S., P. Trieu-Cuot, and H. Bierne.** 2005. Sorting sortases: a nomenclature proposal for the various sortases of Gram-positive bacteria. Res Microbiol **156:**289-97.

15. **Dubreuil, D., R. Lallier, and M. Jacques.** 1990. Immunoelectron microscopic demonstration that *Renibacterium salmoninarum* is encapsulated. FEMS Microbiol Lett **54:**313-5.

16. **Elliott, D. G., R. J. Pascho, and G. L. Bullock.** 1989. Developments in the control of bacterial kidney disease of salmonid fishes. Dis Aquatic Org **6:**201-215.

17. **Elliott, D. G., R. J. Pascho, L. M. Jackson, G. M. Matthews, and J. R. Harmon.** 1997. *Renibacterium salmoninarum* in Spring-Summer Chinook Salmon smelts at dams on the Columbia and Snake Rivers. J Aquatic Animal Health **9:**114-126.

18. **Evelyn, T.** 1993. Bacterial kidney disease - BKD., p. 177-195. *In* V. Inglis, R. J. Roberts, and N. R. Bromage (ed.), Bacterial-diseases-of-fish. Oxford, UK: Blackwell Scientific Publications.

19. **Evelyn, T. P. T.** 1977. An improved growth medium for the kidney disease bacterium ad some notes on using the medium. Bull Int Epiz **87:**511-513.

20. **Evelyn, T. P. T., J. E. Ketcheson, and L. Prosperi-Porta.** 1984. Further evidence for the presence of *Renibacterium salmoninarum* in salmonid eggs and for the failure of povidone-iodine to reduce the intra-ovum infection rate in water-hardened eggs. J Fish Dis **7:**173-182.

21. **Evelyn, T. P. T., L. Prosperi-Porta, and J. E. Ketcheson.** 1986. Experimental intra-ovum infection of salmonid eggs with *Renibacterium salmoninarum* and vertical transmission of the pathogen with such eggs despite their treatment with erythromycin. Dis Aquatic Org **1:**197-202.

22. **Evenden, A. J., T. H. Grayson, M. L. Gilpin, and C. B. Munn.** 1993. *Renibacterium salmoninarum* and bacterial kidney disease-the unfinished jigsaw. Ann Rev Fish Dis **3:**87-104.

23. **Ewing, B., and P. Green.** 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res **8:**186-94.

24. **Ewing, B., L. Hillier, M. C. Wendl, and P. Green.** 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res **8:**175-85.

25. **Fiedler, F., and R. Draxl.** 1986. Biochemical and immunochemical properties of the cell surface of *Renibacterium salmoninarum*. J Bacteriol **168:**799-804.

26. **Flagg, T. A., C. V. W. Mahnken, and J. J. Hard.** 1995. An Assessment of the Status of Captive Broodstock Technology of Pacific Salmon, 1995 Final Report. DOE/BP-55064-1; R&D Project: 199305600; Other: ON: DE97002404; TRN: TRN: 97:004560.

27. **Fryer, J. L., and J. E. Sanders.** 1981. Bacterial kidney disease of salmonid fish. Ann Rev Microbiol **35:**273-98.

28. **Getchell, R. G., J. S. Rohovec, and J. L. Fryer.** 1985. Comparison of *Renibacterium salmoninarum* isolates by antigenic analysis. Fish Path **20:**149-159.

29. **Gordon, D., C. Abajian, and P. Green.** 1998. Consed: a graphical tool for sequence finishing. Genome Res **8:**195-202.

30. **Gordon, D., C. Desmarais, and P. Green.** 2001. Automated finishing with autofinish. Genome Res **11:**614-25.

31. **Grayson, T. H., A. J. Evenden, M. L. Gilpin, K. L. Martin, and C. B. Munn.** 1995. A gene from *Renibacterium salmoninarum* encoding a product which shows homology to bacterial zinc-metalloproteases. Microbiology **141:**1331-41.

