The microevolution of *Renibacterium salmoninarum*

Dissertation for the degree of *Philosophiae Doctor* (PhD)

Ola Brønstad Brynildsrud

Department of Food Safety and Infection Biology
Faculty of Veterinary Medicine and Biosciences
Norwegian University of Life Sciences

Adamstuen 2015

Thesis number 2015:50
ISSN 1894-6402
ISBN 978-82-575-1958-2

"I have yet to see any problem, however complicated, which, when you look at it in the right way, did not become still more complicated."

- Poul William Anderson (science fiction writer, 1926-2001)

Table of Contents

Acknowledgements

The present work was carried out at The Department of Food Safety and Infection Biology at The Norwegian University of Life Sciences (NMBU), formerly The Norwegian School of Veterinary Science (NVH), during the period of 2011-2015. The project was funded primarily through NMBU, with considerable support from The Department for Environment, Food and Rural Affairs (DEFRA), UK, for data acquisition.

I would like to express my gratitude to the many brilliant people who played a part in the creation of this thesis. Firstly, I would like to thank my research supervisor, Jon Bohlin, whose ever-enthusiastic character and playful approach to science in general and bioinformatics in particular has been a source of inspiration. You have always believed in me and have been highly permissive in your supervision; instead of showing me the ropes, you have let me discover the ropes and their intricate ways myself. I am adamant in my belief that this has made me a much better researcher. Thank you.

This work could not have been completed without the exceptional contributions of my many co-authors. In particular, I would like to thank Ed Feil at The University of Bath, UK, David Verner-Jeffreys at The Centre for Environment, Fisheries and Aquaculture Science (CEFAS), Weymouth, UK, and Linda Rhodes at The National Oceanic and Atmospheric Administration, WA, USA. Your excitement for knowledge advancement in microbiology has (ironically?) been highly infectious. Big thanks also go out to David Ussery, currently at Oak Ridge National Laboratory (ORNL), TN, USA, for inviting me to stay with his comparative genomics research group at The Technical University of Denmark (DTU) in 2011, and Lars-Gustav Snipen at Campus Ås for finding time for me in his already overscheduled calendar.

Special regards go out to all my great colleagues at the (now-defunct) EpiCentre at Campus Adamstua. I have enjoyed the spirited debates and discussions we have had on every conceivable topic. Thanks in particular to my excellent internal supervisor Eystein Skjerve, who has seen so many PhD students through the doctorate gauntlet, but still managed to squeeze in time for me. Extra praise must go to my good colleague Jostein Mulder Pettersen for always engaging me in meaningless repartee about everything and nothing. I am ever looking forward to quit our day jobs and perform elaborate art gallery heists instead. Alternatively, starting our long-planned business venture. I can never remember which one we're actually going to do, but here's to a long and prosperous career either way.

It has been a long and arduous ride, and so I must thank my family and friends, whose sustained support has always kept me going. Mother, sister; I know you're not going to understand much of the following, but I hope you enjoy the pictures. Thank you for your love and support, nonetheless. I hope my good friends in "Kontegutta" will enjoy some of the illustrations as well, even if they differ somewhat from our usual drawings. Finally, I would like to thank my dear Michaela, who continues to live with me in spite of my offbeat interests. You have been there every step of the way, and for that I am eternally grateful.

Oslo, May 2015
Ola Brønstad Brynildsrud

Abbreviations

BKD - Bacterial Kidney Disease
bp - base pairs
CNV - Copy Number Variation
EEA - European Economic Area
FAO - Food and Agriculture Organization of the United Nations
GWAS - Genome-wide association study
HGT - horizontal gene transfer
HMM - Hidden Markov Model
LBA - long branch attraction
MCMC - Markov Chain Monte Carlo
ML - Maximum Likelihood
MSA - Major Soluble Antigen (protein)
msa - major soluble antigen (gene)
MRCA - Most Recent Common Ancestor
NGS - Next-Generation Sequencing
NJ - Neighbor-Joining
nt - nucleotides
OIE - Office International des Epizooties (World Organization for Animal Health)
PCR - Polymerase Chain Reaction
qPCR - quantitative Polymerase Chain Reaction
SNP - Single Nucleotide Polymorphism
UK - The United Kingdom of Great Britain and Northern Ireland
UPGMA - Unweighted Pair Group Method with Arithmetic means
US - The United States of America
WGS - Whole-Genome Sequencing

List of Papers

Paper I

Microevolution of *Renibacterium salmoninarum*: Evidence for intercontinental dissemination associated with fish movements

Authors: Brynildsrud O, Feil EJ, Bohlin J, Castillo-Ramirez S, Colquhoun D, McCarthy U, Matejusova I, Rhodes LD, Wiens GD, Verner-Jeffreys DW

Published: The ISME Journal (2014), 8: 746-756

Paper II

CNOGpro: Detection and quantification of CNVs in prokaryotic whole-genome sequencing data

Authors: Brynildsrud O, Snipen L-G, Bohlin J

Published: Bioinformatics (2015), btv070 *(Epub ahead of print)*

Paper III

Identifying copy number variation of the dominant virulence factor *msa* within genomes of the fish pathogen *Renibacterium salmoninarum*

Authors: Brynildsrud O, Gulla S, Feil EJ, Nørstebø SF, Rhodes LD

Submitted

Summary

Renibacterium salmoninarum is the causative agent of bacterial kidney disease (BKD), a chronic infection of cultured and wild salmonids, which can result in acute morbidity or mortality or be a slowly progressive disease causing an often-dramatic decline in growth. BKD is economically important in aquaculture, where it can spread horizontally throughout sea pens or vertically through transferred broodstock or eggs. It is also a concern for conservation and restoration efforts for endangered fish stocks because infections are prevalent among free-ranging Pacific salmon in river and marine systems.

New advances in whole-genome sequencing (WGS) were used to provide previously impossible insights into BKD. We assembled the full genomes of 68 unique *R. salmoninarum* isolates whose origins range widely at spatial, temporal, habitat and host species levels. High-resolution reconstruction of the phylogenomic relationships between strains was possible by using single-nucleotide polymorphisms (SNPs) information.

R. salmoninarum was revealed to be a highly clonal bacterium with a relatively slow rate of evolution. Two main lineages were found to exist, provisionally named lineage 1 and lineage 2. Lineage 1 had a cosmopolitan spatio-temporal distribution, while lineage 2 were restricted to rivers of Eastern Scotland and fjords of Western and Northern Norway. Bayesian evolutionary analyses revealed multiple independent introductions of lineage 1 strains across the Atlantic ocean and from one side of the American continent to the other occurring in the last century-and-a-half, consistent with a hypothesis of anthropogenic spread by movement of fish and ova for aquaculture and recreational angling. The comparatively rare lineage 2 appears to have been long-term enzootic to Europe. Strains from different host species were indistinguishable, suggesting free inter-species transmission.

Peculiarly, all *R. salmoninarum* isolates appear to contain the full complement of genes in the species. However, the copy number of dominant virulence factors *msa* and *p22* varied from two to five and one to five, respectively. This copy number variation was common among North American isolates, rare in Norwegian, and completely absent in British. Analyses suggested that the trait had emerged multiple times in independent populations of North America due to local selection pressures.

In order to detect copy-number variants and quantify the number of paralogs per strain using WGS data we had to develop a completely new method, described in a standalone, full-length paper and available as open-source software.

In summary, the application of WGS and bioinformatic techniques to an important aquaculture pathogen has provided us with unprecedented insights into its genomics, evolutionary processes and transmission dynamics.

Sammendrag (Summary in Norwegian)

Renibacterium salmoninarum er organismen som forårsaker bakteriell nyresjuke (BKD), en kronisk infeksjon hos både oppdrettslaks og vill laksefisk, som kan resultere i akutt morbiditet eller mortalitet eller i andre tilfeller manifestere seg som en sakte progredierende sykdom som gir en dramatisk reduksjon i tilvekst. BKD er økonomisk betydningsfull innen akvakultur, der sykdommen kan spre seg horisontalt mellom merder eller vertikalt gjennom rogn eller flyttede avlsfisk. Sykdommen påvirker også naturvern-, truede dyrearter og artsmangfoldshensyn fordi den er prevalent i villfisk-populasjoner av stillehavslaks i elver og marine systemer.

Fremskritt innen helgenomssekvenseringteknikker (WGS) ble brukt til å oppnå kunnskap om BKD som tidligere ville vært umulig. Vi satte sammen genomene til 68 unike *R. salmoninarum*-isolater hvis opprinnelse varierte svært både geografisk, tidsmessig, artsmessig, samt i habitatet vertsfisken ble isolert fra. Vi kunne rekonstruere detaljert informasjon om fylogenomiske forhold mellom stammene ved å detektere og bruke enkeltbasepolymorfismer (SNPs).

R. salmoninarum viste seg å være en svært klonal bakterie med en relativt langsom evolusjonsrate. To hovedgrener ble oppdaget, og de ble foreløpig navngitt gren 1 og gren 2. Gren 1-stammer hadde en variert opprinnelse både med tanke på rom og tid, mens gren 2 kun ble funnet i områdene Skottland, Vest- og Nord-Norge. Bayesianske evolusjonsanalyser pekte på at det i løpet av det siste halvannet århundre har forekommet flere uavhengige introduksjoner av stammer tilhørende gren 1 på tvers av Atlanterhavet, samt fra den ene siden av det amerikanske kontinentet til den andre, noe som samsvarer med en hypotese om menneskelig spredning av sykdommen gjennom forflytning av fisk og rogn i forbindelse med akvakultur og utsett for hobbyfiske. De forholdsvis sjeldne gren 2-stammene virker derimot å ha vært enzootiske innen Europe i lengre tid. Stammer fra forskjellige arter var genetisk uadskillelige, noe som antyder fri smitte mellom artene.

Overraskende nok viste samtlige *R. salmoninarum*-isolater seg å inneholde en komplett samling av genene som er kjent innen arten. Genene varierte dog i antall. Kopitallet til de dominante virulensfaktorene *msa* og *p22* varierte henholdsvis fra to til fem og én til fem. Denne kopinummervariasjonen var vanlig blant Nord-Amerikanske isolater, sjelden blant norske og ikke til stede i britiske. Analyser viste at kopinummer-mutasjonen har forekommet gjentatte ganger i uavhengige populasjoner, og pekte på lokalt seleksjonspress i Nord-Amerika som en sannsynlig årsak.

For å detektere kopinummervarianter samt å kvantifisere antall paraloge gen per stamme kun ved hjelp av helgenomssekvenserings-data måtte vi utvikle en helt ny metode, som har blitt beskrevet i en egen artikkel, og som er tilgjengelig i programvare med åpen kildekode.

For å oppsummere har helgenomssekvensering og bioinformatiske teknikker blitt anvendt på en patogen som er viktig innen akvakulturen. Dette har gitt oss hittil uovertruffen innsikt i bakteriens genomikk, dens smittedynamikk og de evolusjonære prosessene som former den.

1. Introduction

1.1. Background

1.1.1. Feeding the world in the future

More than 800 million people around the world suffer from chronic malnutrition (1). With the total human population expected to grow to 9.6 billion by the year 2050, the world is faced with an enormous challenge in meeting the food and nutrition requirements of the future. The food production of the future must solve the immense problems of under- and (more importantly) malnutrition, and in order to preserve this world for future generations this must be done in ways that are economically and environmentally sustainable.

Aquaculture is by most measures the fastest growing food-production sector on the planet, and now accounts for roughly half the volume of fish meant for human consumption worldwide (1), expected to grow even further as catches from wild capture fisheries gradually level off. The global fish supply has grown steadily since the 1960s, and since the late 1980s this growth has come almost exclusively from aquaculture. With an annual growth rate of 8.6 percent since 1980, the supply growth has outpaced that of the human population, which is currently at 1.6 percent per annum. This has placed great hope on the aquaculture industry in feeding the future world with high-quality animal protein, essential fatty acids, vitamins and minerals.

However, global aquaculture output growth has declined in recent years, and production continues to be marred by critical obstacles, including fish diseases. Fish are typically reared in overcrowded environments, which can be linked to stress, poor water quality, altered microbial evolutionary pressures and host-microbe interactions, which can afford pathogenic microorganisms an ideal environment in which to spread and cause infectious diseases. This can be detrimental not just in terms of reduced output caused by mortality, reduced growth and downgrading of fillet quality, but can also have wide-ranging social and ecological impacts in the form of tarnished industry reputation and environmental spillover of disease. Furthermore, disease is unacceptable from an animal welfare point of view. It is therefore of the utmost importance that infectious diseases are controlled and that their distribution is reduced to acceptable levels.

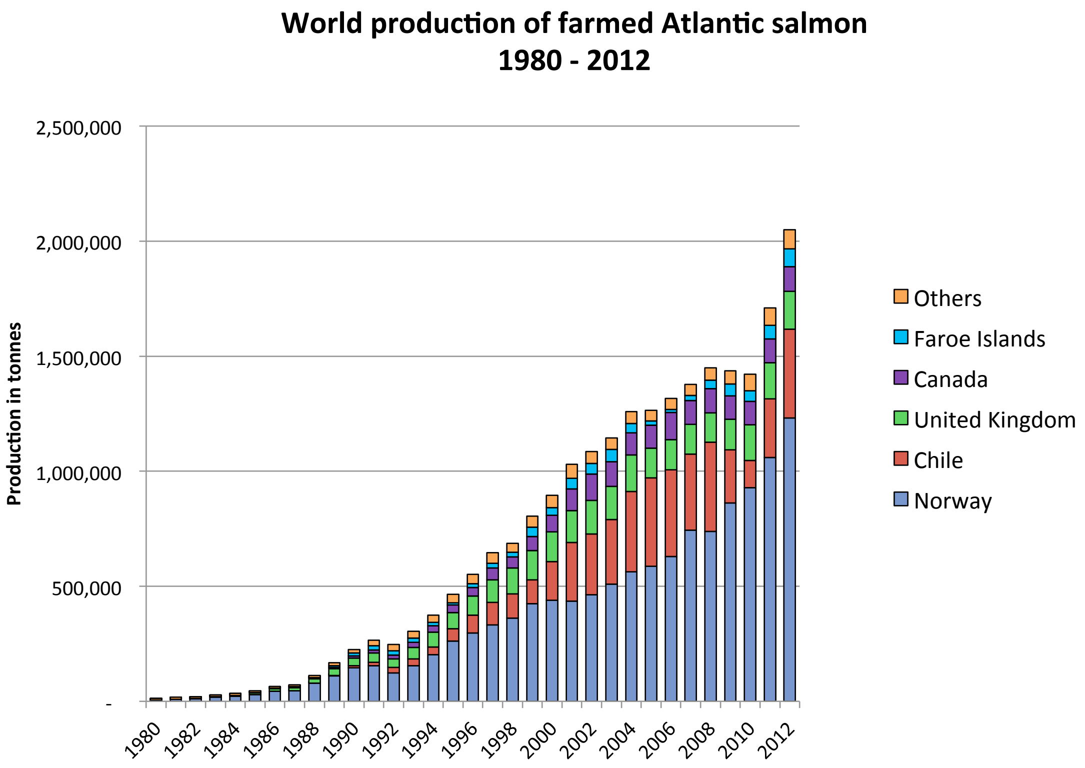


Figure 1. Global production of Atlantic salmon.

Production (in metric tonnes) of farmed Atlantic salmon 1980-2012, sorted by production country. Source: FAO FishStat.

1.1.2. A brief history of salmon aquaculture

Salmon, considered a high-value fish, is an important commodity in the world fisheries trade, currently representing 14 percent of global trade. Roughly two thirds of this share is reared salmon from aquaculture, while the rest is made up of wild catch (1). Numbers vary widely between individual species of salmon; in particular it should be noted that Atlantic salmon (*Salmo salar*), the numerically most important species, come almost exclusively from commercial farming. The history of salmon farming goes back at least to the 18th century (2), a time that saw the development of hatcheries concurrent with a decline in native wild fish populations (3). Unfortunately, the early days of artificial salmon propagation is quite poorly documented. What is known is that the technique was first invented in mid-18th century Germany, but not perfected until nearly a century later in France and Scotland. It was exported to the USA a few years later. The following years saw a dramatic expansion of hatcheries throughout North America and Europe, primarily through the efforts of private small-scale business ventures as well as freshwater stocking for recreational angling purposes. Unfortunately, the details of intra- and intercontinental fish movements have been lost to history, as the entire process was underpinned by a lack of supervision and ecological concern (2,4). To quote Edwin Pister: "*No one really knew what went where, when or why*" (5). Environmentally concerned voices were first raised about a century later, coinciding with the dawn of commercialized aquaculture (5). Rainbow trout was first reared in Norwegian waters as early as 1912, but commercial-scale aquaculture was not initiated until the late 1960s, following the development of sea cages and breeding enhancement of Atlantic salmon stock (6). It was quickly realized that the choice of Atlantic salmon for sea-based farming offered many benefits compared to other salmonids, and today it is by far the most common species in salmonid farming, accounting for ~90% of the farmed salmon market (1,3). Over the last few decades the total world production of Atlantic salmon has exploded in pace with a growing demand from established markets and the emergence of new ones, with a production exceeding 2 million metric tonnes in 2012 (7) (**Figure 1**). The increase in production has been accompanied by a number of emerging infectious diseases.

Infectious diseases may be introduced to farm through a number of routes, including (but not necessarily limited to) horizontal introduction to the farms through introduction of diseased fish, vertical transmission from parent to offspring intra- or extra-ovum, contaminated facilities, equipment, effluent water, or ballast water in transport, naturally occurring pathogens or contaminated feed (8). Although the industry is increasingly vertically structured with unidirectional flow of fish and other material, enormous numbers of live salmon, ova and associated water and equipment are on a daily basis being traded across national and subnational borders between and within aquaculture-keeping nations, making it extremely important that strict biosecurity measures are in place to prevent the unintended spreading of infectious disease.

1.2. Bacterial Kidney Disease

1.2.1. History

In 1933, the Second Interim Report of the Furunculosis Committee reported that during the spring and summer of the previous three years, multiple Atlantic salmon (*Salmo salar*) from Scottish rivers Dee and Spey had been found with mysterious, small necrotic lesions in their spleen (9,10). The disease came to be known as Dee disease. A clinically similar disease with gross pathology dominated by nephritic lesions was reported in various trout species (*Salvelinus fontinalis* - brook trout, *Salmo trutta* - brown trout and *Oncorhynchus mykiss* - rainbow trout) around the same time in North America, where it went under the name "white boil." It was later demonstrated that the same organism was implicated in both cases, and that they were in fact regional variants of the same disease (10), and subsequently the term was changed into the generic "kidney disease in salmonids," which was later changed to "Bacterial Kidney Disease (BKD)." Ordal and Earp were the first to successfully grow the reponsible pathogen; a very small, fastidious Gram-positive diplobacillus, in 1956 (11,12). It was initially considered to belong taxonomically to the *Corynebacterium* group, but in 1980 it was reclassified to its own genus and renamed *Renibacterium salmoninarum* (13).

BKD was eventually reported over most of North America including Canada and Alaska, Europe, Japan, Turkey (14), Taiwan (15), Sri Lanka (16), Iceland, Chile (17) and Venezuela (18), wherever salmonids were cultured (19) (**Figure 2**), with the notable exception of Australia and New Zealand (20). The latter half of the 20th century saw intensified efforts to more precisely understand the disease, and a number of important advances were made to the knowledge of the bacterium's pathogenesis and how to develop control strategies. This along with the emergence of other diseases has somewhat reduced the relative importance of BKD in aquaculture, but the disease remains a highly significant threat to viable fish rearing as well as wild fish worldwide, particularly in the Pacific Northwest and Great Lakes regions of North America.

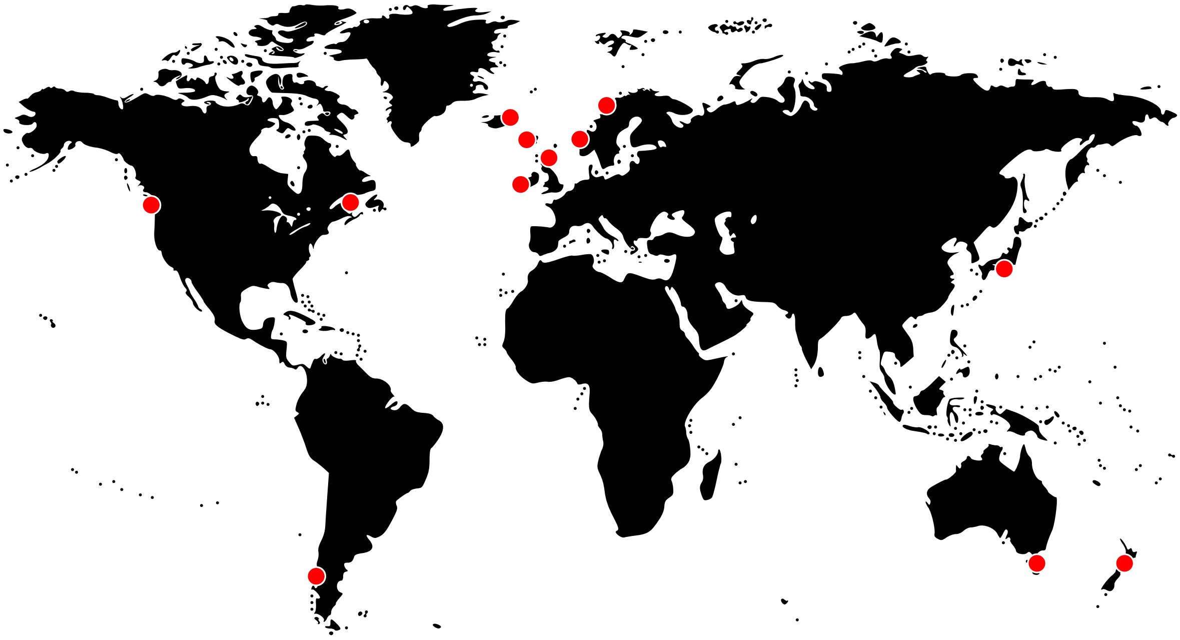


Figure 2. Main coastal regions used for salmon farming

The most important coastal regions extensively used for farming of salmonids are shown as dots. With the exception of Australia and New Zealand, BKD has been reported from all these regions. Source: Own work.

1.2.2. Epidemiology

Bacterial Kidney Disease (BKD) is a chronic, multisystemic, granulomatous disease affecting all species in the family *Salmonidae*. Infection with the causative agent, *Renibacterium salmoninarum* does not invariably result in clinical BKD, and apparently healthy fish may act as vectors of disease. This has led to speculation that the bacterium is an opportunistic pathogen, which may even be part of the normal salmon flora (21), yet others refute this, claiming no evidence exists (22,23). Furthermore, there are significant differences between host species in the degree of innate resistance to the infection, which means that asymptotic infections may be common. Rainbow trout (*Oncorhynchus mykiss*), lake trout (*Salvelinus namaycush*) and eastern brook trout (*Salvelinus fontinalis*) are regarded as being particularly resistant to the disease, with other Pacific species of salmon such as coho (*O. kisutch*), chinook (*O. tshawytscha*) and pink salmon (*O. gorbuscha*) exhibiting a lower degree of resistance. Atlantic salmon (*Salmo salar*) and brown trout (*Salmo trutta*) seems to be somewhere in the middle (22,24–26). External factors also play a role: Water temperature (10,27), chemistry (28,29), salinity, and fish diet (29,30) have all been shown to affect clinical progression. In addition, some specific stocks and transferrin genotypes have increased resistance to the development of disease (31,32), and bacterial strains have been shown to exhibit different virulence properties (33–37). Though it should be noted that the apparent resistance associated with some transferrin genotypes are most likely not related to the transferrin itself, but rather through genetic linkage to some unknown factor (31).

The incubation period is long, ranging from several months up to years. This allows the disease plenty of time to thoroughly infect congregations of hatchery and seawater-confined fish, so that entire stocks of fish may be ruined by the time the disease is discovered. The disease is rarely evident before the age of 6-12 months (19). Infections progress faster at higher temperatures, but this may have an inverse relationship with mortality. According to Wood (38), death occurs 30-35 days after exposure at temperatures greater than 11º C, and after 60-90 days at 7.2-10º C (19). However, total mortality is higher at temperatures below 12º C (27). Owing to its slow progression, BKD is rare in fry, although it is not unheard of in the more susceptible Pacific salmon species (21). It is most common in parr and smolt, and these age cohorts also have the highest mortality from BKD. Adult fish have higher innate resistance, which in most circumstances translates to lower mortalities and a histopathologically more granulomatous pattern of disease. In feral fish, BKD is occasionally observed as a cause of mortality in spawning adults, as a higher proportion of the fish's energy is diverted towards achieving reproductive success, leading to a relative weakening of the immune system (21,39).

R. salmoninarum is notable for being one of the very few bacteria that can be transmitted both horizontally and vertically, and both the horizontal and vertical transmission routes can cause clinical BKD (**Figure 3**). The vertical transmission is of the true, intracellular type, a highly specialized feature not commonly seen in pathogenic organisms. *R. salmoninarum* has the ability to enter, survive and replicate in the intracellular environment within oocytes and phagocytic cells such as macrophages and other leukocytes, and even cells with weak phagocytic qualities such as thrombocytes and endothelial cells (28,40–45). Infected oocytes subsequently fertilized can develop BKD at later life stages. Vertical transmission occurs only through female fish; there is no evidence that the pathogen can be transferred through milt (41).

The bacterium has poor survival rates in free water masses, but fares better in faeces and sediments, where it can live for up to 21 days (46,47). However, even if survival times are short, contaminated water may act as a vector of disease (48). BKD can also be transmitted between fish by direct contact (49). It is not known whether such direct and indirect transmission may also take place between different species (See: Knowledge gaps). Furthermore, *R. salmoninarum* can contaminate feed and infect fish that are fed unpasteurized foodstuffs. However, BKD infection *per os* is not as quick as through skin lesions (50). Finally, *R. salmoninarum* may enter the body and cause BKD via the eye (51).

1.2.3. The role of wild fish

BKD is commonly associated with cultured salmon, but also affects feral fish to varying degrees (52–54). In some regions BKD is thought to be enzootic in wild fish populations, but the prevalence in wild fish populations is in most regions largely unknown. In a 2008 study, the prevalence of BKD in wild fish in the UK was estimated as 0-10%, with the highest prevalence found in grayling (*Thymallus thymallus*) (53). Some river systems in the US have prevalences up to 100%, although on a positive note some wild fish populations in the Great Lakes area have reduced their prevalence in recent years (54). One study found fish capture location to be more important than fish origin in predicting prevalence (55).

It has been shown that diseased wild fish can infect farmed fish stocks (49), and the reverse scenario is also possible. The infection most probably originated in wild fish, and later became a problem when diseased wild fish were collected for establishment of hatchery stocks (19). As an example, it was first diagnosed in Norway in 1980 (21), and all diagnosed sites had used local feral fish as broodstock. As an example of the reverse situation, BKD was likely introduced unintentionally to wild fish in non-native areas such as the Great Lakes region of the US due to stocking of salmonids, particularly trout, at some point prior to 1975 (56).

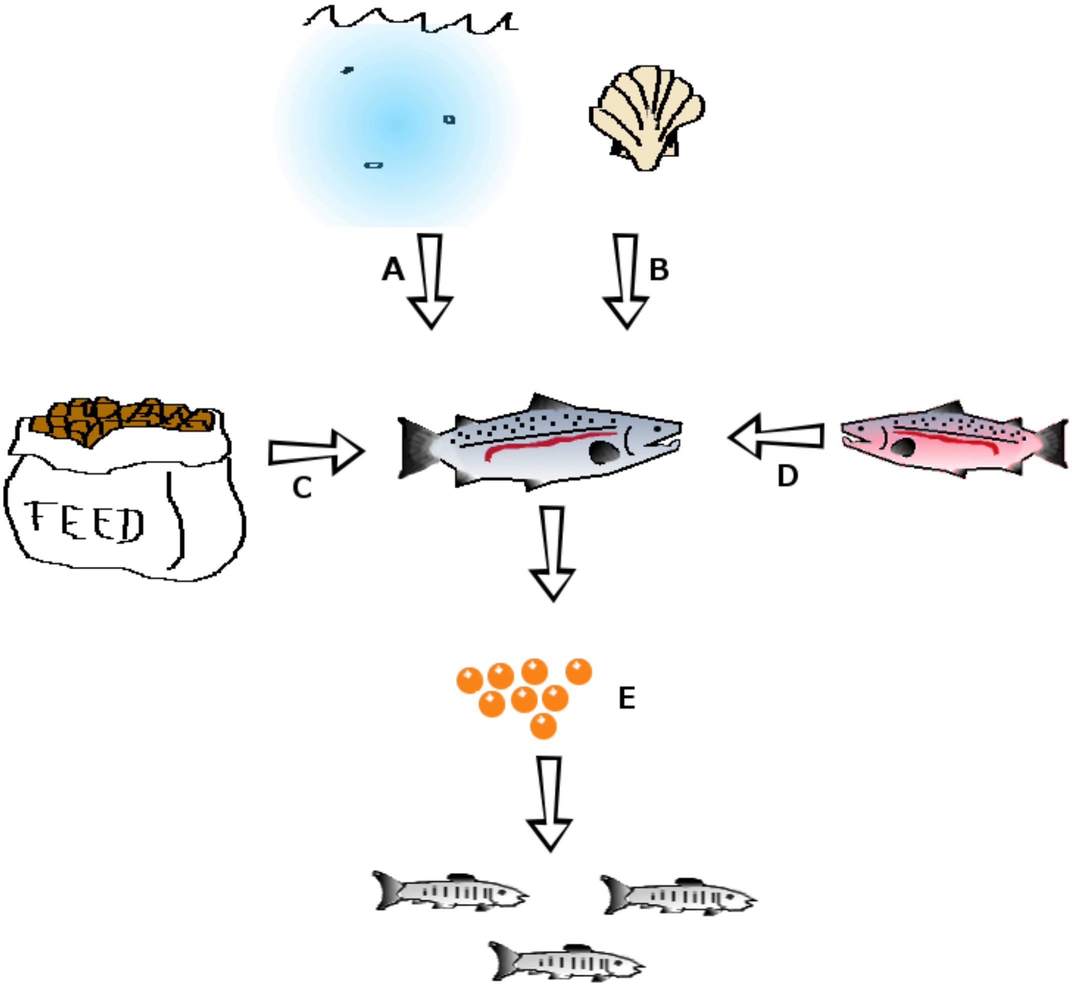


Figure 3. The main transmission routes for BKD

Salmonid fish can be infected in a number of ways: A) The bacterium can survive for a short time in free water masses and longer in faeces and sediments. Fish may then be infected through the skin, perorally or through the eye. B) The bacterium can also be found in non-salmonid hosts, although their importance as disease vectors are unknown. C) Through contaminated or unpasteurized feed. D) Direct transmission from transiently or permanently infected salmonids. E) By vertical transmission of the infection through ova. Source: Own work.

1.2.4. Intermediate hosts

The causative agent, *R. salmoninarum*, is occasionally cultured from non-salmonid fish including distantly related species such as carp (*Cyprinus carpio*) (57), greenling (*Heragrammos otakii*), flathead (*Platycephalus indicus*), Pacific herring (*Clupeo pallasi pallasi*) (58) and even European eel (*Anguilla anguilla*) (53) and sea lamprey (*Petromyzon marinus*) (59), but does not under normal circumstances cause BKD in these species. There have been reports of BKD or BKD-like disease in the non-salmonid lake whitefish (*Coregonus clupeaformis*) (60,61), although knowledge about this remains limited and will not be discussed further.

Moreover, it has been shown that *R. salmoninarum* are taken up and can be detected in certain species of bivalve molluscs that have lived close to salmon culture net pens (62). It is however unknown whether the molluscs can act as intermediate hosts, as these species derive nutrients from the digestion of filtered microorganisms, organic matter and particles.

Likewise, the role of ectoparasites such as *Lepeophtheirus salmonis* is not clear (63). Spread via these common parasites is plausible, but has never been demonstrated, and is likely to be a rare phenomenon, due to the fact that adult, skin-feeding sea lice seldom change hosts.
There has also been speculation that sea birds may in some cases carry the pathogen, but this is regarded as unlikely as the high body temperatures of most birds are not considered to be compatible with the preferred temperature conditions of *R. salmoninarum*.

1.2.5. Pathology

BKD-infected fish may not display overt external symptoms of disease, however late-stage BKD is accompanied by behavioral abnormalities such as erratic swimming, listlessness and disorientation (64). Gross external features of BKD include darkening of the skin, moderate to severe distention of the abdomen ("football appearance"), exophtalmia and ocular lesions, and ulcers and abscesses of the skin (28,51). Internally, the pathology is dominated by ascites, splenomegaly, renomegaly and pseudomembranes covering multiple organs or in other cases encapsulated nodules. The lesions are chronic granulomas, initially in hematopoietic tissue such as the kidney and spleen, but heart, liver and skeletal muscle may also be involved. In more developed cases of the disease all internal organs are affected. In Atlantic salmon, the encapsulation form of the disease is more common, and in this species it is under some circumstances even possible with a full recovery with resolving of lesions (34). In Pacific salmon, however, the pseudomembranous form of the disease is far more common, and encapsulation and removal of lesions is never observed. Histologically, the lesions are chronic granulomas, consisting of necrotic nuclei with infiltrating lymphocytes and macrophages surrounded by epitheloid cells (65). Bacteria are both extracellular and intracellular, residing in phagocytic cells such as macrophages, neutrophils, thrombocytes, monocytes and non-professional phagocytes. The presence of fibroblasts and a fibrous capsule is variable, being more common in aggressive pathogen strains and under more susceptible host-environment interactions (34,66). It seems as if much of the observed pathology should not be attributed solely to detrimental effects of the pathogen itself, but rather to a type III hypersensitivity response with immune complex deposition in the affected tissues of the host (25,67,68). This would at least to some extent explain why chemotherapy and vaccine efforts have often fallen short of their expectations.

1.2.6. Importance

BKD was considered an enormous problem worldwide from its discovery and until the mid-1990s. Despite extensive efforts at control, the disease remains an important source of loss in aquaculture, particularly in the Great Lakes and Pacific Northwest regions of North America (19,25,69). I have not found any serious attempts at quantifying the worldwide economic loss associated with BKD. Costs include those directly associated with outbreaks, such as mortality, fillet downgrading, treatment, diagnostics, vaccination and lower feed-conversion ratios, as well as indirect costs, including (but not limited to) infection of feral fish and environments from spillover, measures to secure the continuous supply of BKD-free broodstock, biosecurity implementation and socio-economic impacts such as decreased end-customer faith in salmon products and the ire of environmentalists and recreational anglers.

In Norway, a huge number of sites were quarantined in the 80s and early 90s (21). Effective stamping out and fallowing efforts in combination with significant changes to the industry (such as increased and better-quality screening efforts and a more vertically structured flow of potential risk-associated materials) around the mid-late 1990s seems to have led to a decline in its relative importance in European aquaculture.

A Norwegian surveillance program ran by the Norwegian Veterinary Institute 2005-2011 did not find a single case of BKD as a result of their monitoring, but sporadic outbreaks were diagnosed by routine health inspections of Atlantic salmon. For example, in 2011 one freshwater site and two on-growing sites in Northern Norway (70) tested positive for BKD. In 2012, there was one confirmed diagnosis in a restocking facility and one from feral fish caught in a stream (71). This indicates a generally low BKD prevalence, and underscores the difficulties of programs monitoring rare diseases.

In the UK, it is similarly considered a disease under control. The UK Fish Health Inspectorate (FHI) surveyed salmon farms in 1993 and 1994, and found a farm-level prevalence of 10%, although due to limitations of this testing it is likely that the true prevalence was higher (72). There has since been a clear downward trend. A recent article estimated the Scottish farm-level prevalence of Atlantic salmon as 0.7% and rainbow trout as ~18% (73). There are several explanations for the much higher prevalence in rainbow trout farms. Firstly, it is viewed as a relatively minor problem in rainbow trout aquaculture due to the relative resistance of this species (72). Secondly, these farms are usually managed with a continuous stocking scheme, meaning there is no fallowing interval between fish batches. This allows a persistent infection to establish at these sites. It has been reported that approximately 20% of trout farms in Scotland are under official control for this reason (74). The possibility that these farms are reservoirs for infection cannot be discarded. In this context is it perhaps interesting to note that BKD has not been found in wild salmonids in Scotland since the 1960s (75).

In contrast, BKD is considered a disease of major importance in the Americas. In Canada, recent estimates have placed the prevalence at 3% in West and East coast Atlantic salmon, and over 5% for west coast Pacific salmon (76). The results ranged up to 30% for one particular location. Central Canada did not have many problems with BKD. It is a massive problem in the Great lakes region with associated river systems, with prevalences approaching 100% in some places (77). As for marine regions, in the Puget Sound region of the Pacific Northwest, wild Pacific salmon prevalence ranged from 11% to 64% in a 2006 study (78).

One factor that greatly contributes to the high economic impact BKD has is that mortality usually peaks in fish older than one year old, after considerable financial investments have already been made (69). Due to the chronic nature of the infection one must assume that the entire stock is infected when the disease first shows. Mortalities can approach 40% in stocks of Atlantic salmon, and up to 80% in Pacific salmon (50,79). Rainbow trout are considered to be very resistant to BKD, and cumulative mortality numbers attributable to BKD are almost invariably low, even after intraperitoneal injection of the causative organism (49).

BKD has no known implications for human health.

1.2.7. Control

Prevention is the preferred control method for BKD (80) (OIE statement, 2013), as eradication is considered hard to impossible once the disease has become enzootic (81). This includes good husbandry practices that limit opportunities for exposure to the bacterium such as site fallowing, single-year class site division and the reduction of fish densities. One early industry measure that immensely contributed to BKD prevention (19) was the implementation of strict pasteurization of feed and viscera, first started in 1960 (28,43). Prior to this, it was common to feed raw salmon viscera, something that greatly exacerbated BKD problems in the 1940s and 1950s (82). Moreover, many countries have implemented legislation that places stringent restrictions on the import of live fish and ova and require licenses to do so, such as the UK Diseases of Fish Act (83) as well as various pieces of the European Economic Area (EEA) requirements in Europe and Federal legislation in the US.

The intracellular nature of the pathogen makes it hard to eliminate. The supply of broodstock and ova that are guaranteed to be BKD-free is of the utmost priority, so a zero tolerance policy should be implemented at that particular stage. At the time of writing, the industry is structured with relatively few suppliers, the most important of which are Aquagen AS, Fanad Fisheries Ltd, Lakeland and Salmobreed AS (84). The meticulous screening of eggs and roe undertaken by these suppliers in combination with the strict import regulations of both eggs and live fish implemented in many countries is no doubt partly responsible for the decline in the number of BKD outbreaks in much of the world during the 1990s.

Measures to handle BKD differ through the stages in the production cycle. The majority of smolt production is done in vertically integrated controlled freshwater environments (84). Norwegian salmon culture exclusively uses onshore tanks for this purpose, but in some other countries raceways and cages connected to lakes and river systems are more widespread, particularly for the production of rainbow trout and coho salmon. This has serious implications for the spread of BKD as well as other diseases and pollutants, and the proper treatment of effluent water is imperative. Norwegian regulations do not allow smolts from BKD-infected facilities to be put in sea cages (85). If BKD is suspected in the sea cage phase, the locality is immediately quarantined, but slaughter is not necessarily initiated immediately. However, strict biosafety will be required at the eventual slaughter, and the locality is required to remain fallowed for some time afterwards.

Iodophor disinfection has been practised since the 1970s to kill viral and bacterial pathogens associated with egg surfaces, coelomic fluid and milt (84). This is effective in treating surface-associated *Renibacterium salmoninarum*, but completely fails at killing bacteria that reside *in ovum* (40,42,43,86,87). Many studies have looked at the efficacy of egg treatment with erythromycin and other antibiotics, but there is currently no scientific consensus on the effectiveness and feasibility of this approach in eliminating intracellular concentrations of the bacterium (43,88).

Multiple studies have investigated the effects of disinfection of hatchery water, which is useful both in the decontamination of effluent and recirculating water and the protection of the environment by eliminating microbial spilling from effluent (43). Austin (89) found ozone to effectively remove *Renibacterium salmoninarum*, and a Danish report also found pH treatment (both acidic and basic), chlorination and heat treatment to be effective (90). I have not found any admissible documentation for the efficiency of UV radiation for this purpose, even though this method is currently permitted for wastewater from hatcheries, slaughterhouses and research facilities under Norwegian legislation (91).

Antimicrobial therapy is possible (92), but not recommended, as there are significant concerns for the development of antimicrobial resistance, and strains with reduced susceptibility to the commonly used antibiotic erythromycin have already been observed (93). Furthermore, due to a combination of bacterial encapsulation, long bacterial generation times and poor target tissue penetration, antibiotic concentrations are thought to drop too rapidly to properly eliminate the bacteria (93,94). Despite this, erythromycin is extensively used in hatcheries and female broodstock rearing facilities in the US. Antibiotic therapy is seldom used for the control of the disease in Europe, as neither erythromycin nor any other effective antibiotic compound have established maximum residue limits (MRL) for use in fish, and as such remain unlicensed (81). The most effective antibiotic for BKD treatment is rifampicin, but this compound is not considered acceptable for this use as it is reserved for the treatment of tuberculosis and other severe chronic infections in humans.

The properties of BKD make it a less than ideal candidate for vaccine development. Humoral antibodies are slow to develop and either are not very effective against the causative bacterium (67,94,95), or may in fact exacerbate or even incite the development of disease. One vaccine is currently commercially available: Renogen® (Novartis AG, Basel, Switzerland) contains a lyophilized live culture of *Arthrobacter davidanieli* (96) (deposition number ATCC 59921 in the American Type Culture Collection), a non-pathogenic bacterium that shares common antigenic determinants with *Renibacterium salmoninarum*. This vaccine is used extensively in North America and Chile (97) but not in Europe (Paul Midtlyng, formerly of Novartis International AG, personal communication, 2015). It should be noted that the vaccine has also been claimed to be effective against salmonid rickettsial septicemia (SRS) (96), so at least some of the use may be attributed to the prevention of that disease.

Despite several efforts to control BKD by selective breeding of Pacific salmon species, this approach has not truly succeeded yet, primarily due to knowledge gaps about genetic markers that are indicative of resistance and shortcomings of methods that mortality rates as the predominant selection criterium. Adding to the difficulties, the heritability of BKD resistance seems to be very low (98,99). According to a 2006 review by Balfry and Brown, there is no conclusive evidence to support breeding efforts as a viable strategy for BKD reduction (69).

1.3. *Renibacterium salmoninarum*

1.3.1. Characterization

R. salmoninarum is a very small (0.3-1.0 x 1.0-1.5 µm), non-acid-fast, non-sporulating, non-motile Gram-positive bacterium (13) whose microscopic appearance ranges from vaguely rod-like to coccoid, often appearing in pairs (**Figure 4**). It grows best at 15-19 ºC, with poor growth at 5 ºC and 22 ºC, and none at all at 30 ºC (25). However, regardless of temperature, it is an extremely slow-growing organism. On agar, the first traces of colonies can be observed after 2-3 weeks, but in some cases up to 12 weeks may be required. This prolonged growing time often allows competing microorganisms to establish on the agar and contaminate the plate by overgrowth. The most widely employed medium for growing the bacterium is based on a recipe called Kidney Disease Medium (KDM), first published by T. P. T. Evelyn (100). KDM contains L-Cysteine, which *R. salmoninarum* has a particular affinity for. The organism will not grow in non-Cysteine-enriched blood or trypticase yeast agar (81). Several improvements on KDM have been made since the original recipe, notably KDM2 that adds a "nursing culture" of stock *R. salmoninarum*, which accelerates growth (101), and the inclusion of the antibiotic compounds cycloheximide, cycloserine, polymyxin B sulphate and oxolinic acid to prevent growth of competing microorganisms in selective KDM (SKDM) (102). Macroscopically, colonies are shiny, smooth, round, raised colonies that in color ranges from white to creamy and in size from pinpoint to ~2mm (81).