32. **Grayson, T. H., M. L. Gilpin, A. J. Evenden, and C. B. Munn.** 2001. Evidence for the immune recognition of two haemolysins of *Renibacterium salmoninarum* by fish displaying clinical symptoms of bacterial kidney disease (BKD). Fish Shellfish Immunol **11:**367-70.

33. **Griffiths, S. G., K. J. Melville, and K. Salonius.** 1998. Reduction of *Renibacterium salmoninarum* culture activity in Atlantic salmon following vaccination with avirulent strains. Fish Shellfish Immunol **1998:**607-619.

34. **Gutenberger, S. K., R. J. Duimstra, J. S. Rohovec, and J. L. Fryer.** 1997. Intracellular survival of *Renibacterium salmoninarum* in trout mononuclear phagocytes. Dis Aquatic Org **28:**93-106.

35. **Hendrickson, E. L., R. Kaul, Y. Zhou, D. Bovee, P. Chapman, J. Chung, E. Conway de Macario, J. A. Dodsworth, W. Gillett, D. E. Graham, M. Hackett, A. K. Haydock, A. Kang, M. L. Land, R. Levy, T. J. Lie, T. A. Major, B. C. Moore, I. Porat, A. Palmeiri, G. Rouse, C. Saenphimmachak, D. Soll, S. Van Dien, T. Wang, W. B. Whitman, Q. Xia, Y. Zhang, F. W. Larimer, M. V. Olson, and J. A. Leigh.** 2004. Complete genome sequence of the genetically tractable hydrogenotrophic methanogen *Methanococcus maripaludis*. J Bacteriol **186:**6956-69.

36. **Hoffnagle, T. L., R. W. Carmichael, and W. T. Noll.** 2002. Grande Ronde Basin spring Chinook salmon captive broodstock program, 1995-2002 project status report. Oregon Fish and Wildlife Department, La Grande, OR.

37. **Ingles, C. J., and G. H. Dixon.** 1967. Phosphorylation of protamine during spermatogenesis in trout testis. Proc Natl Acad Sci U S A **58:**1011-8.

38. **Kumar, S., K. Tamura, and M. Nei.** 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform **5:**150-63.

39. **Marraffini, L. A., A. C. Dedent, and O. Schneewind.** 2006. Sortases and the art of anchoring proteins to the envelopes of gram-positive bacteria. Microbiol Mol Biol Rev **70:**192-221.

40. **McIntosh, D., E. Flano, T. H. Grayson, M. L. Gilpin, B. Austin, and A. J. Villena.** 1997. Production of putative virulence factors by *Renibacterium salmoninarum* grown in cell culture. Microbiology **143:**3349-56.

41. **Mitchum, D. L., and L. E. Sherman.** 1981. Transmission of bacterial kidney disease from wild to stocked hatchery trout. Canadian J Fish Aquatic Sci **38:**547-551.

42. **Mols, M., and T. Abee.** 2008. Role of ureolytic activity in *Bacillus cereus* nitrogen metabolism and acid survival. Appl Environ Microbiol **74:**2370-8.

43. **Mongodin, E. F., N. Shapir, S. C. Daugherty, R. T. DeBoy, J. B. Emerson, A. Shvartzbeyn, D. Radune, J. Vamathevan, F. Riggs, V. Grinberg, H. Khouri, L. P. Wackett, K. E. Nelson, and M. J. Sadowsky.** 2006. Secrets of soil survival revealed by the genome sequence of *Arthrobacter aurescens* TC1. PLoS Genet **2:**e214.

44. **Mukamolova, G. V., O. A. Turapov, K. Kazarian, M. Telkov, A. S. Kaprelyants, D. B. Kell, and M. Young.** 2002. The *rpf* gene of *Micrococcus luteus* encodes an essential secreted growth factor. Mol Microbiol **46:**611-21.

45. **Mukamolova, G. V., O. A. Turapov, D. I. Young, A. S. Kaprelyants, D. B. Kell, and M. Young.** 2002. A family of autocrine growth factors in *Mycobacterium tuberculosis*. Mol Microbiol **46:**623-35.