There is no scientific consensus as to the relative importance of *in ovo* and horizontal infection routes as mechanisms of disease transmission (41,50), although it is clear that the highly specialized ability of true vertical transmission must be important. Despite the importance of this infection route, very little is known about how exactly *R. salmoninarum* manages to enter, survive and replicate within host cells. It has been suggested that *R. salmoninarum* uses antibody and complement, as well as other serum components to activate the respiratory burst system of phagocytizing cells immediately upon contact, thus exhausting these cells, and making uptake and intracellular survival possible (44,103,104). Furthermore, the bacterium seems to be somewhat resistant to the digesting effects of lysozyme (105).

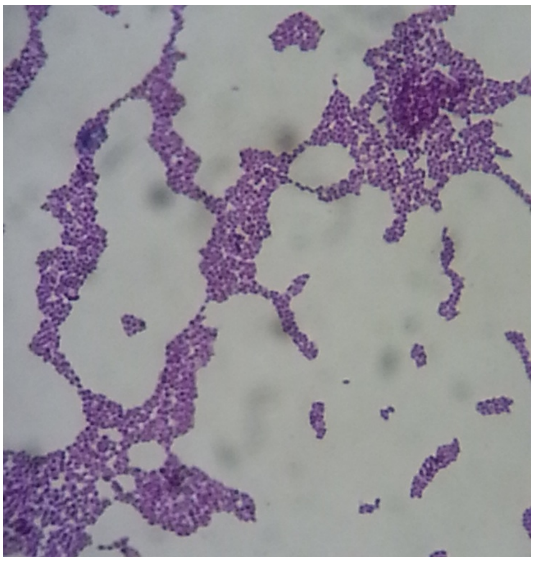


Figure 4. Light microscopy of *R. salmoninarum*

Bacteria are Gram-positive and coccoid. Source: Own research.

Many of the special capabilities of *R. salmoninarum* have been linked to a major soluble and cell-surface associated 57 kDa protein named Major soluble antigen (MSA), p57 or antigen F (68,106–111). MSA is the immunodominant antigen of *R. salmoninarum*, making up 60-70% of all surface protein (107,112), and host immune response is primarily directed against the MSA protein (68). Ironically, MSA is strongly immunosuppressive (106,107). Furthermore, it is involved in agglutination (35,111) and virulence (113), and being the protein towards which most of the immune response is directed, is necessary for the development of clinical BKD (68). Vertically infected fish that have been exposed to MSA in the egg develop an immunotolerance, and do not ever mount a proper immune response towards the protein (106).

There is a dose-response relationship between bacterial cell-surface-associated MSA protein and virulence (33,34). The spontaneous mutant strain MT239, for example, is attenuated, most likely due to a failure of cell wall-anchoring of the MSA protein (114). It has been hypothesized that the MSA protein can form adhesins that project through a polysaccharide capsule to form petrichious fimbriae (115). This adds to the bacterium's hydrophobicity, and may be an important virulence factor.

The second-most abundant surface protein of *R. salmoninarum*, provisionally named p22 (107,116), is another known immunosuppressive protein that reduces antibody production *in vitro* of *in vivo* stimulated cells, but otherwise little is known about it. Other known virulence factors are capsular synthesis, heme acquisition, haemolysins, cytolysins (117), and high hydrophobicity (118).

1.3.2. Taxonomy

The bacterium that caused BKD was initially thought to belong taxonomically to the *Corynebacterium* genus (11), but in 1980 the monospecific genus *Renibacterium* was established (13). Despite the extensive developments that have taken place in microbiology and genomics since that time, no additional species have been added to the genus. *R. salmoninarum* is most closely related to the non-pathogenic soil bacteria of the *Arthrobacter* genus, from whom it evolved by genomic reduction and horizontal gene acquisition (119). It is rather unclear how exactly these bacteria have evolved from each other, as *Renibacterium* does not survive long outside the salmonid host, and not a single member of the *Arthrobacter* genus is known to colonize any host species.

The *Renibacterium* genus is marked by a very high degree of clonality. This has been independently verified by multiple studies for such diverse characteristics as chemical and general microbiological properties (120), serological homology (110), peptidoglycan and cell wall similarity (121), antigenic structure (111), insertion sequence configuration (122), rRNA operons (123) and other molecular markers (124–127).

Despite this high strain homology, there is evidence that strains differ in their virulence properties (26). Much research has therefore been focused on finding exact and appropriate typing techniques. Despite this, the phylogeny of *R. salmoninarum* was relatively roughly described prior to the herein presented studies. A common method was to look for polymorphisms in targeted sequencing of the 16S-23S ribosomal DNA spacer region, which could group isolates into four spacer variants; SV1 from Canada, Norway, Sweden, UK and USA, SV2 from Iceland and Japan, SV3 from Canada and SV4 from Norway and Scotland. Randomly amplified polymorphic DNA (RAPD) could to some extent differentiate isolates and broadly group by geographical region, but without any quantification of patristic distance between isolates. Wiens and Dale used a combination of different techniques, including monoclonal antibody binding, tandem repeats and the presence or absence of a third copy of the *msa* gene, to some success (128). Nevertheless, high-resolution phylogeny relationships between strains remained unclarified.

1.3.3. Genome

The reference strain of *R. salmoninarum* is deposited in the American Type Culture Collection (ATCC) under number 33209. Originally named Lea-1-74, it was isolated from a yearling chinook salmon (*O. tshawytscha*) of the Leaburg hatchery of western Oregon in 1974 (Information from ATCC). The ATCC 33209 genome was fully sequenced in 2008 using capillary array sequencing technology (119).

It was a single circular chromosome of length 3,155,250 bp with no integrated phage or plasmid. Its GC-content was established as 56.3%. The genome was predicted to contain 2,777 protein-encoding open reading frames (ORFs), and an additional 730 pseudogenes. The genome was subject to extensive pseudogenization, as around one in five genes had been inactivated by point mutations, frame shifts, insertion sequences or deletions. This abundant pseudogene organization along with other characteristics, such as frequent insertion elements, size reduction, chromosomal rearrangements, horizontal gene acquisition and a high degree of clonality within the species (119,129), makes the genome of *R. salmoninarum* typical of recently emerged pathogens.

The genome of ATCC 33209 contains abundant insertion sequence elements with a seemingly random distribution in the genome. It contains 69 copies (69 and 67 of *orfA* and *orfB*, respectively) of IS*994*, an IS3-family element with homology to the IS*6110* elements of *Mycobacterium tuberculosis* (122). It additionally possesses 10 copies of IS*Rs2* and 1 copy of IS*Rs3*, which are other insertion sequences with homology to transposases of other pathogenic bacteria such as *Rhodococcus* spp., *Streptomyces* spp. and *Mycobacterium* spp.

It has been shown that two identical copies of the *msa* gene are present in several different *R. salmoninarum* isolates, including ATCC 33209 (114). Duplication appears to be a relatively rare phenomenon in prokaryotes. Interestingly, some isolates even possess a third identical copy of the *msa* gene, and this genotype is associated with increased virulence (36).

There has also been some interest devoted to the presence of several antibiotic resistance factors in the genome of *R. salmoninarum*. Factors include multidrug transporters, beta-lactamases, efflux proteins, tetracycline resistance factors, and genes involved in macrolide resistance such as macrolide glycosyltransferases and ribosomal RNA methyltransferases (119,130). The observation that the *R. salmoninarum* genome contains macrolide resistance factors is particularly interesting since this class of antibiotics is extensively used to control BKD in North America.

1.4. Knowledge gaps

Despite many advances in the understanding of *R. salmoninarum* since its initial discovery, including the 2008 publishing of reference genome sequence ATCC 33209, many questions remained unanswered at the commencement of the present PhD project. (And equally many or more questions remain after its completion.) The following section is meant to represent the contemporary knowledge gaps when my PhD project began in 2011.

Many studies have found a high degree of clonality within the species, but hitherto all tests have been investigating a relatively limited set of markers: The real diversity in terms of gene content and genetic variants is completely unknown. The comparative genomics of *R. salmoninarum* is unexplored territory. Also, more than one in three ORFs in the ATCC 33209 genome is a hypothetical gene, whose existence has been predicted by sequence analysis but for which there is no experimental evidence of transcription and without any function predicted by homology to previously annotated ORFs from public databases. Moreover, absolutely no information pertaining to the regulation of these or other genes are available. Thus, a huge amount of proteomics and transcriptomics work is needed to fully understand gene function and regulation within *R. salmoninarum*.

Furthermore, many papers have explored effective typing schemes for *R. salmoninarum* (127), but there is still hope that a more discriminatory method with high cost- and labor-efficiency could improve diagnostics.

A long-standing unresolved question is whether there are distinguishable strain types between the various host species that contract BKD. This is relevant for a number of reasons. Firstly, it is unknown whether BKD can be readily transmitted between different hosts, particularly between rainbow trout and Atlantic salmon, as this would be a strong argument for more stringent control measures on zoning as well as transportation of fish and affluent/effluent disinfection between neighboring farms that rear different species. Furthermore, although it is known that BKD outbreaks are usually much more severe in Pacific salmon species (except rainbow trout), it is not known if this is primarily due to host or geographical/environmental factors or whether there are subtypes of *R. salmoninarum* with higher tropism for particular hosts, and if such subtypes exists, their associated virulence phenotypes. Finally, there is incomplete knowledge about whether distinguishable and cross-infecting subtypes circulate between wild and farmed fish, and between saltwater and freshwater habitats.

The global phylogeny of *R. salmoninarum* isolates is similarly scarcely described, although a regional strain distribution has been confirmed in several studies. It is unknown how genetically different these types are though, and to a large degree whether there are phenotypic differences. One interesting mutation that has been observed in some geographically constrained strains is the substitution of an alanine with glutamic acid at position 139 in the MSA protein, as this mutation seems to have substantial implications for agglutination activity (37). It is thus of major interest to properly map other isolates for this mutation, as this could inform decisions on BKD restriction policy, treatment and vaccinology.

As a final point, one long-standing hypothesis on BKD transmission history states that it has been severely influenced by early days' relatively uncritical transport of live fish, ova and unpasteurized offal by professionals and hobbyists alike. These hypotheses have largely come about from theorycrafting and extrapolation from other pathogens such as *Aeromonas salmonicida* ssp. *salmonicida* and *Yersinia ruckeri*, both of which were largely disseminated by anthropogenic means (131–133). To the best of our knowledge there have been no studies describing the global dissemination history of BKD.

2. A short treatise on relevant bioinformatic methods

From the previous chapter it is evident that many of the unanswered questions about *R. salmoninarum* relate to functions that are encoded in its genome (such as for example gene regulatory networks) or information that can be deduced from a collection of genomes (such as phylogeny or strain typing recommendations.) The advent of various types of sequencing technologies allows for careful study of both. However, in this thesis I will focus on the latter, namely information that can be inferred by comparing a collection of genomes. In the following chapter I will therefore leave the subject of *R. salmoninarum* and rather discuss technologic and algorithmic solutions that have been the basis for all the data I have generated and the conclusions I have reached.

2.1. DNA sequencing

DNA sequencing is the process of determining the composition and order of nucleotides in a DNA molecule. It was first pioneered in the 1970s, but due to astronomic costs in terms of time and money, did not see extensive use until much later. In 1995, the bacterium *Haemophilus influenza* was published (134); making it the first-ever sequenced full-length DNA molecule that was not of viral or organelle origin. Since that time DNA sequencing has become indispensable in all biosciences in sync with price and time drops, more automated pipelines, and increased data output. Today, DNA sequencing is quick and inexpensive (**Figure 5**), but with the enormous quantities of data produced it has become an increasingly more labor-intensive job to curate, process and analyze the results. This has increased the demand for bioinformaticists and called for more efficient algorithms as well as increased computer memory and speed.

A central tenet of all current-generation DNA sequencing technologies is that due to technological constraints it is unfortunately impossible to read all nucleotides of an entire DNA molecule in one go. Rather, the genome is fragmented into templates that are individually sequenced to produce *reads*, which are continuous stretches of sequence. Later, the reads are sorted and reassembled computationally into *contigs*, which in this context means contiguous sequence, overlapping DNA fragments that together represent a consensus region. When the distance (in terms of number of base pairs) between two contigs as well as their relative orientation is known, but the sequence between the contigs is unknown, the two can be connected into a *scaffold*. Sequence assembly is a complex computational problem, and complete assembly is usually impossible due to the fact that most genomes contain multiple regions that are identical to each other, known as *repeats*. Problems arise when the repeats are longer than the reads. In order to completely reconstruct the genome, reads need to span across repeat regions, otherwise the assemblies collapse into contigs and cannot be joined. Therefore, the longer the read, the better.

A notable concept is that of *paired-end sequencing* (135). If a template is sequenced from both sides, and the size of the template is known, then by deduction the distance between the reads is also known. Since a known distance (which might be subject to variation) links the paired reads, the effective read length becomes equal to the sum of the read lengths plus the length of the gap between them. Paired-end sequencing is thus an effective way of overcoming technical constraints on read lengths, which is important since the only way of resolving repetitive sequences is by having reads span across them.

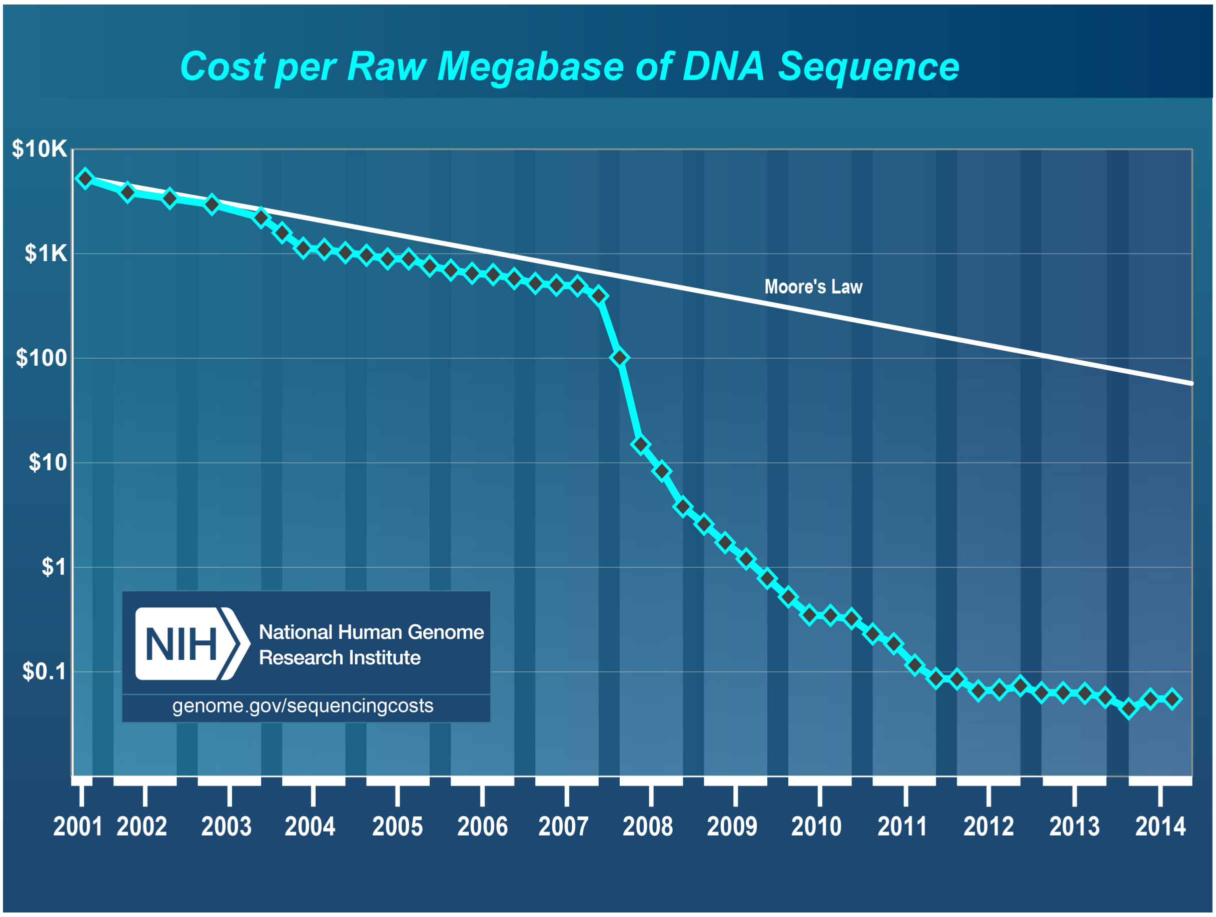


Figure 5. Sequencing costs

Costs associated with sequencing, measured in price per megabase. Development followed Moore's law (cost halved every two years) until late 2007, when costs dropped dramatically as new technologies hit the market. Source: NIH.

In the following section I will briefly introduce the most important DNA sequencing platforms. The technologies are rather confusingly categorized according to their proprietary companies as well as different methods of template preparation, sequencing, imaging, and data analysis.

2.1.1. First-generation: Sanger sequencing

This technology is based on DNA replication in an *in vitro* environment; DNA strands are separated and incubated with primers, polymerase, normal nucleotides and chain-terminating dideoxynucleotides that are labeled with dyes or radioactive phosphorous (136,137). The latter, when incorporated into the newly synthesized DNA strand interrupt strand elongation. Sequence fragments are then read after separation by size using gel electrophoresis. This method is only effective for strands as long as 100-1.000 base pairs long. There are two principal methods of sequencing longer stretches:

Primer walking sequentially shifts between sequencing of DNA fragments up to ~1.000 bp and designing new and appropriate primers by examining the terminal nucleotides of the previous read. It is then possible to design primers that start a new "step" from where the previous iteration stopped. Primer walking therefore finds consecutive stretches of DNA sequence.

Shotgun sequencing randomly divides the DNA molecule into shorter fragments that are sequenced to produce *reads*. Since the fragmentation is random, by performing several rounds of fragmentation and sequencing one will obtain multiple partially overlapping reads. A further advancement of the technology incorporates a hierarchical process where the DNA molecule is first sheared into several medium-sized pieces that are then internally ordered. Further shearing and sequencing can then proceed domain-wise. Note that shotgun sequencing technology has been adapted for Next-Generation sequencing as well. It is almost always used when the goal is to sequence an entire genome (*whole-genome sequencing* - WGS) rather than just a specific piece.

2.1.2. Next-Generation Technologies

Sanger sequencing was the leading sequencing technology for more than 20 years, but has now largely been replaced by technology with higher throughput (i.e. resolving a higher number of bases/hour). The main selling point of these technologies is that they are "massively parallel", meaning that they output enormous amounts of data. This comes at the cost of read length and per-base accuracy. However, due to massive redundancy, the accuracy skyrockets when considering the consensus at each base.

Pyrosequencing (454 sequencing) is a type of sequencing by synthesis (138). It works by capturing shotgun fragments to arrayed primer-coated beads (i.e. no cloning is necessary), and nucleotides are added sequentially, and for each elongation step pyrophosphate is released and measured by luminometry. A major advantage of 454 sequencing is the relatively long read length (~700 bp) compared to many other Next-Generation methods. However, it is quite expensive and has issues with homopolymer sequences.

Illumina dye sequencing, in contrast, produces shorter reads (50-300 bp), but much higher total output, resulting in high accuracy. It is another type of sequencing by synthesis (139). The process begins with binding DNA molecules to primers on a slide, and amplifies that DNA into "DNA clusters". Termination-types of all four nucleotides are then added, and these compete for binding. The non-bound nucleotides are then washed away, and a laser reads the dye of the incorporated base. Sequencing thus proceeds base-by-base.

Single molecule real time (SMRT/PacBio) sequencing is a third type of sequencing by synthesis (140). Sequencing is done on a chip that contains many wells, each well containing a single DNA polymerase and a single DNA molecule. Dyed nucleotides are incorporated as corresponding to the template DNA strand, and the dyes are cleaved off upon binding and read by a detector. PacBio sequencing is expensive but creates very long reads (up to 10.000-40.000 bp; Read lengths actually follow an exponential distribution, so most reads are shorter than this), making it invaluable for *de novo* sequencing and for genomes with many repeated elements. It also suffers from a high raw error rate (~14%), but since the error model is stochastic high qualities can still be achieved in the consensus sequences.

Ion semiconductor sequencing is a final type of sequencing by synthesis (141). It also uses microwells containing DNA fragments, but the detection of nucleotide binding is based on hydrogen ion release upon polymerization. It is rapid and cheap, but has some accuracy and homopolymer issues. Reads are up to 400 bp.

SOLiD sequencing instead employs sequencing by ligation (142). DNA is immobilized on a bead or other solid surface, and sequencing proceeds by the preferential binding of certain oligonucleotides (8-mers) to the DNA strand by DNA ligase. Fluorescent dyes are cleaved off to inform on which oligonucleotide were bound. SOLiD sequencing is cheap, but slow. It also comes with short reads (25-80 bp) and issues with palindromic sequences.

2.2. Genome assembly

An intuitive approach to DNA sequencing would be to take one genome and read it from start to end, as one would read a book. But as explained in the previous chapter, things are unfortunately not this simple. A more appropriate analogy would be that we have a large stack of copies of the same book, which are next shredded to tiny little pieces. Our goal is then to reconstruct an original copy of the book by gluing together miniscule page scraps that perhaps only contain a few words (**Figure 6**). Worse, the book contains many similar or identical paragraphs, and the shredder may have introduced typos into some of the words. Finally, the author may have suffered from severe writer's block, having borrowed entire sections from other books and in some cases having written sentences that are complete gibberish.