46. **O'Farrell, C. L., and M. S. Strom.** 1999. Differential expression of the virulence-associated protein p57 and characterization of its duplicated gene *msa* in virulent and attenuated strains of *Renibacterium salmoninarum*. Dis Aquatic Org **38:**115-23.

47. **O'Riordan, K., and J. C. Lee.** 2004. *Staphylococcus aureus* capsular polysaccharides. Clin Microbiol Rev **17:**218-34.

48. **Overbeek, R., N. Larsen, T. Walunas, M. D'Souza, G. Pusch, E. Selkov, Jr., K. Liolios, V. Joukov, D. Kaznadzey, I. Anderson, A. Bhattacharyya, H. Burd, W. Gardner, P. Hanke, V. Kapatral, N. Mikhailova, O. Vasieva, A. Osterman, V. Vonstein, M. Fonstein, N. Ivanova, and N. Kyrpides.** 2003. The ERGO genome analysis and discovery system. Nucleic Acids Res **31:**164-71.

49. **Pallen, M. J., and B. W. Wren.** 2007. Bacterial pathogenomics. Nature **449:**835-42.

50. **Raymond, C. K., S. Subramanian, M. Paddock, R. Qiu, C. Deodato, A. Palmieri, J. Chang, T. Radke, E. Haugen, A. Kas, D. Waring, D. Bovee, R. Stacy, R. Kaul, and M. V. Olson.** 2005. Targeted, haplotype-resolved resequencing of long segments of the human genome. Genomics **86:**759-66.

51. **Raymond, K. N., E. A. Dertz, and S. S. Kim.** 2003. Enterobactin: an archetype for microbial iron transport. Proc Natl Acad Sci U S A **100:**3584-8.

52. **Rhodes, L. D., A. M. Coady, and R. K. Deinhard.** 2004. Identification of a third *msa* gene in *Renibacterium salmoninarum* and the associated virulence phenotype. Appl Environ Microbiol **70:**6488-94.

53. **Rhodes, L. D., A. M. Coady, and M. S. Strom.** 2002. Expression of duplicate *msa* genes in thesalmonid pathogen *Renibacterium salmoninarum*. Appl Environ Microbiol **68:**5480-7.

54. **Rhodes, L. D., C. Durkin, S. L. Nance, and C. A. Rice.** 2006. Prevalence and analysis of *Renibacterium salmoninarum* infection among juvenile Chinook salmon *Oncorhynchus tshawytscha* in North Puget Sound. Dis Aquat Organ **71:**179-90.

55. **Rhodes, L. D., T. H. Grayson, S. M. Alexander, and M. S. Strom.** 2000. Description and characterization of IS*994*, a putative IS*3* family insertion sequence from the salmon pathogen, *Renibacterium salmoninarum*. Gene **244:**97-107.

56. **Rhodes, L. D., O. T. Nguyen, R. K. Deinhard, T. M. White, L. W. Harrell, and M. C. Roberts.** In Press. Characterization of Renibacterium salmoninarum with reduced susceptibility to macrolide antibiotics by a standardized antibiotic susceptibility test. Dis Aquat Org.

57. **Rhodes, L. D., C. K. Rathbone, S. C. Corbett, L. W. Harrell, and M. S. Strom.** 2004. Efficacy of cellular vaccines and genetic adjuvants against bacterial kidney disease in chinook salmon (*Oncorhynchus tshawytscha*). Fish Shellfish Immunol **16:**461-74.

58. **Rockey, D. D., P. S. Turaga, G. D. Wiens, B. A. Cook, and S. L. Kaattari.** 1991. Serine proteinase of *Renibacterium salmoninarum* digests a major autologous extracellular and cell-surface protein. Canadian J Microbiol **37:**758-63.

59. **Salonius, K., C. Siderakis, A. M. MacKinnon, and S. G. Griffiths.** 2005. Use of *Arthrobacter davidanieli* as a live vaccine against *Renibacterium salmoninarum* and *Piscirickettsia salmonis* in salmonids. Dev Biol (Basel) **121:**189-97.

60. **Sanders, J. E., and J. L. Fryer.** 1980. *Renibacterium salmoninarum* gen. nov., sp. nov., the causative agent of bacterial kidney disease in salmonid fishes. Int J Syst Bacteriol **30:**496-502.

61. **Sorum, U., B. Robertsen, and L. Kenne.** 1998. Structural studies of the major polysaccharide in the cell wall of *Renibacterium salmoninarum*. Carbohydrate Res **306:**305-14.

62. **Starliper, C. E., D. R. Smith, and T. Shatzer.** 1997. Virulence of *Renibacterium salmoninarum* to Salmonids. J Aquatic Animal Health **9:**1-7.

63. **Stinear, T. P., T. Seemann, S. Pidot, W. Frigui, G. Reysset, T. Garnier, G. Meurice, D. Simon, C. Bouchier, L. Ma, M. Tichit, J. L. Porter, J. Ryan, P. D. Johnson, J. K. Davies, G. A. Jenkin, P. L. Small, L. M. Jones, F. Tekaia, F. Laval, M. Daffe, J. Parkhill, and S. T. Cole.** 2007. Reductive evolution and niche adaptation inferred from the genome of Mycobacterium ulcerans, the causative agent of Buruli ulcer. Genome Res **17:**192-200.

64. **Sudheesh, P. S., S. Crane, K. D. Cain, and M. S. Strom.** 2007. Sortase inhibitor phenyl vinyl sulfone inhibits *Renibacterium salmoninarum* adherence and invasion of host cells. Dis Aquat Organ **78:**115-27.

65. **Turaga, P. S. D., G. D. Wiens, and S. L. Kaattari.** 1987. Analysis of *Renibacterium salmoninarum* antigen production *in situ*. Fish Path **22:**209-214.

66. **Turaga, P. S. D., G. D. Wiens, and S. L. Kaattari.** 1987. Bacterial kidney diseases: the potential role of soluble protein antigen(s). J Fish Biol **31:**191-194.

67. **White, O.** 2004. Bacterial Genome Annotation at TIGR, p. 33-46. *In* C. M. Fraser, Read, T.D., Nelson, K.E. (ed.), Microbial Genomes. Humana Press, Totowa, New Jersey.

68. **Wiens, G. D.** 2006, posting date. Bacterial Kidney Disease. . CAB International. [Online, http://www.cabicompendium.org/ac.]

69. **Wiens, G. D., and S. L. Kaattari.** 1999. Bacterial Kidney Disease (*Renibacterium salmoninarum*), p. 269-301. *In* P. T. K. Woo and D. W. Bruno (ed.), Fish Diseases and Disorders: Viral, Bacterial and Fungal Infections, vol. 3. CAB International.

70. **Wiens, G. D., and S. L. Kaattari.** 1989. Monoclonal antibody analysis of common surface protein(s) of *Renibacterium salmoninarum*. Fish Path **24:**1-7.

71. **Wiens, G. D., and S. L. Kaattari.** 1991. Monoclonal antibody characterization of a leukoagglutinin produced by *Renibacterium salmoninarum*. Infect Immun **59:**631-7.

72. **Wiens, G. D., R. Pascho, and J. R. Winton.** 2002. A single Ala139-to-Glu substitution in the Renibacterium salmoninarum virulence-associated protein p57

results in antigenic variation and is associated with enhanced p57 binding to chinook salmon leukocytes. Appl Environ Microbiol **68:**3969-77.

73. **Wood, D. W., J. C. Setubal, R. Kaul, D. E. Monks, J. P. Kitajima, V. K. Okura, Y. Zhou, L. Chen, G. E. Wood, N. F. Almeida, Jr., L. Woo, Y. Chen, I. T. Paulsen, J. A. Eisen, P. D. Karp, D. Bovee, Sr., P. Chapman, J. Clendenning, G. Deatherage, W. Gillet, C. Grant, T. Kutyavin, R. Levy, M. J. Li, E. McClelland, A. Palmieri, C. Raymond, G. Rouse, C. Saenphimmachak, Z. Wu, P. Romero, D. Gordon, S. Zhang, H. Yoo, Y. Tao, P. Biddle, M. Jung, W. Krespan, M. Perry, B. Gordon-Kamm, L. Liao, S. Kim, C. Hendrick, Z. Y. Zhao, M. Dolan, F. Chumley, S. V. Tingey, J. F. Tomb, M. P. Gordon, M. V. Olson, and E. W. Nester.** 2001. The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. Science **294:**2317-23.