In reassembling the original book, it would help immensely if we had a copy of a very similar book, say, the first edition, as we would not expect very much to have changed between editions. Most of the structure would probably be identical, and entire chapters may deviate from each other by only a few sentences. This would allow us to use the first-edition book as a template, so that we only needed to find which book section most closely resembled any individual shred. It might also be possible to reconstruct chapters that were not in the previous edition, as long as there is some unique overlap to old book sections. This is analogous to *alignment* of reads to a known reference genome, which is a lot less complex than *de novo assembly* of reads, where (in naïve algorithms) all pairs of reads need to be compared. In the book analogy this would be similar to taking the first shred piece and next checking every other shred to see if there is some overlap between the two. In this case we might not even know whether we are assembling a biography or a science book, or even what language the book is written

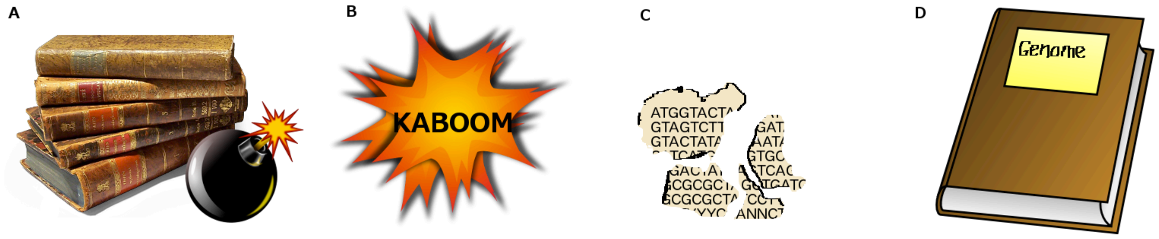


Figure 6. Exploding book analogy

A) Multiple copies of a book exist in a sample. B) A bomb explodes, tearing pages into tiny, charred fragments. C) Overlapping fragments that create meaningful sentences can be assembled into paragraphs, pages and chapters. D) Our goal is to recreate a full copy of the initial book. Source: Own work.

in, however, such information could potentially help us simplify the assembly procedure.

In designing solutions to this problem, we need to prepare for enormous amounts of data (ranging into terabytes), non-simple assembly due to repeats (exponentially increasing algorithmic time and space complexities) and errors in reads from the sequencing instruments.

2.2.1. Assembly strategies

All currently used algorithms use one of two paradigms in sequence assembly, both of which use terms, notation and concepts from graph theory:

The Overlap-Layout-Consensus (OLC) starts by identifying pairs of reads that overlap sufficiently well. It then organizes every read as a node, and every overlap as an edge between node pairs. The sequence is inferred by walking Hamiltonian paths (paths visiting all the nodes exactly once) in the graph. A variation where the graph is simplified by removing redundant information such as transitive edges (i.e. edges that do not contribute to the reachability of each individual node) is called a *string graph*.

De Bruijn graphs model overlaps between substrings of each read, called a *k-mer* because the substring length is set to be k. Edges between two nodes indicate a k-1 length overlap between them. The sequence is inferred by walking Eulerian paths (paths visiting all edges exactly once) in the graph. De Bruijn graphs have been most successful in assembling short reads with very high accuracy, such as Illumina data.

The high memory requirements associated with running these algorithms on datasets whose size range into Terabytes has led to the demand for efficient data structures to reduce computational demands. In particular, the introduction of *FM-index* (143), which is a way of compressing text based on *Burrows-Wheeler transformation* (144), has led to large runtime improvements. The FM-index compresses text into a memory-efficient data structure that catalogues the number of occurrences of all recurring (and non-recurring) patterns in the text as well as the position of the patterns.

Despite huge advances in computational power and algorithmic improvements, the problem of optimal assembly is still considered "computationally intractable", meaning that a prohibitive amount of computer resources would be needed to guarantee an optimal solution. Most assembly software therefore work by heuristic principles rather than more rigorous approaches such as brute force algorithms, which select the best solution from all possible solutions.

2.2.2. Collapse in high-identity sequence fragments

As mentioned in chapter 2.1, a central problem in sequence assembly is dealing with repeats, which are genomic regions that are identical or highly similar to each other. Crucially, repeats that are longer than the effective read length (i.e. after paired-end information has been taken into account) cannot be traversed, breaking the assembly up into contigs (145).

During assembly, inadequate read lengths will in most algorithms lead to a collapse of identical or similar genome regions. If a sequenced genome G contains, say, three copies of a region A, all flanked by unresolvable repeat sequences, our assembly will be unable to decide if there are one, three or a hundred copies of A. (Figure: coverage). The algorithm may walk the cycle in the directed graph an arbitrary number of times, in which case our assembly will contain the same number of copies as the number of cycles walked. Alternatively, an algorithm may decide to walk the shortest possible path, or filter out more repetitive sequence before walking the graph. In these cases, our assembly contig will contain exactly one copy.

A related scenario occurs when aligning reads from G against a reference genome R. If G contains six copies of region A but R only contains one, a regular alignment algorithm will not be able to detect the duplications present in G.

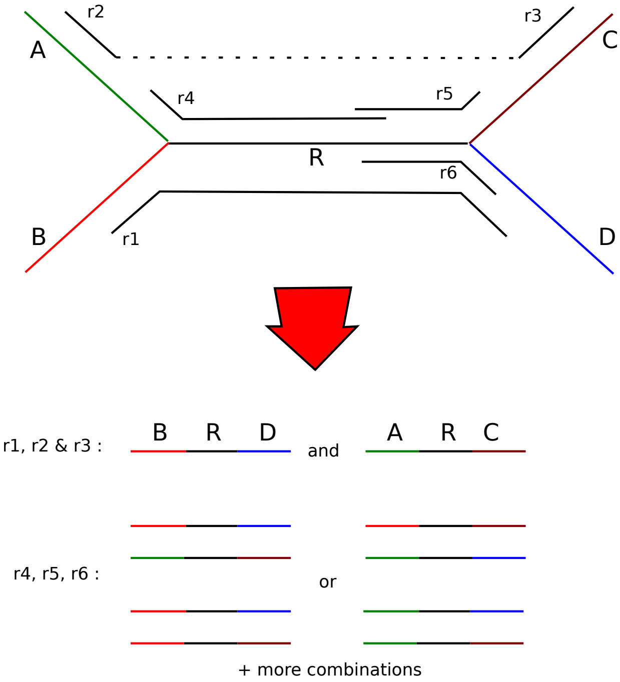


Figure 7. Importance of read length

A, B, C and D are unique parts of the genome, and R is a repeat structure between A and C and B and D, respectively. Read r1 is long enough to traverse the repeat, resolving the correct configuration. Paired reads r2 and r3 can also resolve the correct configuration. The remaining reads (r4, r5, r6) cannot determine the correct configuration, alone nor together. Source: Own work. Adapted from Nagarajan and Pop, 2013.

In both the previously described scenarios, we can gain additional information about the structure of G by investigating the number of reads mapped to each region.

2.2.3. Fold coverage

The word coverage is unfortunately used to refer to several very different but still related concepts. In the current thesis it is taken to mean the following: The absolute number of reads mapped to a particular nucleotide, *or* number of reads mapped to a genomic region times the read length, divided by the length of the genomic region. This is called the read depth or the fold coverage. This is slightly different from the theoretical coverage calculated before sequencing, which is the expected number of reads times the read length divided by the length of the sequenced target, because this latter metric is affected by unclonable regions of the target, sequencing errors and unmappable reads. As if to ensure total confusion, the word coverage is sometimes also applied to the percentage of a genome covered by reads, i.e. the total assembly length divided by the target length. When the word coverage is used in the present thesis it is never referring to this definition.

For a whole genome, fold coverage of 30 X means that each nucleotide has been sequenced an average of 30 times. For a specific nucleotide, it means that 30 reads added information about this nucleotide.

The concept of fold coverage is important for a number of reasons. First, high coverage is associated with increased confidence levels of the sequence being correct, and higher certainty when discovering variants such as *single nucleotide polymorphisms* (SNPs). Second, coverage is, on average, uniform across the target genome, which means that deviations can be used to predict errors in the assembly. It can be especially informative on any collapses in high-identity segments, as well as structural variants such as duplications, de-duplications and deletions (**Figure 8**).

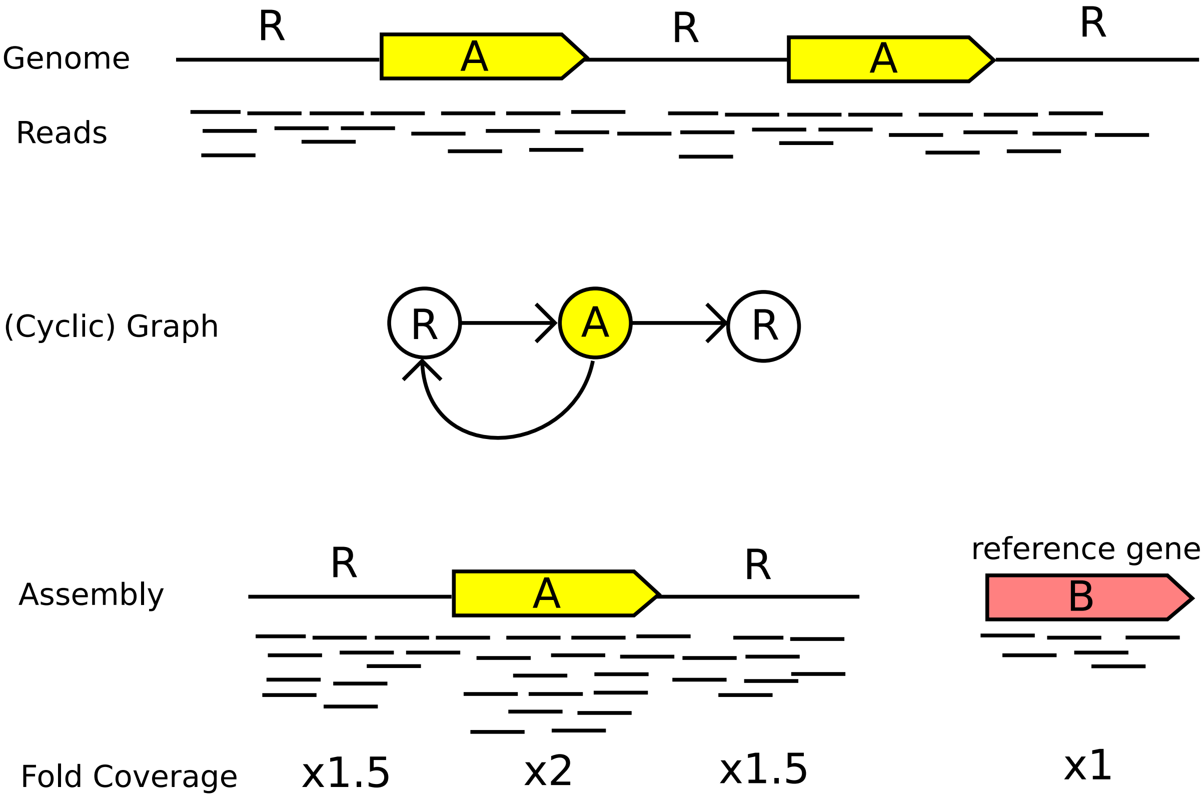


Figure 8. Coverage and structural variants

A genome containing two identical genes A with interspersed repeat structures R is sequenced. Regular graph assembly algorithms cannot resolve the correct structure because the cyclic graph can be traversed any number of times. After assembly or reference alignment, read coverage hints at the original number of copies when compared to a reference gene B whose copy number is definitely one. Source: Own work.

The concept of fold coverage is related to but very different from k-mer coverage. In De Bruijn graph algorithms, reads are broken into k-mers to provide *perfect coverage*, meaning k is chosen so that every possible substring of length k in the sequence is represented at least once (**Figure 9**).

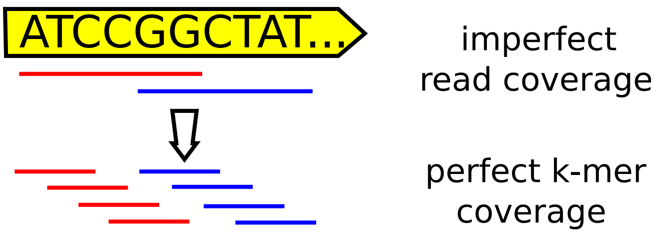


Figure 9. Perfect coverage

Reads can be broken up into smaller k-mers so that every subsequence of length k from the read is represented, called perfect coverage. This makes assemblies less memory-intensive, but potentially increases non-resolvable graph structures. Source: Own work.

2.3. Phylogenetics

Phylogenetics is the study of the evolutionary relationship between organisms. The term is distinguished slightly from phylogenomics, which is simply to signify that we are using information from the whole genome rather than a single gene, as was historically more common. In the present work I am always using information from multiple genomic sites rather than a single gene, but will still be using the terms interchangeably.

Unfortunately, we cannot use time machines to travel back in time to study evolutionary history in real-time. However, by comparing the similarities and differences between organisms or groups of organisms, we may deduce the most probable sequence of events that led to the current distribution of characteristics. In order to create such a *phylogenetic tree*, we will try to construct a graph where we position organisms more closely to each other the more intimately related they are.

The most immediate problem in constructing a phylogenetic tree is that we need some measure of the relatedness between organisms. Although there are many features we could use to compare bacteria, such as morphology or biochemical properties, a much more specific feature to use are the sequences of biological macromolecules such as DNA, RNA and proteins.

2.3.1. Sequence alignment

Sequence alignment is the process of manipulating two or more character strings through sorting, insertion and deletion of characters, so that identical or similar characters are lined up with each other. It is a central and well-studied problem in bioinformatics. A scoring matrix containing pairwise scores for character similarity (or penalties for character non-similarity) can help determine optimal alignments.

It is generally considered too computationally intensive to examine every possible alignment to see which is the highest scoring. Comparison of a small number of sequences use dynamic programming (solving sub-problems) approaches such as the Needleman-Wunsch (146) or Smith-Waterman (147) algorithms, both of which are guaranteed to find the best possible alignment. These approaches require computational resources that increase exponentially with the number of sequences compared. Therefore, when comparing many sequences, it is necessary to trade the optimality guarantee for speed, using heuristic algorithms such as BLAST (148).

Sequence alignments are valuable in phylogenetic studies because they follow the principle of Occam's razor: "What are the least improbable sequence of mutations that would have to occur in string A for it to become identical to string B?" This is a useful and parsimonious way of modeling evolution.

2.3.2. Phylogenetic trees

Generally speaking, the degree to which sequences differ from each other are related to the evolutionary distance between them. High identity suggests that the sequences diverged not too long ago, while the opposite is true when the sequences have many differences. Central to this is the assumption that the sequences *coalesce* (149), *i.e.* that they can be traced back to a (hypothetical) common ancestor.

A visual way of examining the internal relatedness of a group of isolates is through phylogenetic trees. A phylogenetic tree is basically a graph that states something about the relationships and distances of a number of sequences, where the sequences represent different organisms. The graph consists of edges, more commonly called branches, and nodes. End-nodes are called leaves, and usually correspond to the sequences that we observe. The tree can be *rooted*, in which case the graph is explicit about the direction of evolution. The root of the tree represents the *Most Recent Common Ancestor* (*MRCA)*, the term applied to the most recent individual organism from whom all organisms in the tree have descended. The term is also used with the same implied meaning in sub-trees. Alternatively, the tree is *unrooted*, in which case the tree does not make any implicit assumptions about the direction of descent.

The *molecular clock hypothesis* states that the difference between two related sequences is roughly proportional to the time since they diverged (150–152). However, in a phylogenetic tree the rates and effects of mutation and selection may vary, and this means that the relationship between sequence divergence and coalescence time is not necessarily constant in all branches of the tree. Models that allow for rate-variation between the branches of a tree are called *relaxed* molecular clock models.

Many different approaches exist for creating phylogenetic trees:

Algorithms based on *Neighbor-Joining (NJ)* (153) and *Unweighted Pair Group Method with Arithmetric mean (UPGMA)* (154) techniques rely instead on pre-computed distance-matrices that come from multiple sequence alignments. NJ trees are unrooted and can use relaxed molecular clocks, while UPGMA trees are rooted and require a constant-estimate molecular clock. Both are so-called greedy algorithms, meaning that they iteratively add an ancestor node to the two most closely related leaves, and thus build the tree directly rather than iterating over a range of possible topologies at each stage. They are not guaranteed to find the best-fitting tree.

Maximum parsimony (155,156) incorporates the idea of Occam's razor. Internal nodes are assigned in order to minimize the number of evolutionary events leading from an ancestor node to the observed state. Maximum parsimony algorithms evaluate many different trees and decide according to optimality criteria. However, it is essentially a heuristic approach and cannot guarantee that the output represents the optimal fitting tree.

Finally, we have the probabilistic methods *Maximum Likelihood (ML)* (157) and *Bayesian inference* (158). These more advanced models require *substitution models* that specify the expected rates and probabilities of different types of mutations. They are both rather computationally intensive. They allow different clock rates across different branches and sequence sites.

ML works by inferring tree parameters through maximization of the likelihood function, which is a function describing how likely the data (the alignment) is for a given tree T and substitution model **Ω**, Pr(D | T,**Ω**). Most implementations use heuristic approaches, exploring a subset of the many possible tree topologies and varying branch lengths to decide on the tree that maximizes the probability of the data.

Bayesian inference also uses the likelihood estimator, but additionally involve prior distributions on the branch lengths, substitution parameters and topology of a tree. The trees are then evaluated on its posterior probability. In order to avoid analytical calculation of the posterior probability, sampling methods such as *Markov Chain Monte Carlo (MCMC)* (159) algorithms are used instead.
In this framework, it is possible to explore a great number of tree configurations and parameters as well as different models of evolution, effectively searching for the most probable explanation given the data and the prior assumptions.

3. Aims

The overall aim of this project was to clarify genetic diversity within *R. salmoninarum* and to demonstrate the applicability of NGS technology as a multi-purpose tool in the study of animal pathogens.

Objectives

1. Resolve the phylogenetic relationships between diverse *R. salmoninarum* isolates, and if possible determine sub-population descriptive metadata correlations, assess risk of cross-species transfer and reconstruct the transmission history of BKD. (Paper I)
2. Catalogue mutations and determine variation patterns, particularly in the dominant virulence factors (Paper I, Paper III)

During the work with structural mutations, the need for a better method for detection of copy-number variation became clear and resulted in the implementation of such a method into open-source software accompanied by a full-length paper describing it (Paper II).

4. Materials and methods

4.1. Materials

The results and conclusions presented in the current body of work are almost exclusively based on whole-genome sequencing data. In early 2011 a collection of 68 *R. salmoninarum* isolates were sequenced using the Illumina HiSeq 2000 platform at The Genome Analysis Centre (TGAC), Norwich, UK. Isolates were selected from all international contributors' isolate banks with criteria chosen to maximize allowable inference space. The isolates represented the full temporal and spatial distribution from available data, with one isolate most likely first isolated by US fish disease researcher Ken Wolfe in the 1950s or early 1960s. Similarly, care was taken to ensure that multiple host species and capture habitats, as well as both wild and farmed fish, were represented in the data. In order to allow inferences about possible transmission between neighboring farms as well as to or from wild fish from local marine and river systems, a number of isolates of closely related origin were also included. A few isolates were included based on perceived novelties; Examples of this are the non-pathogenic isolate MT239 and the (anecdotally) high-pathogenic isolate Carson5b.

The following institutions contributed isolates for sequencing as part of this project: Centre for Environment, Fisheries and Aquaculture Science (CEFAS), UK; Norwegian Veterinary Institute (NVI), Norway; Marine Scotland, UK; United States Department of Agriculture (USDA), USA; National Oceanic and Atmospheric Administration (NOAA), USA.

Since these culture collections were founded and maintained independently, the available metadata information varied widely. Some isolates had likely been frozen, thawed, exchanged with other laboratories and subsequently re-plated and re-frozen or dried, and this information was largely unavailable. Similarly, although a majority of the samples were initially isolated from clinical BKD outbreaks, exact and standardized information about disease morbidity and mortality were in most cases absent. In order to present equal data richness for all isolates some information was therefore collapsed to the detail level of the least informative isolate.

A full description of the isolates included in this project can be found in Paper I.

4.2. Methods

Briefly, reads were paired, assessed for quality and pre-filtered. Quality score criteria were used to guide sequence alignment towards reference isolate ATCC33209. Non-aligning reads were assembled with a De Bruijn graph approach. All SNP sites with acceptable quality scores were used for phylogenetic inference. The most likely phylogenetic reconstruction indicated tree-like evolution without significant recombination. A consensus tree was constructed using bayesian inference. Because of the relative length of the longest branch compared to the rest of the tree the tree could uncontroversially be rooted at midpoint. Reconstruction of character states at internal nodes in the tree was done with both parsimony and ML models. A dated phylogeny was created using a relaxed molecular clock model.

All reference alignments were explored for evidence of gene copy number variation by examining variations in read coverage. The coverage signal was normalized for biases introduced by the Illumina sequencing protocol as well as the alignment procedure. Variations were called using HMM techniques, and confidence intervals for copy numbers were found using gene-wise bootstrapping of coverage observations. This functionality was integrated into CNOGpro, a tool specifically developed by the authors to detect CNV in prokaryotes.

By inspection of the phylogenetic tree the CNV status did not seem to follow simple inheritance patterns, and therefore the correlation between isolate pairs' CNV status to the patristic as well as the geographic distance between them was investigated using Mantel correlograms.

The methods are described in far more detail in the attached papers.