74. **Wood, P. A., and S. L. Kaattari.** 1996. Enhanced immunogenicity of *Renibacterium salmoninarum* in chinook salmon after removal of the bacterial cell surface-associated 57 kDa protein. Dis Aquatic Org **25:**71-79.

75. **Young, C. L., and G. B. Chapman.** 1978. Ultrastructural aspects of the causative agent and renal histopathology of bacterial kidney disease in brook trout (*Salvelinus fontinalis*). J Fish Res Board Canada **35:**1234-1248.

Figure Legends

Figure 1. Circular diagram of the *Rs* ATCC 33209 genome showing gene and gene similarity to *Arthrobacter spp*. Circle order from inside to outside: 1) GC skew (blue = low, red = high); 2) GC content (red ≥ 50%, blue < 50%) ; 3) disrupted ORFs (red), 4) IS elements (green); 5) ORFs with similarity to *Arthrobacter spp.* FB24 with a score of < 1 e-10 (orange); 6) ORFs with similarity to *A. aurescens* TC1 with a score of < 1 e-10 (purple); 7) *Rs* ORFs shown by strand and colored by assigned function (see Supplemental Figure 4 for color key).

Figure 2. Examples of disrupted ORF's in *Rs*. The four panels show ORF maps from different regions of the genome, compared to syntenous regions of the *Arthrobacter* FB24 chromosome (panels A and C), or that of *Streptomyces coelicolor* (panel B). The *Rs* sequence is in the top box in each panel, and homologous ORFs are shown with the same shading in each panel. Nonhomologous ORFs are represented in white in each panel. Panel A shows a dipeptide permease operon in which each ORF in *Rs* is inactivated by frameshifts. There is also a complete IS*994* sequence (black) inserted into the center of the operon. Panel B shows two divergently oriented citrate synthase genes that are arranged similarly in the related species, *S. coelicolor*. One of the citrate synthase ORFs is truncated by serial frameshift mutations. An IS*994* sequence is indicated in black. Panel C shows a *tetP* sequence, encoding a candidate tetracycline resistance protein, that is interrupted in *Rs* via a frameshift mutation. Panel D shows a candidate fibronectin binding protein that may have been a sortase substrate, in which the

sortase signal is separated from the majority of the protein by a frameshift mutation. The horizontal bar in each panel is positioned over a 500 bp region of the chromosome that was sequenced in 5 different *R. salmonarum* samples, including 2 samples amplified directly from diseased fish collected at an Oregon hatchery. In 18/20 samples, the genome sequence of these regions was identical to the assembled *Rs* 33209 genome sequence.

Figure 3. *Rs* is closely related to *Arthrobacter spp*. by 16S rRNA sequence (A) and protein cluster analysis (B). (A) 16S sequences were aligned using ClustalW and the tree was constructed using the Neighbor-Joining method with 1000 bootstrap replications. Red dots indicate pathogenic bacterial species with evidence of genome reduction. Bar indicates nucleotide substitutions. The large arrows denote collapsed branches containing the following *Streptomyces* or *Corynebacterium* 16S rRNA sequences: *C. diphtheriae* NCTC 13129 (NC_002935); *C. jeikeium* K411 (NC_007164); *C. efficiens* YS-314 (NC_004369); *C. glutamicum* ATCC 13032 (NC_006958); *Streptomyces coelicolor* A3(2) (NC_003888); and *S. avermitilis* MA-4680 (NC_003155). Ten additional 16S rRNA sequences from completed genomes were included in the alignment but are not shown here. (B) Venn diagram displaying numbers of common and unique protein clusters between *Rs* and the sequenced *Arthrobacter spp.* The numbers of protein clusters and total ORFs within the clusters are shown.