5. Summary of papers

Paper I

Microevolution of *Renibacterium salmoninarum*: Evidence for intercontinental dissemination associated with fish movements

We used WGS to generate genome-wide single-nucleotide polymorphism (SNP) data from 68 diverse *R. salmoninarum* isolates representing broad geographical and temporal ranges and different host species. Phylogenomic analysis robustly delineated two lineages (lineage 1 and lineage 2); furthermore, dating analysis estimated that the time to the most recent ancestor of all the isolates is 1239 years ago (95% credible interval 444–2720 years ago). Our data revealed the intercontinental spread of lineage 1 over the last century, concurrent with anthropogenic movement of live fish, feed and ova for aquaculture and stocking of recreational fisheries, whilst lineage 2 appeared to have been endemic in wild Eastern Atlantic salmonid stocks before commercial activity. SNP-based analyses allowed us to separate closely related isolates linked to neighboring fish farms, indicating that they formed part of single outbreaks. The main lineage 1 subgroup of *R. salmoninarum* isolated from Norway and the UK were found to represent an introduction to these areas 40 years ago.

Paper II

CNOGpro: Detection and quantification of CNVs in prokaryotic whole-genome sequencing data

Current tools and pipelines for the analysis of DNA copy number variation (CNV) have shortages that make them unattractive or intractable for prokaryote data, as most tools are designed specifically for diploid, human genomes. However, there are several idiosyncrasies in prokaryotic WGS data. Here we describe a step-by-step method for detection and quantification of copy number variants in prokaryotes specifically. We aligned WGS reads to a reference genome, counted reads in sliding windows and normalized counts for bias introduced by differences in GC content. We then investigated the coverage in two fundamentally different ways: (I) Employing a Hidden Markov Model, and (II) by repeated sampling with replacement (bootstrapping) on each individual gene. To demonstrate our method we applied it to real and simulated WGS data, and benchmarked it against two popular methods for CNV detection. We also presented the open-source software CNOGpro, written entirely in the R programming language.

Paper III

Identifying copy number variation of the dominant virulence factor *msa* within genomes of the fish pathogen *Renibacterium salmoninarum*

Here we used the coverage depth from genomic sequencing and real-time quantitative PCR to detect copy number variation (CNV) among the genes of *R. salmoninarum*. CNV was limited to the known dominant virulence factors *msa* and *p22*. Among 68 isolates representing the United Kingdom, Norway, and North America, the *msa* gene ranged from two to five identical copies and the *p22* gene ranged from one to five copies. CNV for these two genes co-occurred, suggesting they may be functionally linked. Isolates carrying CNV were phylogenetically restricted, and originated predominantly from sites in North America, rather than the United Kingdom or Norway. Although both phylogenetic relationship and geographic origin were found to correlate with CNV status, geographic origin was a much stronger predictor than phylogeny, suggesting a role for local selection pressures in the repeated emergence and maintenance of this trait.

6. Results and general discussion

Close to 80 years has passed since the initial discovery of *R. salmoninarum*, and yet many basic questions remained unanswered. In the following chapter I will describe my contribution to the body of knowledge on *R. salmoninarum*, followed by my contribution to the field of computational genomics. Towards the end of the chapter I will discuss methodological limitations and assess the impact of this work.

6.1. *R. salmoninarum* population structure

Our first and main objective was to resolve a high-resolution phylogenetic cladogram and attempt to quantify genomic variance levels within the genus. We reported in Paper I that *R. salmoninarum* is indeed a highly clonal bacterium, in agreement with all previous studies. Despite this low level of variation, there were clear schisms in the tree. We discovered that isolates belonged to either of two major lineages, which we uncreatively named lineage 1 and lineage 2 (**Figure 10**). Lineage 2 isolates represented only 10% of the collection, and were restricted geographically to Northern Europe and host-wise to the *Salmo* genus, while lineage 1 has a catholic distribution. We showed that lineage 1 could be further subdivided into 1A and 1B, the former of which, again, contained isolates from a wide range of places, while the latter strictly contained isolates of North American origin.

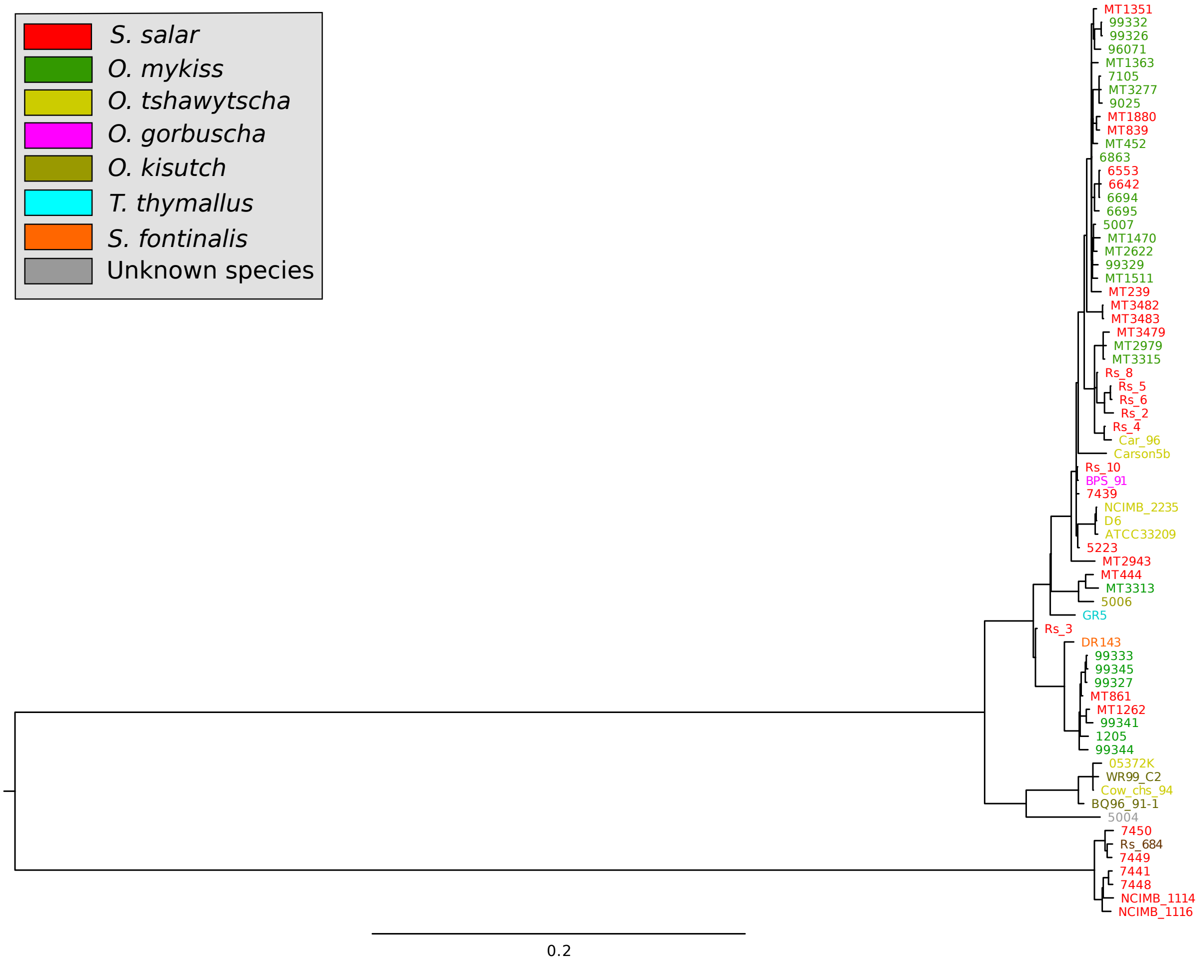


Figure 10. *R. salmoninarum* phylogeny

The phylogenetic tree of all 68 isolates, colored by host species. The major branch divides lineages 1 and 2, and the major branch of lineage 1 divides 1A and 1B. Reprinted from Brynildsrud *et al.*, 2014.

The inferred population structure agrees with previous publications that have studied *Renibacterium* phylogeny with methods such as intergenic spacer variation (ITS), and randomly amplified polymorphic DNA (RAPD) (122–127), although with a lot more discriminatory power. In the last stages of writing Paper I, I was made aware of a similar paper that used multilocus variable-number tandem repeats (VNTR) to discriminate *R. salmoninarum* isolates (160). Fortunately, these authors reported a phylogeny that was highly concordant to ours, although they report polytomies where we have resolved branches with 100% posterior probability support. Further, even though they have included both lineage 1 and lineage 2 isolates, they were unable to appreciate the full phylogenetic distance between these.

We have inferred the phylogenetic relationships between the major circulating lineages in Europe and North America. However, the tree must not be considered exhaustive. In fact, by looking at old *R. salmoninarum* typing studies it seems likely that our tree is hiding several long branches. As an example, Grayson *et al.* found in an ITS study most *R. salmoninarum* isolates to belong to the SV1 group, but with alternative genotypes of Japanese (SV2), Icelandic (SV2) and Northwest territory, Canadian (SV3) origin (125). We cannot say how and where these isolates would fit in with our phylogenetic tree, but we can infer from other studies such as Wiens and Dale (128), that SV1 corresponds to our lineage 1 (both 1A and 1B) while SV4 corresponds to lineage 2. Seeing as the full diversity range of 1A and 1B is collapsed to SV1, it seems likely that SV2 and SV3 isolates would represent rather long branches in our tree. This finding, if verified, could have significant impacts on some of the main conclusions that we have drawn. Instinctively, these relatively remote regions (especially NWT, Canada), have probably seen much less cross-clade interaction from both anthropogenic activities and natural migration events, thereby greatly reducing BKD transmission risk. They may represent unspoilt, long-time endemic strains of *R. salmoninarum*. This is theoretically supported by their apparent genetic deviance from lineage 1 and lineage 2, as the bacterium's rather slow rate of evolution does not normally allow such differences to emerge rapidly over short time frames.

6.2. General patterns of transmission

In our phylogenetic tree we saw tightly clustered isolates from different host species and in some cases from neighboring Atlantic salmon/rainbow trout farms. These observations answered the long-standing question of whether different host species infect each other with BKD: There are no genetic factors that suggest host-specific subtypes. Or at the very least, some subtypes very likely cross-infect between different host species. This observation has major implications. Sea- and river systems aquaculturists need to be aware of the risk of contagion from proximal farms even if they rear different *Salmonidae* species. Similarly, wild fish or even alternative vectors such as ectoparasites may indiscriminately introduce BKD to naïve farms. We cannot say how important cross-species BKD transmission is, only that it is possible. To some extent, this finding contradicts recent risk evaluations that do not consider co-localization of Atlantic salmon and rainbow trout as especially important (72,74). As a final note, our findings kill the hypothesis that pathogen-related factors are the major source for the observed variations in host species outbreak mortalities (24). The apparent susceptibility of many Pacific salmon species must stem from host- or environment factors or a combination of host-environment-pathogen interactions.

In Paper I we further show that near-identical isolates can be harvested from neighboring farms rearing different species, both in near-concurrent disease outbreaks and years apart. This suggests that strains can sometimes become endemic to an area. However we do not know the exact mechanism of establishment, as *R. salmoninarum* is not thought to live long in free water masses or sediment, and the role of wild fish and bivalve vectors is unclear. Hydrodynamic contamination is only thought to be important between sites on the same farm. Other vectors are likely involved in inter-farm disease transmission.

This is all testimony to something that has long been known: That there is a need for tight biosecurity, sensitive screening and compartmentalization in fish farms. The increasingly vertical integration and unidirectional flow of material that has gradually been implemented in the last few decades have undoubtedly contributed to the decreased impact of BKD in Northern Europe. Proper egg treatment and screening and hatchery placement on land or in isolated freshwater environments should continue to diminish its importance.

6.3. Reconstructing BKD transmission history

A major objective of Paper I was to reconstruct the modern transmission history of BKD, as has been performed for several major human pathogens (161–163) and notably the aquatic pathogen Infectious Salmon Anemia Virus (164).

First, to be able to make inferences about evolutionary direction we needed to root the tree, which can sometimes be a complicated issue. Fortunately, the comparatively long branch between lineage 1 and lineage 2 made it possible to uncontroversially use midpoint rooting on this branch, as it would require extreme violations of any clock-like evolution for this to be erroneous. This allowed us to make direct evolutionary inferences.

We discovered that North American isolates had a strong tendency to appear basally to sub-lineage clusters, suggesting North American origin for these clusters. Next, we used parsimony and likelihood methods to reconstruct the character states at internal nodes in the phylogenetic tree, equivalent to finding the most likely transmission pattern of our isolate collection. This analysis corroborated our initial finding of a North American origin for many of the lineage 1A clusters, but was inconclusive with regards to lineage 1B and 2 origins.

In order to more thoroughly investigate this more thoroughly, we attempted to estimate the effective population size of the largest 1A sub-population. We found evidence for a significant population expansion posterior to the initial state switch from North America to UK, hinting at a clonal expansion following importation of this lineage into the UK.

Finally, we attempted to label all internal MRCAs in the tree according to how long ago they existed. According to our results, lineage 1 and 2 split between 444 and 2720 years ago, while 1A and 1B split some time between 150 to 700 years ago. Both splits predate modern aquaculture. Looking at the most credible time to MRCA for some smaller also helped us identify at least two modern clade introductions with direction from North America to Europe within the last century.

There is a cautionary tale to be found in these results. The observed time-correlation between geographic switch and subsequent population expansion to known periods of Western aquaculture expansion is not exclusive to *R. salmoninarum* but extends what has been suspected and/or demonstrated for other aquaculture pathogens such as A. *salmonicida* ssp. *salmonicida* (132) (Garcia et al, 2000) and *Yersinia ruckeri* (133,165).

We concluded that anthropogenic activities significantly contributed to the proliferation of BKD over the last century-and-a-half, but noted the possibility that region-specific lineages were already endemic to recipient areas prior to the dawn of modern aquaculture in the latter 19th century. As evidence for the latter we cited the limited geographical distribution of lineage 2, an increased inter-strain variability in this lineage inconsistent with recent emergence, and the fact that BKD was first observed in Scottish rivers Dee and Spey. We consider it probable that these outbreaks were from strains related to our lineage 2 strains (specifically NCIMB1114 and NCIMB1116 from river Dee). Furthermore, the high incidence of BKD outbreaks in Norway during the 1980s, chiefly of lineage 2 origins, came reportedly after establishment of brood stock from capture of wild Atlantic salmon. Since the mid 1990s outbreaks have been sporadic and of lineage 1 origin. It is curious that clinical BKD was nearly eradicated in Norway with the establishment of highly effective screening, vertical integration and strict biosecurity. The origin of lineage 1 is unclear. It may be interesting to note that the split between the major lineages 1200 years ago clearly predate anthropogenic spread and subsequent population isolation. One hypothesis states that *R. salmoninarum* must have co-evolved with Atlantic salmon longer than with Pacific salmon, which would help explain the much higher innate resistance against BKD in Atlantic salmon. (However, this hypothesis fails to address the Pacific species that have the highest innate resistance - rainbow trout). We will thus refrain from speculation on the evolutionary origin of lineage 1 and limit ourselves to conclusion about transmission rates in the last few hundred years.

We consider the above conclusion to be most parsimonious from our data. However, we cannot rule out the possibility that reverse transmission rates, i.e. from Europe to North America and elsewhere, have been important as well. Quintessentially, such inferences are limited by isolate sampling, as it is obviously impossible to sample every possible genotype, and we therefore have to accept imperfect phylogenetic trees. In our case, we find evidence of transmission rates from North America to Europe, which seems true for *these* isolates, but we are aware of some central caveats. Firstly, we have uneven sample origin representation. It is possible that we could have detected North American isolates that were descendants of European ones had we sampled more extensively from that region. Secondly, we have collapsed detailed geographic information into unspecific tokens "USA", "Canada", UK" and "Norway". This may be unfair towards the North American countries, since they are many times larger than the comparatively small North Atlantic region homing UK and Norway. Consequently, higher variance is expected in North America due to more varied environments and selection pressures, and selection will bias the phylogenetic tree. It might also be that rather small subdivisions of these countries represent the true epicenter of dissemination, rather than the nations in their entirety.

Given *R. salmoninarum*'s apparently slow rate of evolution it is highly unlikely that we are observing saturation or convergence in our sequences, and *horizontal gene transfer* (HGT) (See chapter 6.6) seems to be a very rare event. We found that the possibility of a limited degree of recombination could not be excluded, but that it was not likely to significantly influence our results. Selection is the major concern. We do not really know much about site-variant selection pressure but assume that it is insignificant. One supportive finding for this is our observation that SNPs are evenly distributed across the entire genome, with no particular hotspots or highly constrained regions.

These models are all based on a set of assumptions, and it is probably impossible to prove that they are not violated. In our analysis we have tried to relax prior assumptions as much as possible, setting conservative, relaxed or non-informative priors where no plausible information were available. However, in some analyses the assumptions are definitely violated, such as the assumption of no population structure in the estimation of effective sample size. However, it is impossible to reliably quantify how much the assumptions are off, and so it is not clear exactly how this impacts results. To summarize, although our methods and results come with imperfect precision and potential "chinks in the armor", the following words from English writer Douglas Adams come to mind: "*If it looks like a duck, and quacks like a duck, we have at least to consider the possibility that we have a small bird of the family Anatidae on our hands* (166)."

6.4. Comparative genomics

The high genetic homogeneity of the *Renibacterium* genus has been known since the mid-1980s (167). However, most taxonomic and typing studies have relied on fingerprinting of a few genomic markers and thus have not been able to quantify how diverse the bacteria really are with sufficient precision. In Paper I we disambiguated the homogeneity question by reporting the differences between major lineages down to SNP differences between individual strains. We found a mutation number as low as one SNP per 876 bases, or 3.600 SNPs in total in our collection of 68 isolates.

Unfortunately a lack of detailed metadata meant that we were not able to perform genome-wide association studies (GWAS), as was initially planned. We were however able to map previously known mutations with important phenotype implications to specific lineages. As an example, we found that the *msa*.Ala139Glu mutation, associated with enhanced MSA binding to Chinook leukocytes (37) was unique to and ubiquitous throughout lineage 2. The clinical importance of this increased agglutination activity in lineage 2 is unknown.

Part of our original plan was to assemble non-reference sequence to explore the pan-genome of the *Renibacterium* genus. It was therefore baffling to find that the gene content was exactly the same in all our isolates. Or more precisely, not a single isolate had a single non-homologue gene that was not present in ATCC33209. We were struck with the uncomfortable realization that a *Renibacterium* comparative genomics paper would become dreadfully boring, and this conception would likely have persisted had we not lucked into the following finding: Although non-homologue gene content appears to be constant, some strains have multiple identical or near-identical copies of certain genes.

6.4.1. Copy number variation

In this study, we discovered that although gene content did not appear to vary between *R. salmoninarum* strains, some strains have multiple copies of genes, some of which are involved in pathogenicity. This phenomenon is called copy-number variation (CNV), and is an incompletely described mechanism for phenotypic variation in prokaryotes. However, gene duplication events can be advantageous over both short and long time frames (168–170). CNV seems to be a dominant strain variation factor for *R. salmoninarum*.

We discovered CNV in the *msa* and *p22* genes, with copy numbers ranging from one (two for *msa*) to five. These paralogs appear to be completely identical to one another, suggesting recent duplication events or functional constraints. Furthermore, CNV in these factors were co-occurring. Duplication of either *msa* or *p22* without the other was not observed. This could indicate a common duplication mechanism such as genetic linkage or co-localization on a plasmid or phage. Alternatively, it could indicate some form of mutualism, where duplication of one factor is redundant without the other. The latter would perhaps make sense, since the MSA and p22 proteins have similar functions and together constitute most of the membrane-associated protein of the species. It is unknown if the two operate together or independently though. It is likely that this question could be answered using pathogenicity screening profiles on gene knockout and knock-in clones.

Because of the many interspersed identical repeats in the *R. salmoninarum* genome, we were unable to determine the genomic location and orientation of the *msa* and *p22* paralogs. Although this could possibly be answered with blotting techniques such as Southern blotting, a more surefire approach would be to supplement our WGS reads with long-read sequencing technology such as PacBio, Nanopore, or Illumina's mate-pair protocol.

Similarly, it is unknown whether these extra copies, like their reference counterparts, are integrated in the main chromosome or associated with plasmids, nor the mechanics of their duplication (or de-duplication). IS sequences and transposases flanking the reference genes suggest that a homologous recombination-type expansion is likely, but on the other hand (partially degraded) conjugation-related genes are also abundant and could point to HGT as a dominant force.

We confirmed previous reports (36,128) that CNV seems to be far more common in North American isolates compared to European, and particularly UK, isolates. Furthermore, we found that CNV was not constricted to a monophyletic group, but present in seemingly distantly related isolates. This seems to point towards multiple independent introductions of the trait rather than a single duplication event at one point in history. In fact, we provided metrics that suggested geographic origin was far more correlated to CNV status than was phylogenetic information. We concluded that this points to local selection pressures, chiefly in North America, as the main drive in CNV emergence and maintenance.

Unfortunately, we have not yet researched the phenotypic effects of CNV in the *msa* and *p22* genes. Still, it does not seem like a huge leap to believe that there are tangible effects.

Firstly, this trait has emerged multiple times and appears to be actively maintained in some populations. This signals that extra copies are not redundant at all. After all, it was already known that

R. salmoninarum had *two* identical transcriptionally active copies of the *msa* gene. Although the two ORFs are identical, we discovered a perfect 91 bp rho-independent terminator-encoding palindrome at the 3' with a polymorphism (between the two ORFs) in the loop region. This could at the very least indicate differential regulation of the two copies. Additional copies may allow fine-tuning of protein output and, in the absence of tightly controlled negative feedback loops, provide a quantitative boost to protein output (168). Over longer time frames, auxiliary copies of essential genes are free to accumulate mutations and, in time, evolve into new genes with new functions (171).

Secondly, (below a threshold inoculum dose) a virulence-increase phenotype has already been shown for *msa3*-positive isolates (36). It does not seem like a stretch to extrapolate this finding to >3 *msa* copies.

Nevertheless, more research is needed to definitely determine the effects of these mutations, and future infection trials may want to address this.