Figure 4. Comparison of *Renibacterium* and *Arthrobacter* FB24 capsular biosynthesis genes suggests horizontal gene acquisition. Panel A shows a set of capsular synthesis

genes (RSal33209_1436- RSal33209_1439; gray) that differ from *Arthrobacter* FB24 ORFs (white) within an otherwise fully conserved region of the genome. The same-colored ORFs from each genome are homologous between the organisms. Panel B shows evidence for horizontal acquisition of a set of capsular synthesis genes (RSal33209_1338- RSal33209_1342; black) located between a complete and truncated IS994 sequence (hatched), inserted into a region of the genome that is otherwise syntenous with the *Arthrobacter* genome (white). The G+C % values for the ORFs in the region are indicated within each ORF.

Table 1. Oligonucleotide primers used for amplification of four fragmented RS ORFs from recent clinical isolates and clinically infected Chinook salmon.

Oligonucleotide name	Gene position	Oligonucleotide sequence
dppD/F (F)	RSal33209_1660	5’ GGC.CGC.GAT.GTG.TCG.ATG.GTT.TT
dppD/F (R)	RSal33209_1659	5’ GCC.CAA.GTG.CGG.CAC.TGC.AGC
Citrate Synthase (F)	RSal33209_2900	5’ GCG.CGG.AGA.GAA.GTT.CTT.GTG
Citrate Synthase (R)	RSal33209_2898	5’ CGA.TGC.GGT.GCG.ACG.TTT.T
tetP (F)	RSal33209_1874	5’ GCC.TAG.CGA.CGC.AAA.AG
tetP (R)	RSal33209_1873	5’ ATA.GTG.ACT.AAG.CAA.TCG.GTG
fbp (F)	RSal33209_1552	5’ CTG.ACG.CCA.ACG.GTA.AAT.ACA.CC
fbp (R)	RSal33209_1551	5’ GGC.GGA.TTC.TCA.ACA.CTC.ACG

Table 2. Genome statistics of *Renibacterium salmoninarum* strain 33209, *Arthrobacter aurescens* TC1, and *Arthrobacter sp.* FB24.

	***Rs* 33209**[a]		***Art.* aur. TC1**		***Art.* sp. FB24**	
Data Category[a]	**Counts**	**%**	**Counts**	**%**	**Counts**	**%**
DNA total sequenced, bases	3,155,250	100	5,226,648	100	5,070,478	100
DNA coding sequence, bases	2,854,595	90.5	4,587,903	87.8	4,538,003	89.5
DNA G+C content, bases	1,775,501	56.3	3,262,989	62.4	3,315,507	65.4
ORFs total	3,507	100	4,587	100	4,506	100
ORFs with assigned function	2,620	74.7	2,891	63.0	2,986	66.3
ORFs, no assigned function	887	25.3	1,696	37.0	1,520	33.7
ORFs, no function or similarity	240	6.8	262	5.7	147	3.3
ORFs, no function, with similarity	647	18.5	1,434	31.3	1,373	30.5
ORFs in paralog clusters	592	16.9	1,000	21.8	1,067	23.7
ORFs in COGs	2,428	69.2	3,382	73.7	3,426	76.0
ORFs with Pfam matches	2,002	57.1	3,273	71.4	2,700	59.9
COGs	1,064		1,262		1,287	

[a] ORF data are publicly available on ERGO Light (http://www.ergo-light.com/index.cgi/).

Table 3. Disrupted *Rs* ORFs: category, number, and percentage.

Category	**Total**	**Partial**	**Percent**
Transport/Binding proteins	346	152	43.9
Fatty acid/phospholipid metabolism	100	31	31.0
Central intermediary metabolism	170	48	28.2
Energy metabolism	304	86	28.3
Regulatory Functions	240	53	22.1
DNA metabolism	108	23	21.3
Cellular processes	90	18	20.0
Biosynthesis of cofactors/prosthetic gps.	109	24	22.0
Protein fate	152	28	18.4
Amino Acid Biosynthesis	111	18	16.2
Cell envelope	215	27	12.6
Signal transduction	6	1	16.7
Hypothetical/Unclassified/Unknown Fx	1182	198	16.8
Nucleotides	61	6	9.8
Protein Synthesis	128	9	7.0
Transcription	34	2	5.9
Mobile and extrachromosomal elements	151	6	4.0
Total ORFs	3507	730	20.8

Table 4. Comparison of ORF functional role categories between *Rs* and *A. aurescens* TC1.

	Rs[a]		*A. aurescens*TC1[b]	
Functional Role Category Breakdown	Number	Percent	Number	Percent
Transport and binding proteins	372	10.60	517	12.50
Energy metabolism	311	8.86	739	17.87
Regulatory functions	255	7.27	332	8.03
Cell envelope	220	6.27	340	8.22
Central Intermidiary metabolism	180	5.13	117	2.83
Protein fate	158	4.50	167	4.04
Mobile and extrachromosomal element function	158	4.50	35	0.85
Protein synthesis	135	3.85	123	2.97
Synthesis of cofactors, prosthetic groups, and carriers	112	3.19	141	3.41
Amino Acid biosynthesis	112	3.19	136	3.29
DNA metabolism	110	3.14	111	2.68
Cellular processes	93	2.65	186	4.50
Fatty acid and phospholipid metabolism	107	3.05	123	2.97
Purines, pyrimidines, nucleosides, and nucleotides	62	1.77	82	1.98
Transcription	34	0.97	89	2.15
Signal transduction	7	0.20	54	1.31
ORFS Total	3507		4136	

[a] ORF categories were determined by using Manatee and manual annotation (Supplementary Data Table 1). Partial ORFs were also included in these totals but it is unclear whether they maintain function.

[b] ORF numbers and categories are from (43)

Supplementary Information

Figure S1 *Rs* strain ATCC 33209 causes mortality in challenged chinook salmon. Five to seven gram chinook salmon were injected with 1 x 10^6 *Rs* ATCC 33209 resulting in 70% mortality by 65 days post-challenge. Data represent averages from duplicate tanks (n=75 fish/tank). This is a typical mortality curve for fish experimentally infected *Rs*.

Figure S2. Phylogenetic analysis using five highly conserved proteins indicates a close relationship between *R. salmoninarum* and *Arthrobacter spp*. Predicted protein sequences aligned included the LSU ribosomal protein L2P, chaperone protein dnaK, ATPsynthaseF1, Elongation Factor Tu (EF-TU) and gyrB. Sequences were concatenated and aligned using ClustalW and the tree was constructed using the Neighbor-Joining method with 1000 boostrap replications. Genome accession numbers: *A. aurescens TC1* (NC_008711); *A. sp. FB24* (NC_008541); *R. salmonianrum* (NC_010168); *K. rhizophila* (NC_010617); *M. luteus* (NZ_ABLQ00000000); *L. xyli subsp. xyli* (NC_006087); *B. linens* (NZ_AAGP00000000); *K. radiotolerans* (NC_009664); *S. coelicolor* (NC_003888); *C. diphtheriae* (NC_002935); *C. glutamicum* (NC_006958); *M. avium paratuberculosis K10* (NC_002944); *M. marinum* (NC_010612); *M. leprae TN* (NC_002677); *M. bovis* (NC_002945); *M. tuberculosis* (NC_002755, NC_000962); *B. subtilis* (NC_000964).

Figure S3. Multiple whole-genome alignment of *Rs* with two strains of *Arthrobacter*. The *Rs* genome is smaller and 67 co-linear gene blocks were identified between the three

strains. Alignment was constructed using Mauve 2.0 (Multiple Genome Alignment, asap.ahabs.wisc.edu/software/mauve/).

Supplementary Data Table 1. ORF annotations with worksheets including the following: consortium annotations, putative pseudogenes, secreted proteins, stress-related genes, sortase substrates, and ORFs without similarity or function.

Supplementary Data Table 2. ORF clustering in ERGO Workbench between *Rs*, *Arthrobacter sp.* FB24 and *Arthrobacter aurescens* TC1. Worksheets include ORFs common to all three species, ORFs common to two of the three, and unique ORFs for each species.