6.5. CNOGpro

A major challenge in this project was the detection and description of CNVs from WGS data. Although it appeared at first glance to be a relatively simple problem, careful study of the available methods and tools revealed substantial shortcomings that meant they would not properly handle our particular problem. Existing tools' crude customizability and lack of support for haploid organisms were particularly relevant for us. It became clear that we needed to design, implement and publish our own method in order to adequately answer the questions we wanted.

We developed and published a tool called CNOGpro, which is an acronym for "*Copy numbers of genes in prokaryotes*." The tool was written entirely in the R programming language, and is available as open-source software at the Comprehensive R Archive Network (CRAN - cran.r-project.org).

CNOGpro uses read coverage to call CNVs. It requires an annotated reference sequence for use as a backbone in sequence alignment. The alignment process needs to be completed by other software, but CNOGpro imports read alignment coordinates and use this to calculate read densities in sliding windows. This count is the central parameter that allows copy number inferences to be made. The program further parses counts according to annotated elements (typically genes and rRNA/tRNAs) from the reference file, and makes CNV inferences element-wise. This gene-wise parsing bypasses the complex problem of signal segmentation (breaking the coverage signal into segments of equal copy number value), which can be heavily influenced by the many biases affecting coverage. Our program corrects bias resulting from unequal GC-content, which is the quantitatively most important bias on the most common sequencing platforms (172–174). There are other important biases that affect read coverage, such as batch effects, genomic mappability (low-complexity regions have fewer reads mapped to them), replication cycle skew (some chromosomes are replicating, meaning that genes closer to the origin have higher copy numbers on average) and sequencing probe GC-content. We consider our method of GC-normalization and subsequent parsing of read counts into windows defined by gene boundaries to avoid major systematic errors, i.e. count is an unbiased estimator of the true copy number.

CNOGpro determines copy number in two different and non-interacting ways: (i) By allowing the coverage probabilities to vary according to overdistributed poisson functions with nested means, using Hidden Markov Models (HMMs) to determine copy numbers. (ii) By performing bootstrap iterations on coverage observations that have been binned gene-wise, using the resultant percentiles to form confidence intervals for the true copy number. In order to draw from the strengths of both methods while mitigating weaknesses, we strongly encourage integrated human interpretation of results.

CNOGpro is no magic black box. Both input and output require some degree of manual curation, and commands and parameters must be soundly chosen based on biological as well as computational proficiency. It is especially dependent on high-quality alignments with sensible choices of parameters. While it is, for example, clear that the use of a distantly related reference organism will introduce bias and be detrimental to results (175), it is rather hard to draw the line between a sufficiently and insufficiently close relation. Low-complexity and repeated regions in the genome can distort some of the central assumptions of our program, such as the assumption that coverage is approximately equal across equal-copy number regions. CNOGpro integrates a number of assumptions that are usually true, but not always, such as the assumption that most genes in a chromosome have copy number equal to one. Furthermore, CNOGpro might be vulnerable against unreasonable parameter settings. The probabilities of changing copy number states, *i.e.* the transition matrix in the HMM, must be set manually. From our own testing, results seem to be quite robust, but we can obviously not guarantee that this is the case for every possible dataset. It is possible that a parameter estimation method such as for example the Baum-Welch or segmental K-means algorithms (176) would better insulate the results from haphazard settings.

We tested CNOGpro on real WGS data and data simulated with ART (177), the latter manipulated by us so that the size, location, orientation and multiplicity of all structural mutations were known to us. Under both these scenarios, our method performed favorably to existing CNV detection tools cn.MOPS (178) and CNV-seq (179). These tools were chosen based on their perceived high esteem (personal observation) and broad applicability. Unlike us, their applicability is not restricted to the prokaryotic kingdom. Furthermore, both have individual strengths which were not considered by us: The sensitivity and specificity of cn.MOPS increases with the number of input samples; we used only one at a time. CNV-seq is very fast; we did not consider runtimes in our comparison. CNV-seq also lacks innate CNV quantification methods. It is primarily a tool for CNV detection that gives associated p-values and 2-log-transformed signal intensities. It is highly likely that we could have created alternative setups where these tools had outperformed CNOGpro, which is why we limit or superiority claim to a narrow set of conditions while noting that we have not tested alternate setups thoroughly enough.

Regardless, I believe CNOGpro is a valuable addition to the bioinformatician's toolbox. It was created out of frustration with not being able to do what I wanted to do with my data using existing tools. One of the most common problems I ran into was outdated compatibility or circular library dependencies. It was therefore one of my main goals to create a tool that would be able to do the same job in five or ten years as it could now. CNOGpro has no specific library requirements, and currently only uses methods from one other R package. (We plan to write our own standalone version of this method for an upcoming version.) The GenBank format and the R programming language are not likely to change radically in the near future. New paradigms in WGS technology can of course make our method irrelevant, but at the time of writing short-read technology have excellent prospects for years to come.

A more detailed description and discussion of our method is found in Paper II.

6.6. Methodological considerations

Even if bioinformatic methods have spearheaded great scientific breakthroughs, careful attention must be attuned to the caveats of these methods, lest we might end up in a situation where the famous quote has evolved into "Lies, damned lies and bioinformatics."

Like most scientific disciplines, the obtained results and conclusions may vary depending on sampling, including which organisms, isolates, sequences and macromolecule sites are used for comparison. However, phylogenetics have traditionally not been bound by equally strict requirements of independent observations as are found in many other statistical sciences. Inferences are often limited to those isolates, sequences and sequence sites observed in the experiment, rather than generalized as being valid for entire populations, and sampling breadth and strength is therefore more relevant to the precision or completeness of the study rather than for avoiding bias for the estimation of population characteristics. (If population parameter values are estimated however, non-representative sampling becomes an issue. An example is estimating the mutation rate of a species; this could vary between branches, genes and sequence sites and the population may be highly structured or of unequal size.) In fact, extensive sampling of related isolates is generally considered a good thing as it tends to improve phylogenetic accuracy (180,181), and incomplete representation of all taxa in a phylogenetic study does not systematically skew results (182). This means that even though a large fraction of our samples are phylogenetically clustered isolates of UK origin and not one sample hails from Asia, our inferred phylogenetic tree is not wrong, just incomplete.

The methods I have used assume that sequences evolve in a tree-like manner, that there is little to no *recombination*. This may not always be the case in bacteria, which can be quite promiscuous. Non-vertical transfer of genetic material in bacteria is called *horizontal gene transfer* (*HGT*) and includes transformation (uptake of free genetic material from the surroundings), transduction (DNA transfer by phage vectors) and conjugation (transfer of DNA between bacteria via a plasmid). Possible recombination events are typically ignored in phylogenetic studies, which may lead the researcher to draw incorrect conclusions about the evolution of the sequence, especially in the case of recent HGT events (183). A number of techniques to detect recombination have been suggested, such as examining discordance between evolutionary trees inferred from different genes or loci, or phylogenetic network theory (184), but the usefulness of these methods is debated. In Paper I, our analyses indicated an almost complete absence of recombination in lineage 1, and a limited but insignificant degree of recombination in lineage 2.

Another thing that can hamper inferences through computational phylogenetics is homoplasy or convergent evolution, which means that characters are evolving to become more rather than less similar. This will cause sequences to appear to be more closely related to each other than they really are in a phylogenetic tree, a phenomenon called *long branch attraction (LBA)*. A further exacerbation of this problem occurs between distantly related sequences, because in DNA/RNA/protein analyses each character can take a limited number of different states. As time passes the probability of reverse mutations, i.e. a character mutating back to its original state, increases. Over long evolutionary time frames reverse mutations may significantly reduce our ability to predict the age of a tree bifurcation, a phenomenon called *saturation*. *R. salmoninarum* has a slow mutation rate, an age of the MRCA that was found to be only around 1200 years, as we have sequenced the full 3.15 MBp genome. If we assume that all (or at least more than a handful of) sites in the genome are free to mutate, this means that saturation is completely irrelevant in this study.

Since we are mostly using heuristic algorithms for phylogenetic inference, we cannot guarantee that there are no trees that suit our data better than the one we have found. Even extremely commonly used tools such as the BLAST algorithm can fail to detect sequence homology and thereby misinform evolutionary analysis. This will improve as computational power becomes more readily and cheaply available concurrent with algorithmic improvements that can make these problems more tractable as data continue to grow.

There is also hope that future improvements in sequencing technology itself can alleviate many of the current problems with sequence assembly and analysis. As mentioned, longer reads are good since it reduces assembly complexity increases contig lengths. In addition to read lengths, technologies should improve on error rates, biases, affordability as well as physical size of the sequencing instrument. An optimal technology would not require the genome to be shotgunned at all, but rather read an entire chromosome as a single circular read. Single molecule technologies such as nanopore sequencing have in fact been under development for at least two decades (185), but unfortunately still have not reached the market.

6.7. Impact of work

Knowledge about *R. salmoninarum* has been slow to accumulate, partly owing to the bacterium's exceedingly slow growth and fastidious nature. This has caused many researchers to lose their patience and move on to other, less particular, organisms (L. D. Rhodes, personal communication). The decline of BKD's relevance in Europe has also shifted attention towards other important diseases in salmon aquaculture. However, BKD is still a looming threat to salmonids worldwide, and, in the words of A. J. Evenden, an "unfinished jigsaw (22)."

Many of the frustrations of growing *R. salmoninarum* can be avoided by corroborating or substituting lab experiments with *in silico* experiments. This project's application of WGS to a wide and representative set of strains has allowed for exploration of many hypotheses that would have been hard to answer with other methods. These data are now freely available through the European Bioinformatics Institute's sequence read archive, allowing other scientists to continue the work that we have started.

This project has significantly contributed to knowledge about *R. salmoninarum*. In recent years there have been many case reports and risk evaluation studies, but the genomics of *R. salmoninarum* has been largely overlooked. In Paper I and Paper III we greatly expand current knowledge of phylogeny and comparative genomics for this bacterium, and answer long-standing questions in the field such as how limited the worldwide genetic diversity really is. Our finding of free between-host transmission and opportunities for long-distance transfer should have implications for aquaculturists and policy makers. The deregulation of BKD that has taken place in UK trout culture, for example, must be viewed as a risky move in light of these results. A worst-case (or at the very least, a bad-case) scenario would be the establishment of a dissemination pattern similar to that seen in North America, where BKD seems to be propagating at undiminished rates through wild fish. Eradication of BKD in the US is becoming less likely as the disease prevalence increases and new, hitherto thought unsusceptible, species are infected. Our observation of free-species transfer and the potential for enzootic establishment has major implications for any eventual eradication or prevalence reduction programs. Furthermore, increased-virulence biotypes circulating and perhaps also originating in North America makes the situation here even more precarious. Add this to larger landmasses, a more thoroughly infected wild fish population, complex sub-national legislation, low innate Pacific salmon resistance (except *O. mykiss*), and in some cases non-transparent networks of trade, and the North American situation looks even harder. A more thorough understanding of the disease is the foundation in making this better.

Sometimes it is possible to control outbreaks without fully understanding disease though, as John Snow infamously proved when he removed the pump handle from the *Vibrio cholerae*-containing well on Broad Street (186). In this project we uncovered evidence of long-term geographic structuring with several instances of more recent transnational and transatlantic spread. The stricter biosecurity measures implemented in Norway around the early-mid 1990s were likely responsible for a sharp decline in BKD outbreaks around the same time. Our results provide support to the idea that lineage 2 outbreaks were avoided after the practice of using captured fish as broodstock was halted.

There seems to be reason to discriminate between BKD outbreaks from different biotypes. Finding a lineage 2 strain in North America would be surprising and possibly indicate biosecurity breach through imports. Also, all screened lineage 2 strains have the *msa*.Ala139Glu mutation that is associated with *in vitro* increased agglutination of leukocytes and erythrocytes. However, this must be weighed against the possibly higher pathogenicity of strains with CNV in *msa*. All this will be relevant information that can be easily gained from typing efforts.

Of course, *R. salmoninarum* is just one of many important pathogens. The present effort can also be viewed as simply a demonstration of the utility of WGS-directed bioinformatics for high-resolution epidemiology and comparative genomics efforts of important pathogens. Such studies have been done for many human pathogens such as MRSA (161), *V. cholerae* (162) and *Shigella sonnei* (163), but we were the first to pioneer this for an aquaculture pathogen. I predict many similar studies in the future as sequencing costs continue to drop, methodology becomes gradually more streamlined and user-friendly, and the global importance of aquaculture increases. Our methods could be generalizable to other bacterial pathogens. As previously mentioned, *R. salmoninarum* is not unique in having a dissemination history interwoven with anthropogenic activity; evidence exists for similar dissemination of other aquaculture pathogens as well. The results are therefore to some degree generalizable to other diseases, and could inform future policies on trade, farm placement, different species co-habitation, disinfection of affluent and effluent water, transport, and slaughter considerations.

The observation that *R. salmoninarum* contains up to five copies of seemingly identical genes has interest beyond the world of aquaculture microbiology, as this appears to be quite rare. Further studies are required to determine the significance of these and other duplications and the evolutionary pressures driving them.

It is too early to say anything meaningful about the impact of our CNV-calling method (Paper II). Our main contributions in this field were: (i) The overdispersed Poisson parameterization of coverage probabilities under different copy number tokens in the HMM. Many other methods used overdispersed Poisson models as well (175,178,187), but not in conjunction with HMM. (ii) Using a bootstrap approach to strengthen the HMM predictions and create copy number confidence intervals. Although this does not represent a huge theoretical leap, to the best of my knowledge our method is the first to provide meaningful statistics of copy number confidence. (iii) Replacement of segmentation algorithms with gene-wise coverage tallies, mitigating specificity problems resulting from incomplete bias correction. (iv) The streamlining of all necessary steps into a widely available and (almost) self-sufficient R-package, specifically engineered for haploid/prokaryotic organisms. I hope the accessibility of our method and the novelty of our results will inspire new interest in the study of copy-number mutations in bacteria, as so far the surface has just barely been scratched.

7. Main conclusions

The full genomes of *R. salmoninarum* isolates from a wide distribution of hosts, habitats and spatio-temporal environments were sequenced and assembled, and the genomic variance and average mutation rate was quantified.

Phylogenetically, strains belonged to either of two major lineages, lineage 1 and lineage 2. The majority of isolates belonged to lineage 1. Lineage 2 appeared to be restricted to the North/East Atlantic Ocean, while lineage 1 had a cosmopolitan distribution. Data revealed long-term geographical structuring. Lineage 1 and 2 diverged 444-2720 years ago, prior to the dawn of aquaculture, suggesting historical allopatry.

Polytomies in the tree pointed to local transmission dynamics as important for isolates from neighboring farms. Such polytomies did in several cases include isolates from different host species, suggesting free transmission between different species unhindered by high tropism. In some cases isolates from rainbow trout and Atlantic salmon were indistinguishable, which in our opinion advocate stronger restrictions on the co-localization of these species.

A most parsimonious transmission dynamic in our phylogenetic reconstruction indicated North American ancestry of major lineage 1 sub-trees with predominantly European isolates, and molecular clock analyses indicated that these character switches occurred within the last century-and-a-half. This appears to have been a period of great diversification and population founding within the genus. Lineage 1 isolates demonstrated free transfer between the Pacific and Atlantic Oceans, in spite of the slim chances of populations from these oceans interacting naturally. Lineage 2 on the other hand appeared to have been endemic in North/East Atlantic Ocean populations for a long time. The above results were indicative of anthropogenic involvement in the dissemination history of *R. salmoninarum*. Trade and movement of live fish, ova and unpasteurized feed for aquaculture and angling have afforded plenty of opportunity for long-distance dispersion outside of normal fish migration patterns.

R. salmoninarum was found to have a very clonal, constant genome. Idiosyncratically, its pan-genome was in this study equal to its core, and copy-number variation (CNV) appeared to be the main source of gene content variation. Several lineage 1 isolates had CNV in dominant virulence factor genes *msa* and *p22*. The CNV mutation is probably associated with an increased-virulence phenotype. The mutation was found in roughly half of our North American isolates, and in one Norwegian isolate out of twelve, but was not seen in any of the thirty-five isolates from the UK. Analysis indicated repeated emergence of the CNV trait due to local selection pressures, most importantly in North America.

In order to properly detect and count CNV mutations from WGS data of haploid organisms we had to design and implement our own method, implemented as open-source software. Our method compared favorably to existing programs on test data. The method and software were described in a full-length paper, and represents advancement in the methodology for studying CNV in prokaryotes.

8. Future perspectives

We have characterized the major lineages of *R. salmoninarum* circulating in North America and North-Western Europe, but from previous studies (125) we can infer that our isolates do not encompass the full diversity of the species. It should be clarified how isolates from other BKD-infected regions of the world relate to the lineages we have described, as well as their particular characteristics of biochemical and antigenic nature and antibiotic resistance-profiles. Isolates from Iceland, Japan, Greenland and the Northwest Territories of Canada are particularly interesting, as these locations are known grounds for phylogenetic variants not described in this thesis, but other regions of the world such as South America and the Mediterranean Sea may be as diverse and important for full phylogenetic representation and accuracy.

Relevant differences between strains and lineages need to be addressed methodically. Infection trials would of course be the ultimate method to establish strain differences in pathogenicity, but the number of fish needed for proper independent multi-strain trials combined with the long prepatent period, protracted disease, and possibly miniscule differences means such trials are expensive and raise significant animal welfare issues. Efforts to develop alternative *in vivo* models such as zebrafish (*Danio rerio*) to study BKD have unfortunately not been very successful (188). Thus, a major objective should be the development of good experimental models for the study of BKD pathogenesis and strain pathogenicity differences.

Strains with CNV in *msa* and *p22* are especially good candidates for such experiments, as three copies of *msa* (as opposed to the normal two), have already been proven to have clinical significance (36). It would be useful to establish whether these extra copies are transcriptionally active under *in vivo* conditions, under what conditions the total protein output is increased, and whether or not they are individually regulated. For this it would presumably be helpful to know the genomic locations and regulatory regions of these extra copies, which, again, could probably be determined by sequencing with sufficiently long reads.

We have shown unequivocally that strains infecting Atlantic salmon and rainbow trout are in some cases genetically indistinguishable. This means that host factors must be largely responsible for the observed differences in mortality between these two species. If these host factors become better known it is possible that breeding (and in the future possibly cloning) efforts could help culturists in developing fish with a higher degree of innate resistance towards BKD.

The currently accepted hypothesis is that *R. salmoninarum* evolved by reductive evolution from an *Arthrobacter* ancestor (119). The specifics of how, when, where and why this happened however is unclear. In addition, some of the most interesting components in the *R. salmoninarum* genome such as for instance the *msa* gene do not have any known homologues elsewhere in nature. It seems unlikely that such a highly specialized protein evolved from nothing. Equally mystifying is the fact that there are no other known species in the *Renibacterium* genus or any "missing link" between *Renibacterium* and *Arthrobacter*, despite the global distribution and still very clonal nature of the former. It would be interesting to explore the macro-scale evolutionary history of the bacterium and, in particular, some of its most interesting genes.

Major unsolved mysteries still surround the pathogenesis of BKD. In this regard the ability to enter and subsequently survive and replicate within various types of phagocytic cells, some of whose specific job description is to incapacitate and kill invading microbes, is especially interesting, as this highly specialized and characteristic function of *R. salmoninarum* is incompletely described.

Further studies are needed to determine the impact BKD might have in the future. Globalization itself is probably irrelevant for BKD, but increased global temperatures could have wide-ranging consequences. It is well established that there is an association between increased temperature and BKD-related mortality. Increased temperatures could also drastically affect patterns of salmon and other fish migration, viability of water systems for aquaculture, and through increased ice melting and altered precipitation patterns probably even the landscapes themselves.

To my knowledge, no research has ever quantified the global costs of BKD. The importance of the disease is not purely economical; it also has social, environmental and animal welfare implications. Good studies documenting these costs could help shape future policies related to combating the disease.

Finally, the methodology outlined in the current thesis and the attached papers can and should be used for studies of other animal or human infections. The high-resolution phylogenetic information makes it possible to rapidly detect outbreak sources at micro- and macro-scales, as evidenced by their ability to detect person-by-person transmission within a hospital (161,189) as well as population-scale dissemination across continents throughout history (190).

9. References

1. Food and Agricultural Organization of the United Nations. The state of world fisheries and aquaculture - Opportunities and challenges. Rome; 2014.

2. Halverson A. An entirely synthetic fish: How rainbow trout beguiled America and overran the world. New Haven, CT, USA: Yale University Press; 2011.

3. Veterinary Medicine Advisory Committee to the US FDA. An overview of Atlantic salmon, its natural history, aquaculture, and genetic engineering [Internet]. [updated 2015 Mar 20 cited 2015 Apr 23]. Available from http://www.fda.gov/AdvisoryCommittees/CommitteesMeetingMaterials/VeterinaryMedicineAdvisoryCommittee/ucm222635.htm

4. Crawford SS, Muir AM. Global introductions of salmon and trout in the genus *Oncorhynchus*: 1870–2007. Rev Fish Biol Fish. 2008;18(3):313–44.

5. Pister EP. Wilderness fish stocking: History and perspective. Ecosystems. 2001;4(4):279–86.

6. Knapp G, Roheim CA, Anderson JL. The great salmon run: Competition between wild and farmed salmon. Washington DC: TRAFFIC North America; 2007

7. Food and Agricultural Organization of the United Nations. FIGIS. FishStat (Database). 2015. Available from data.fao.org/ref/ babf3346-ff2d-4e6c-9a40-ef6a50fcd422.html

8. Bartley DM, Bondad-Reantaso MG, Subasinghe RP. A risk analysis framework for aquatic animal health management in marine stock enhancement programmes. Fish Res. 2006;80(1):28–36.

9. Mackie TJ, Arkwright JA, Pryce-Tannatt TE, Mottram JC, Johnston WDD, Menzies WJM, et al. Second interim report of the Furunculosis Committee. London: Ministry of Agriculture and Fisheries; 1933.

10. Smith IW, Department of agriculture and fisheries for Scotland. The occurrence and pathology of Dee disease. Freshwater Salmon Fish Res. 1964;34:1-12

11. Ordal EJ, Earp BJ. Cultivation and transmission of etiological agent of kidney disease in salmonid fishes. Proc Soc Exp Biol Med. 1956;92(1):85–8.

12. Earp BJ, Ellis CH, Ordal EJ. Kidney disease in young salmon. Olympia, State of Washington, Dept. of Fisheries; 1953.

13. Sanders JE, Fryer JL. *Renibacterium salmoninarum* gen. nov., sp. nov., the causative agent
of bacterial kidney disease in salmonid fishes. Int J Syst Evol Microbiol. 1980;30(2):496–502.

14. Savas HI, Altinok I, Cakmak E, Firidin S. Isolation of *Renibacterium salmoninarum* from cultured Black Sea salmon (*Salmo trutta labrax*): first report in Turkey. Bull Eur Assoc Fish Pathol. 2006;26(6):238-46.

15. Cheng LT, Lin WH, Wang PC, Tsai MA, Ho PY, Hsu JP, et al. Epidemiology and phylogenetic analysis of Taura syndrome virus in cultured Pacific white shrimp *Litopenaeus vannamei* B. in Taiwan. Dis Aquat Organ. 2011;97(1):17–23.

16. OIE. Quarterly aquatic animal disease report October - December 2007 (Asia and Pacific region). Paris; 2007 p. 39.

17. Toranzo AE, Magariños B, Romalde JL. A review of the main bacterial fish diseases in mariculture systems. Aquaculture. 2005;246(1–4):37–61.

18. OIE. World Animal Health in 2006. Paris; 2006.

19. Fryer JL, Lannan CN. The history and current status of *Renibacterium salmoninarum*, the causative agent of bacterial kidney disease in Pacific salmon. Fish Res. 1993;17(1–2):15–33.

20. Anderson C, Knowles G. Surveillance of New Zealand salmonids for *Renibacterium salmoninarum*. Surveillance. 1999;26(4):1.

21. Dale OB. Bakteriell nyresyke, infeksjon med *Renibacterium salmoninarum*. In: Poppe T, editor. Fiskehelse og fiskesykdommer. Oslo: Universitetsforlaget; 1999. p. 115–20.

22. Evenden AJ, Grayson TH, Gilpin ML, Munn CB. *Renibacterium salmoninarum* and bacterial kidney disease — the unfinished jigsaw. Annu Rev Fish Dis. 1993;3:87–104.

23. Evenden AJ. The use of gene cloning techniques in the study of the fish pathogen *Renibacterium salmoninarum* [PhD Dissertation]. University of Plymouth; 1993.

24. Starliper CE, Smith DR, Shatzer T. Virulence of *Renibacterium salmoninarum* to salmonids. J Aquat Anim Health. 1997;9(1):1–7.

25. Jansson E. Bacterial kidney disease in salmonid fish: development of methods to assess immune functions in salmonid fish during infection by *Renibacterium salmoninarum*. PhD Dissertation. Swedish University of Agricultural Sciences: Uppsala; 2002.

26. Dale OB, Gutenberger SK, Rohovec JS. Estimation of variation of virulence of *Renibacterium salmoninarum* by survival analysis of experimental infection of salmonid fish. J Fish Dis. 1997;20(3):177–83.

27. Sanders JE, Pilcher KS, Fryer JL. Relation of water temperature to bacterial kidney disease in coho salmon (*Oncorhynchus kisutch*), sockeye salmon (*O. nerka*), and steelhead trout (*Salmo gairdneri*). J Fish Res Board Can. 1978;35(1):8–11.

28. Fryer JL, Sanders JE. Bacterial kidney disease of salmonid fish. Annu Rev Microbiol. 1981;35(1):273–98.

29. Lall SP, Paterson WD, Hines JA, Adams NJ. Control of bacterial kidney disease in Atlantic salmon, *Salmo salar* L., by dietary modification. J Fish Dis. 1985;8(1):113–24.

30. Thorarinsson R, Landolt ML, Elliott DG, Pascho RJ, Hardy RW. Effect of dietary vitamin E and selenium on growth, survival and the prevalence of *Renibacterium salmoninarum* infection in chinook salmon (*Oncorhynchus tshawytscha*). Aquaculture. 1994;121(4):343–58.

31. Winter GW, Schreck CB, McIntyre JD. Resistance of different stocks and transferring genotypes of coho salmon, *Oncorhynchus kisutch*, and steelhead trout, *Salmo gairdneri*, to bacterial kidney disease and vibriosis. Fish Bull. 1980;77(4):795-802.

32. Withler RE, Evelyn TPT. Genetic variation in resistance to bacterial kidney disease within and between two strains of coho salmon from British Columbia. Trans Am Fish Soc. 1990;119(6):1003–9.

33. Bruno DW. The relationship between auto-agglutination, cell surface hydrophobicity and virulence of the fish pathogen *Renibacterium salmoninarum*. FEMS Microbiol Lett. 1988;51(2-3):135–9.

34. O'Farrell CL, Elliott DG, Landolt ML. Mortality and kidney histopathology of chinook salmon *Oncorhynchus tshawytscha* exposed to virulent and attenuated *Renibacterium salmoninarum* strains. Dis Aquat Organ. 2000;43(3):199–209.

35. Senson PR, Stevenson RMW. Production of the 57 kDa major surface antigen by a non-agglutinating strain of the fish pathogen *Renibacterium salmoninarum*. Dis Aquat Organ. 1999;38(1):23–31.

36. Rhodes LD, Coady AM, Deinhard RK. Identification of a third *msa* gene in *Renibacterium salmoninarum* and the associated virulence phenotype. Appl Environ Microbiol. 2004;70(11):6488–94.

37. Wiens GD, Pascho R, Winton JR. 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. Appl Environ Microbiol. 2002;68(8):3969–77.

38. Wood JW. Diseases of Pacific salmon: Their prevention and treatment. Olympia, Washington, USA: Hatchery division, Washington Department of Fisheries; 1974.

39. Iida T, Takahashi K, Wakabayashi H. Decrease in the bactericidal activity of normal serum during the spawning period of rainbow trout. Nippon Suisan Gakkaishi. 1989;55(3):463–5.

40. Evelyn TPT, Ketcheson JE, Prosperi-Porta L. 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. 1984;7(3):173–82.

41. Evelyn TPT, Prosperi-Porta L, Ketcheson JE. 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 Aquat Organ. 1986;1(3):197–202.

42. Evelyn TPT, Prosperi-Porta L, Ketcheson JE. Persistence of the kidney-disease bacterium, *Renibacterium salmoninarum*, in coho salmon, *Oncorhynchus kisutch* (Walbaum), eggs treated during and after water-hardening with povidone-iodine. J Fish Dis. 1986;9(5):461–4.

43. Elliott DG, Pascho RJ, Bullock GL. Developments in the control of bacterial kidney disease of salmonid fishes. Dis Aquat Organ. 1989;6(3):201–15.

44. Gutenberger SK, Duimstra JR, Rohovec JS, Fryer JL. Intracellular survival of *Renibacterium salmoninarum* in trout mononuclear phagocytes. Dis Aquat Organ. 1997;28(2):93–106.

45. Gutenberger SK. Phylogeny and intracellular survival of *Renibacterium salmoninarum* PhD Dissertation. Oregon State University: Corvallis, Oregon, USA; 1993.

46. Bullock GL, Herman RL. Bacterial kidney disease of salmonid fishes caused by *Renibacterium salmoninarum*. 1988 [cited 2015 Apr 23]; Available from: http://agris.fao.org/agris-search/search.do?recordID=US201300301795

47. Austin B, Rayment JN. Epizootiology of *Renibacterium salmoninarum*, the causal agent of bacterial kidney disease in salmonid fish. J Fish Dis. 1985;8(6):505–9.

48. Evelyn TPT. Bacterial kidney disease - BKD. In: Inglis V, Roberts RJ, Bromage NR, editors Bacterial diseases of fish. Oxford: Blackwell; 1993;p177–95.

49. Mitchum DL, Sherman LE. Transmission of bacterial kidney disease from wild to stocked hatchery trout. Can J Fish Aquat Sci. 1981;38(5):547–51.

50. Balfry SK, Albright LJ, Evelyn TPT. Horizontal transfer of *Renibacterium salmoninarum* among farmed salmonids via the fecal-oral route. Dis Aquat Organ. 1996;25(1-2):63-9.

51. Hoffmann R, Popp W, Graaff S van de. Atypical BKD [bacterial kidney disease] predominantly causing ocular and skin lesions. Bull Eur Assoc Fish Pathol. 1984;4(1):7–9.

52. Rucker RR, Earp BJ, Ordal EJ. Infectious diseases of Pacific salmon. Trans Am Fish Soc. 1954;83(1):297–312.

53. Chambers E, Gardiner R, Peeler EJ. An investigation into the prevalence of *Renibacterium salmoninarum* in farmed rainbow trout, *Oncorhynchus mykiss* (Walbaum), and wild fish populations in selected river catchments in England and Wales between 1998 and 2000. J Fish Dis. 2008;31(2):89–96.

54. Faisal M, Schulz C, Eissa A, Brenden T, Winters A, Whelan G, et al. Epidemiological investigation of *Renibacterium salmoninarum* in three *Oncorhynchus* spp. in Michigan from 2001 to 2010. Prev Vet Med. 2012;107(3–4):260–74.

55. Rhodes LD, Rice CA, Greene CM, Teel DJ, Nance SL, Moran P, et al. Nearshore ecosystem predictors of a bacterial infection in juvenile Chinook salmon. Mar Ecol Prog Ser. 2011;432:161–72.

56. Bronte CR, Ebener MP, Schreiner DR, DeVault DS, Petzold MM, Jensen DA, et al.

Fish community change in Lake Superior, 1970–2000. Can J Fish Aquat Sci. 2003;60(12):1552–74.

57. Sakai M, Ogasawara K, Atsuta S, Kobayashi M. Comparative sensitivity of carp, *Cyprinus carpio* L. and rainbow trout, *Salmo gairdneri* Richardson, to *Renibacterium salmoninarum*. J Fish Dis. 1989;12(4):367–72.

58. Traxler GS, Bell GR. Pathogens associated with impounded Pacific herring *Clupea harengus pallasi*, with emphasis on viral erythrocytic necrosis (VEN) and atypical *Aeromonas salmonicida*. Dis Aquat Organ. 1988;5(2):93–100.

59. Eissa AE, Elsayed EE, McDonald R, Faisal M. First record of *Renibacterium salmoninarum* in the sea lamprey *(Petromyzon marinus)*. J Wildl Dis. 2006;42(3):556–60.

60. Faisal M, Loch TP, Brenden TO, Eissa AE, Ebener MP, Wright GM, et al. Assessment of *Renibacterium salmoninarum* infections in four lake whitefish (*Coregonus clupeaformis*) stocks from northern Lakes Huron and Michigan. J Gt Lakes Res. 2010;36, Suppl 1:29–37.

61. Fenichel EP, Tsao JI, Jones MG. A model of *Renibacterium salmoninarum* dynamics in Great Lakes fish populations: Implications for bacterial kidney disease management and research in Lake Michigan. Unpublished completion report to Great Lakes Fisheries Trust / Michigan State University; 2007.

62. Sakai M, Kobayashi M. Detection of *Renibacterium salmoninarum*, the causative agent of bacterial kidney disease in salmonid fish, from pen-cultured coho salmon. Appl Environ Microbiol. 1992;58(3):1061–3.

63. Nylund A, Bjørknes B, Wallace C. *Lepeophtheirus salmonis* - A possible vector in the spread of diseases on salmonids. Bull Eur Assoc Fish Pathol. 1991;11(6):213–6.

64. Price CS. Factors affecting the saltwater-entry behavior and saltwater preference of Chinook salmon, *Oncorhynchus tshawytscha.* PhD Dissertation. Oregon State University: Corvallis, OR, USA; 2002.

65. Young CL, Chapman GB. Ultrastructural aspects of the causative agent and renal histopathology of bacterial kidney disease in brook trout (*Salvelinus fontinalis*). J Fish Res Board Can. 1978;35(9):1234–48.

66. Bruno DW. Histopathology of bacterial kidney disease in laboratory infected rainbow trout, *Salmo gairdneri* Richardson, and Atlantic salmon, *Salmo salar* L., with reference to

naturally infected fish. J Fish Dis. 1986;9(6):523–37.

67. Kaattari S, Turaga P, Wiens GD. Development of a vaccine for bacterial kidney disease in salmon - Final report to Bonneville Power Administration, Portland, OR, Contract 84-AI-16480, Project 84-46, 323 electronic pages (BPA Report DOE/BP-16480-5). 1989.

68. Grayson TH, Cooper LF, Wrathmell AB, Roper J, Evenden AJ, Gilpin ML. Host responses to *Renibacterium salmoninarum* and specific components of the pathogen reveal the mechanisms of immune suppression and activation. Immunology. 2002;106(2):273–83.

69. Balfry SK, Brown LL. Feasibility of selective breeding for resistance to bacterial kidney disease: Current state of knowledge. Vancouver Fish Health Management Committee: British Columbia Center for Aquatic Health Sciences; 2006.

70. Nilsen H, Jensen BB, Sunde EB, Rørvik S. The surveillance and control programme for bacterial kidney disease (BKD) in Norway 2011. Oslo, Norway: Norwegian Veterinary Institute; 2012.

71. Hjeltnes B. Fiskehelserapporten 2013. Oslo, Norway: Norwegian Veterinary Institute; 2013.

72. Murray AG, Munro LA, Wallace IS, Peeler EJ, Thrush MA. Bacterial kidney disease: Assessment of risk to Atlantic salmon farms from infection in trout farms and other sources. Scott Mar Freshw Sci: 2011;2(3).

73. Hall LM, Duguid S, Wallace IS, Murray AG. Estimating the prevalence of *Renibacterium salmoninarum*-infected salmonid production sites. J Fish Dis. 2014;38(2):231–5.

74. Murray AG, Hall M, Munro LA, Wallace IS. Modelling management strategies for a disease including undetected sub-clinical infection: Bacterial kidney disease in Scottish salmon and trout farms. Epidemics. 2011;3(3–4):171–82.

75. Bruno DW. Prevalence and diagnosis of bacterial kidney disease (BKD) in Scotland between 1990 and 2002. Dis Aquat Organ. 2004;59(2):125–30.

76. BC Centre for Aquatic Health Sciences. Evaluation of bacterial kidney disease (BKD) impacts on the Canadian salmon aquaculture industry. Final report to Fisheries and Oceans Canada; 2010.

77. Eissa AE, Elsayed EE, Faisal M. Prevalence and shedding of *Renibacterium salmoninarum* in brook trout (*Salvelinus fontinalis*) in Michigan. Nat Sci. 2007;5(1):8-17.

78. Rhodes LD, Durkin C, Nance SL, Rice CA. Prevalence and analysis of *Renibacterium salmoninarum* infection among juvenile chinook salmon *Oncorhynchus tshawytscha* in North Puget Sound. Dis Aquat Organ. 2006;71(3):179–90.

79. Densmore C. Bacterial kidney disease and its effect on the salmonid immune response. PhD dissertation. Virginia Polytechnic Institute and State University: Blacksburg, VA, USA; 1997.

80. OIE. Diagnostic manual for aquatic animal diseases. Paris, France; 2000.

81. European Commission. Scientific Committee on Animal Health and Animal Welfare. Bacterial kidney disease. Brussels; 1999.

82. Wood JW, Wallis, J. Kidney disease in adult chinook salmon and its transmission by feeding to young chinook salmon. Res Briefs. 1955;6(2):32–40.

83. Hill BJ. National legislation in Great Britain for the control of fish diseases. Rev Sci Tech Int Off Epizoot. 1996;15(2):633–45.

84. Marine Harvest. Salmon farming industry handbook. 2014. Available from http://www.marineharvest.com/globalassets/investors/handbook/handbook-2014.pdf

85. Colquhoun DJ. Fakta om: Bakteriell nyresjuke (BKD). Norwegian Veterinary Institute: Oslo, Norway; 2015. Available from: http://www.vetinst.no/Faktabank/Bakteriell-nyresjuke-BKD

86. Bruno DW, Munro ALS. Observations on *Renibacterium salmoninarum* and the salmonid egg. Dis Aquat Organ. 1986;1(2):83–7.

87. Bovo G, Hill B, Husby A, Håstein T, Michel C, Olesen NJ, et al. Work package 3 report: Pathogen survival outside the host, and susceptibility to disinfection. Veterinærmedisinsk Oppdragssenter AS: Oslo, Norway; 2005 p. 41.

88. Brown LL, Albright LJ, Evelyn TPT. Control of vertical transmission of *Renibacterium salmoninarum* by injection of antibiotics into the maturing female coho salmon *Oncorhynchus kisutch*. Dis Aquat Organ. 1990;9(2):127–31.

89. Austin B. Effectiveness of ozone for the disinfection of laboratory effluent. FEMS Microbiol Lett. 1983;19(2–3):211–4.

90. Skall HF, Olesen NJ. Treatment of wastewater from fish slaughterhouses. Evaluation and recommendation for hygienization methods. Danish Veterinary Institute; 2011.

91. Loncarevic S. Metoder godkjent for desinfeksjon av vann til/fra akvakulturrelatert virksomhet. Norwegian Veterinary Institute: Oslo, Norway; 2011.

92. Evelyn TPT. Prevention of vertical transmission of the bacterial kidney disease agent *Renibacterium salmoninarum* by broodstock injection with erythromycin. Dis Aquat Organ. 1994;18(1):1–4.

93. Rhodes LD, Nguyen OT, Deinhard RK, White TM, Harrell LW, Roberts MC. Characterization of *Renibacterium salmoninarum* with reduced susceptibility to macrolide antibiotics by a standardized antibiotic susceptibility test. Dis Aquat Organ. 2008;80(3):173–80.

94. Rhodes LD, Rathbone CK, Corbett SC, Harrell LW, Strom MS. Efficacy of cellular vaccines and genetic adjuvants against bacterial kidney disease in chinook salmon (*Oncorhynchus tshawytscha*). Fish Shellfish Immunol. 2004;16(4):461–74.

95. Alcorn S, Murray AL, Pascho RJ, Varney J. A cohabitation challenge to compare the efficacies of vaccines for bacterial kidney disease (BKD) in chinook salmon *Oncorhynchus tshawytscha*. Dis Aquat Organ. 2005;63(2-3):151–60.

96. Salonius K, Siderakis C, MacKinnon AM, Griffiths SG. Use of *Arthrobacter davidanieli* as a live vaccine against *Renibacterium salmoninarum* and *Piscirickettsia salmonis* in salmonids. Dev Biol. 2005;121:189–97.

97. Bravo S, Midtlyng PJ. The use of fish vaccines in the Chilean salmon industry 1999-2003. Aquaculture. 2007;270(1-4):36–42.

98. Beacham TD, Evelyn TPT. Genetic variation in disease resistance and growth of chinook, coho, and chum salmon with respect to vibriosis, furunculosis, and bacterial kidney disease. Trans Am Fish Soc. 1992;121(4):456–85.

99. Gjedrem T, Gjøen HM. Genetic variation in susceptibility of Atlantic salmon, *Salmo salar* L., to furunculosis, BKD and cold water vibriosis. Aquac Res. 1995;26(2):129–34.

100. Evelyn TPT. An improved growth medium for the kidney disease bacterium and some notes on using the medium. Bull Int Epizoot. 1977;87(5-6):511–3.

101. Evelyn TPT, Bell GR, Prosperi-Porta L, Ketcheson JE. A simple technique for accelerating the growth of the kidney disease bacterium *Renibacterium salmoninarum* on a commonly used culture medium (KDM2). Dis Aquat Organ. 1989;7:231–4.

102. Austin B, Embley T m., Goodfellow M. Selective isolation of *Renibacterium salmoninarum*. FEMS Microbiol Lett. 1983;17(1-3):111–4.

103. Campos-Pérez JJ, Ellis AE, Secombes CJ. Investigation of factors influencing the ability of *Renibacterium salmoninarum* to stimulate rainbow trout macrophage respiratory burst activity. Fish Shellfish Immunol. 1997;7(8):555–66.

104. Hardie LJ, Ellis AE, Secombes CJ. In vitro activation of rainbow trout macrophages stimulates inhibition of *Renibacterium salmoninarum* growth concomitant with augmented generation of respiratory burst products. Dis Aquat Organ. 1996;25(3):175–83.

105. Yousif AN, Albright, LJ, Evelyn TPT. In vitro evidence for the antibacterial role of lysozyme in salmonid eggs. Dis Aquat Organ. 1994;19:15–9.

106. Brown LL, Iwama GK, Evelyn TPT. The effect of early exposure of Coho salmon (*Oncorhynchus kisutch*) eggs to the p57 protein of *Renibacterium salmoninarum* on the development of immunity to the pathogen. Fish Shellfish Immunol. 1996 Apr;6(3):149–65.

107. Fredriksen Å, Endresen C, Wergeland HI. Immunosuppressive effect of a low molecular weight surface protein from *Renibacterium salmoninarum* on lymphocytes from Atlantic salmon (*Salmo salar* L.). Fish Shellfish Immunol. 1997;7(4):273–82.

108. Rockey DD, Turaga PS, Wiens GD, Cook BA, Kaattari SL. Serine proteinase of *Renibacterium salmoninarum* digests a major autologous extracellular and cell-surface protein. Can J Microbiol. 1991;37(10):758–63.

109. Turaga P, Wiens G, Kaattari S. Bacterial kidney disease: the potential role of soluble protein antigen(s). J Fish Biol. 1987;31(Suppl A):191–4.

110. Getchell RG, Rohovec JS, Fryer JL. Comparison of *Renibacterium salmoninarum* isolates by antigenic analysis. Fish Pathol. 1985;20(2/3):149–59.

111. Wiens GD, Kaattari SL. Monoclonal antibody characterization of a leukoagglutinin

produced by *Renibacterium salmoninarum*. Infect Immun. 1991;59(2):631–7.

112. Wood P, Kaattari S. Enhanced immunogenicity of *Renibacterium salmoninarum* in chinook salmon after removal of the bacterial cell surface-associated 57 kDa protein. Dis Aquat Organ. 1996;25(1-2):71–9.

113. Coady AM, Murray AL, Elliott DG, Rhodes LD. Both *msa* genes in *Renibacterium salmoninarum* are needed for full virulence in bacterial kidney disease. Appl Environ Microbiol. 2006;72(4):2672–8.

114. O'Farrell CL, Strom M. Differential expression of the virulence-associated protein p57 and characterization of its duplicated gene *msa* in virulent and attenuated strains of *Renibacterium salmoninarum*. Dis Aquat Organ. 1999;38(2):115–23.

115. Dubreuil JD, Jacques M, Graham L, Lallier R. Purification, and biochemical and structural characterization of a fimbrial haemagglutinin of *Renibacterium salmoninarum*. J Gen Microbiol. 1990;136(12):2443–8.

116. Fredriksen Å, Bakken V. Identification of *Renibacterium salmoninarum* surface proteins by radioiodination. FEMS Microbiol Lett. 1994;121(3):297–301.

117. McIntosh D, Flaño E, Grayson TH, Gilpin ML, Austin B, Villena AJ. Production of putative virulence factors by *Renibacterium salmoninarum* grown in cell culture. Microbiology. 1997;143(10):3349–56.

118. Ellis AE. Immunity to bacteria in fish. Fish Shellfish Immunol. 1999;9(4):291–308.

119. Wiens GD, Rockey DD, Wu Z, Chang J, Levy R, Crane S, et al. Genome sequence of the fish pathogen *Renibacterium salmoninarum* suggests reductive evolution away from an environmental *Arthrobacter* ancestor. J Bacteriol. 2008;190(21):6970–82.

120. Bruno DW, Munro ALS. Uniformity in the biochemical properties of *Renibacterium salmoninarum* isolates obtained from several sources. FEMS Microbiol Lett. 1986;33(2–3):247–50.

121. Fiedler F, Draxl R. Biochemical and immunochemical properties of the cell surface of *Renibacterium salmoninarum*. J Bacteriol. 1986;168(2):799–804.

122. Rhodes LD, Grayson TH, Alexander SM, Strom MS. Description and characterization of IS*994*, a putative IS *3* family insertion sequence from the salmon pathogen, *Renibacterium salmoninarum*. Gene. 2000;244(1):97–107.

123. Grayson TH, Alexander SM, Cooper LF, Gilpin ML. *Renibacterium salmoninarum* isolates from different sources possess two highly conserved copies of the rRNA operon. Antonie Van Leeuwenhoek. 2000;78(1):51–61.

124. Grayson TH, Atienzar FA, Alexander SM, Cooper LF, Gilpin ML. Molecular diversity of *Renibacterium salmoninarum* isolates determined by randomly amplified polymorphic DNA analysis. Appl Environ Microbiol. 2000;66(1):435–8.

125. Grayson TH, Cooper LF, Atienzar FA, Knowles MR, Gilpin ML. Molecular differentiation of *Renibacterium salmoninarum* isolates from worldwide locations. Appl Environ Microbiol. 1999;65(3):961–8.

126. Alexander SM, Grayson TH, Chambers EM, Cooper LF, Barker GA, Gilpin ML. Variation in the spacer regions separating tRNA genes in *Renibacterium salmoninarum* distinguishes recent clinical isolates from the same location. J Clin Microbiol. 2001;39(1):119–28.

127. Alexander SM. The molecular differentiation of *Renibacterium salmoninarum* isolates PhD Dissertation. University of Plymouth: Plymouth, UK; 2002.

128. Wiens GD, Dale OB. *Renibacterium salmoninarum* p57 antigenic variation is restricted in geographic distribution and correlated with genomic markers. Dis Aquat Organ. 2008;83(2):123-31.

129. Stinear TP, Seemann T, Pidot S, Frigui W, Reysset G, Garnier T, et al. Reductive evolution and niche adaptation inferred from the genome of *Mycobacterium ulcerans*, the causative agent of Buruli ulcer. Genome Res. 2007;17(2):192–200.

130. Burnett JR. Genetic variation of *Renibacterium salmoninarum* genes in infected salmonids Baccalaureate of Science in Bioresource Research, Biotechnology. Oregon State University: Corvallis, OR, USA; 2008.

131. Davies RL. Clonal analysis of *Yersinia ruckeri* based on biotypes, serotypes and outer membrane protein-types. J Fish Dis. 1991;14(2):221–8.

132. Garcia JA, Larsen JL, Dalsgaard I, Pedersen K. Pulsed-field gel electrophoresis analysis of *Aeromonas salmonicida* ssp. *salmonicida*. FEMS Microbiol Lett. 2000;190(1):163–6.

133. Wheeler RW, Davies RL, Dalsgaard I, Garcia J, Welch TJ, Wagley S, et al. *Yersinia*

ruckeri biotype 2 isolates from mainland Europe and the UK likely represent different clonal groups. Dis Aquat Organ. 2009;84(1):25–33.

134. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science. 1995;269(5223):496–512.

135. Roach JC, Boysen C, Wang K, Hood L. Pairwise end sequencing: a unified approach to genomic mapping and sequencing. Genomics. 1995;26(2):345–53.

136. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci. 1977;74(12):5463–7.

137. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975;94(3):441–8.

138. Ahmadian A, Gharizadeh B, Gustafsson AC, Sterky F, Nyrén P, Uhlén M, et al. Single-nucleotide polymorphism analysis by pyrosequencing. Anal Biochem. 2000;280(1):103–10.

139. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, et al. A large genome center's improvements to the Illumina sequencing system. Nat Methods. 2008;5(12):1005–10.

140. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323(5910):133–8.

141. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature. 2011;475(7356):348–52.

142. Pandey V, Nutter RC, Prediger E. Applied Biosystems SOLiD™ system: Ligation-based sequencing. In: Janitz M, editor. Next Generation Genome Sequencing. Weinheim Wiley-VCH Verlag; 2008. p. 29–42.

143. Ferragina P, Manzini G. Opportunistic data structures with applications. In: Proceedings of the 41st Symposium on Foundations of Computer Science (FOCS 2000). Los Alamitos, CA, USA: IEEE Computer Society; 2000. p. 390–8.

144. Burrows M, Wheeler DJ. A block sorting lossless data compression algorithm. Digital Equipment Corporation: Palo Alto, CA, USA; 1994. (SRC Research Report No. 124.)

145. Nagarajan N, Pop M. Sequence assembly demystified. Nat Rev Genet. 2013;14(3):157–67.

146. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970;48(3):443–53.

147. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147(1):195–7.

148. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.

149. Kingman JFC. The coalescent. Stoch Process Appl. 1982;13(3):235–48.

150. Tajima F. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics. 1993;135(2):599–607.

151. Bromham L, Penny D. The modern molecular clock. Nat Rev Genet. 2003;4(3):216–24.

152. Zuckerkandl E. On the molecular evolutionary clock. J Mol Evol. 1987;26(1-2):34–46.

153. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4(4):406–25.

154. Sokal RR, Michener C. A statistical method for evaluating systematic relationships. Univ Kansas Sci Bull. 1958;38:1409–38.

155. Farris JS. Methods for computing Wagner trees. Syst Zool. 1970;19(1):83–92.

156. Fitch WM. Toward Defining the Course of Evolution: Minimum change for a specific tree topology. Syst Biol. 1971;20(4):406–16.

157. Felsenstein J. Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol. 1981;17(6):368–76.

158. Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. Bayesian inference of phylogeny and its impact on evolutionary biology. Science. 2001;294(5550):2310–4.

159. Hastings WK. Monte Carlo sampling methods using Markov chains and their

applications. Biometrika. 1970;57(1):97–109.

160. Matejusova I, Bain N, Colquhoun DJ, Feil EJ, McCarthy U, McLennan D, et al. Multilocus variable-number tandem-repeat genotyping of *Renibacterium salmoninarum*, a bacterium causing bacterial kidney disease in salmonid fish. BMC Microbiol. 2013;13(1):285.

161. Harris SR, Feil EJ, Holden MTG, Quail MA, Nickerson EK, Chantratita N, et al. Evolution of MRSA during hospital transmission and intercontinental spread. Science. 2010;327(5964):469–74.

162. Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, et al. Evidence for several waves of global transmission in the seventh cholera pandemic. Nature. 2011;477(7365):462–5.

163. Holt KE, Baker S, Weill F-X, Holmes EC, Kitchen A, Yu J, et al. *Shigella sonnei* genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. Nat Genet. 2012;44(9):1056–9.

164. Plarre H, Nylund A, Karlsen M, Brevik Ø, Sæther PA, Vike S. Evolution of infectious salmon anaemia virus (ISA virus). Arch Virol. 2012;157(12):2309–26.

165. Horne MT, Barnes AC. Enteric redmouth disease (*Yersinia ruckeri*). In: Woo PTK, Bruno DW, editors. Fish Diseases and Disorders. Vol. 3. Viral Bacterial and Fungal Infections. Wallington: CABI Publishing; 1999. p. 455–77.

166. Adams D. Dirk Gently's Holistic Detective Agency. London, UK: William Heinemann; 1987.

167. Goodfellow M, Embley TM, Austin B. Numerical taxonomy and emended description of *Renibacterium salmoninarum*. J Gen Microbiol. 1985;131(10):2739–52.

168. Kondrashov FA. Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc R Soc B Biol Sci. 2012;279(1749):5048–57.

169. Riehle MM, Bennett AF, Long AD. Genetic architecture of thermal adaptation in *Escherichia coli*. Proc Natl Acad Sci. 2001;98(2):525–30.

170. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science. 2007;315(5813):848–53.

171. Conant GC, Wolfe KH. Turning a hobby into a job: How duplicated genes find new functions. Nat Rev Genet. 2008;9(12):938–50.

172. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. 2008;36(16):e105–e105.

173. Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 2011;12(2):R18.

174. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012;40(10):e72

175. Nijkamp JF, van den Broek MA, Geertman J-MA, Reinders MJT, Daran J-MG, de Ridder D. *De novo* detection of copy number variation by co-assembly. Bioinformatics. 2012;28(24):3195–202.

176. Juang B-H, Rabiner L. The segmental K-means algorithm for estimating parameters of hidden Markov models. IEEE Trans Acoust Speech Signal Process. 1990;38(9):1639–41.

177. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. Bioinformatics. 2011;28(4):593–4.

178. Klambauer G, Schwarzbauer K, Mayr A, Clevert D-A, Mitterecker A, Bodenhofer U, et al. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res. 2012;40(9):e69.

179. Xie C, Tammi MT. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinformatics. 2009;10(1):80.

180. Zwickl DJ, Hillis DM. Increased taxon sampling greatly reduces phylogenetic error. Syst Biol. 2002;51(4):588–98.

181. Heath TA, Hedtke SM, Hillis DM. Taxon sampling and the accuracy of phylogenetic analyses. J Syst Evol. 2008;46(3):239–57.

182. Rosenberg MS, Kumar S. Incomplete taxon sampling is not a problem for

phylogenetic inference. Proc Natl Acad Sci U S A. 2001;98(19):10751–6.

183. Posada D, Crandall KA. The effect of recombination on the accuracy of phylogeny estimation. J Mol Evol. 2002;54(3):396–402.

184. Bloomquist EW, Suchard MA. Unifying vertical and nonvertical evolution: A stochastic ARG-based framework. Syst Biol. 2010;59(1):27–41.

185. Kasianowicz JJ, Brandin E, Branton D, Deamer DW. Characterization of individual polynucleotide molecules using a membrane channel. Proc Natl Acad Sci U S A. 1996;93(24):13770–3.

186. Hempel S. John Snow. Lancet. 2013;381(9874):1269–70.

187. Sepúlveda N, Campino SG, Assefa SA, Sutherland CJ, Pain5 A, Clark TG. A Poisson hierarchical modelling approach to detecting copy number variation in sequence coverage data. BMC Genomics. 2013;14(1):128.

188. Hulbig VA. Developing a model for bacterial kidney disease in the zebrafish, *Danio rerio*. Master of Science. University of Maine: Orono, ME, USA; 2007.

189. Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Henderson DK, Palmore TN, et al. Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. Sci Transl Med. 2012;4(148):148ra116.

190. Morelli G, Song Y, Mazzoni CJ, Eppinger M, Roumagnac P, Wagner DM, et al. Phylogenetic diversity and historical patterns of pandemic spread of *Yersinia pestis*. Nat Genet. 2010;42(12):1140–3.

10. Scientific papers I - III

I

OPEN

The ISME Journal (2013), 1–11
© 2013 International Society for Microbial Ecology All rights reserved 1751-7362/13
www.nature.com/ismej

ORIGINAL ARTICLE

Microevolution of *Renibacterium salmoninarum*: evidence for intercontinental dissemination associated with fish movements

Ola Brynildsrud[1], Edward J Feil[2], Jon Bohlin[1], Santiago Castillo-Ramirez[2], Duncan Colquhoun[3], Una McCarthy[4], Iveta M Matejusova[4], Linda D Rhodes[5], Gregory D Wiens[6] and David W Verner-Jeffreys[7]

[1]*EpiCentre, Department of Food Safety and Infection Biology, Norwegian School of Veterinary Science, Oslo, Norway;* [2]*Department of Biology and Biochemistry, University of Bath, Bath, UK;* [3]*Section for Bacteriology, Department of Laboratory Services, Norwegian Veterinary Institute, Oslo, Norway;* [4]*Marine Scotland Science, Aberdeen, Scotland, UK;* [5]*NOAA, Northwest Fisheries Science Center, Seattle, WA, USA;* [6]*USDA, National Centre for Cool and Coldwater Aquaculture, Leetown, WV, USA and* [7]*Cefas Weymouth Laboratory, The Nothe, Weymouth, UK*

***Renibacterium salmoninarum* is the causative agent of bacterial kidney disease, a major pathogen of salmonid fish species worldwide. Very low levels of intra-species genetic diversity have hampered efforts to understand the transmission dynamics and recent evolutionary history of this Gram-positive bacterium. We exploited recent advances in the next-generation sequencing technology to generate genome-wide single-nucleotide polymorphism (SNP) data from 68 diverse *R. salmoninarum* isolates representing broad geographical and temporal ranges and different host species. Phylogenetic analysis robustly delineated two lineages (lineage 1 and lineage 2); futhermore, dating analysis estimated that the time to the most recent ancestor of all the isolates is 1239 years ago (95% credible interval (CI) 444–2720 years ago). Our data reveal the intercontinental spread of lineage 1 over the last century, concurrent with anthropogenic movement of live fish, feed and ova for aquaculture purposes and stocking of recreational fisheries, whilst lineage 2 appears to have been endemic in wild Eastern Atlantic salmonid stocks before commercial activity. The high resolution of the SNP-based analyses allowed us to separate closely related isolates linked to neighboring fish farms, indicating that they formed part of single outbreaks. We were able to demonstrate that the main lineage 1 subgroup of *R. salmoninarum* isolated from Norway and the UK likely represent an introduction to these areas ~40 years ago. This study demonstrates the promise of this technology for analysis of micro and medium scale evolutionary relationships in veterinary and environmental microorganisms, as well as human pathogens.**

The ISME Journal advance online publication, 31 October 2013; doi:10.1038/ismej.2013.186

Subject Category: Microbial population and community ecology

Keywords: fish pathogens; microbial evolution; next-generation sequencing; *Renibacterium salmoninarum*; transmission history

Introduction

The causative agent of bacterial kidney disease (BKD) in salmonids, *Renibacterium salmoninarum,* is a Gram-positive slow-growing facultative intracellular pathogen. BKD, a chronic, progressive granulomatous infection, is a major threat to both farmed and wild salmonid fish species worldwide (Fryer and Sanders, 1981; Evelyn, 1993; Evenden *et al.*, 1993; Fryer and Lannan, 1993; Wiens, 2011). It was first reported in the wild Atlantic salmon (*Salmo salar* L.) from rivers in Scotland and in brook and brown trout from the East coast of US in the 1930s (Earp *et al.*, 1953; Smith, 1964).

The genome of *R. salmoninarum* consists of a single circular 3.15-Mbp chromosome with no known plasmids or phage elements (Wiens *et al.*, 2008). As with other specialized intracellular pathogens, there is evidence of genome reduction (Wiens *et al.*, 2008), and it has evolved mechanisms to evade detection by the host immune system (Grayson *et al.*, 2002). *R. salmoninarum* survives, and possibly also replicates, within the macrophages of the kidney (Young and Chapman, 1978; Gutenberger *et al.*, 1997). The bacterium is able to spread horizontally between fish hosts as well as vertically

Correspondence: DW Verner-Jeffreys, Cefas Weymouth Laboratory, CEFAS, The Nothe, Barrack Rd, Weymouth DT4 8UB, UK.
E-mail: david.verner-jeffreys@cefas.co.uk

Received 5 June 2013; revised 4 September 2013; accepted 9 September 2013

via the ova (Evelyn *et al.*, 1986). Overt symptoms may not be seen until several months post infection, thus providing ample opportunity for horizontal transmission through stocks. Transovarial transmission additionally provides a mechanism for global dissemination of *R. salmoninarum* via commercial activity. These factors, coupled with a lack of efficient therapy and vaccine regimens for this disease (Elliott *et al.*, 1989), have, in some countries, resulted in the imposition of movement controls on premises confirmed as positive for BKD, or the destruction of infected animals, disinfection and fallowing of premises (Richards, 2011).

There is continued speculation relating to the likely origins and spread of BKD. It is frequently detected in wild and hatchery-bred Pacific salmon (*Oncorhynchus* spp.) in both freshwater and oceanic phases (Banner *et al.*, 1986; Kent *et al.*, 1998; Meyers *et al.*, 2003; Arkoosh *et al.*, 2004). Paterson *et al.* (1979) also detected *R. salmoninarum* in the kidneys of Atlantic salmon returning to rivers in Eastern Canada using an indirect fluorescent-antibody technique. However, the significance of reservoirs of infection in wild and feral salmonid populations in Western Europe is unclear. A PCR-based survey of wild fish from six rivers in England and Wales reported an infection prevalence of 8% in grayling (*Thymallus thymallus*) and 4.8% in Atlantic salmon (Chambers *et al.*, 2008). It has also not been determined whether the original outbreaks of the 'Dee Disease' in Scottish rivers (Smith, 1964) were caused by introduction of the pathogen from elsewhere (for example, via ova imported from North America), or the represented reoccurrence of a disease that had long been endemic in European populations of Atlantic salmon. In the UK there has also been debate as to the extent to which farmed rainbow trout infected with *R. salmoninarum* pose a risk to farmed (and wild) Atlantic salmon and the best ways to control these potential risks (Murray *et al.*, 2011; Richards, 2011; Murray *et al.*, 2012).

Highly discriminatory *R. salmoninarum* genotyping tools are required to address these questions. Several different molecular typing methods have been developed for *R. salmoninarum* (Grayson *et al.*, 1999, 2000; Rhodes *et al.*, 2000; Alexander *et al.*, 2001). The data generated by these studies point to *R. salmoninarum* being highly clonal with very limited phenotypic and genetic variation. There is, therefore, a need for more powerful methods to resolve sub-lineages within the population and, in doing so, to reconstruct recent evolutionary history and patterns of transmission. To this end, we generated genome-wide single-nucleotide polymorphism (SNP) data using a next-generation sequence platform for 68 strains of *R. salmoninarum* isolated from different host species, and wide temporal and geographical ranges. The data were analyzed using methods pioneered for use with important human pathogens (for example, Harris *et al.*, 2010; Mutreja *et al.*, 2011; Holt *et al.*, 2012). To our knowledge, this study represents the first application of this technology to a strictly animal pathogen.

Materials and methods

Bacterial strains and growth conditions

The sample was composed as follows: two of the original isolations from wild Atlantic salmon from the River Dee in 1960, 12 isolates from Norway, 7 isolates from New Brunswick on the East Coast of Canada, 11 isolates from the West Coast of the USA and Canada (including Washington, Oregon and British Columbia), 22 isolates from Scotland, 12 isolates from England and Wales, one single isolate from an eastern brook trout from Alberta, Canada and one isolate from a grayling from Montana, USA. Full details of these isolates are given in Table 1.

Cultures were grown on KDM solid media (Evelyn, 1977) at 15 °C, with colonies isolated and grown for DNA extraction.

DNA extraction and sequencing

Each participating laboratory separately prepared DNA from the isolates they supplied for the study. In brief, DNA was extracted from freshly grown bacteria harvested directly from solid media, resuspended in 500 µl sterile deionized water, and then centrifuged at 14 000 *g* for 15 min. Or, in other cases, cultures of *R salmoninarum* were grown in KDM broth to an OD_{525} and centrifuged at 10 000 *g* for 20 min. In all cases, the resultant bacterial pellets were then resuspended in 1 ml 10 mM Tris and the DNA extracted as described by Wiens *et al.,* 2002. There were no noticeable differences in the DNA quality in preparations from either solid or broth media (data not shown).

All the isolates (Table 1) were submitted for whole-genome, paired-end sequencing to The Genome Analysis Centre, Norwich, UK. DNA quality and yield was first determined by fluorometry using a Qubit fluorometer (Invitrogen) with QUANT-iT dsDNA assay (Broad Range). The samples were also assessed for RNA contamination using a Qubit fluorometer with QUANT-iT RNA assay (Invitrogen). DNA TruSeq libraries were constructed for each isolate and were run on the Illumina (San Diego, CA, USA) HiSeq 2000 platform in pools of up to 12 libraries per lane.

DNA TruSeq library construction

The Illumina TruSeq DNA Sample Preparation was used to prepare pooled-indexed paired-end libraries of genomic DNA for subsequent cluster generation (Illumina cBot) and DNA sequencing using the reagents provided in the Illumina TruSeq DNA Sample Preparation v2 Kit. The samples (starting material of ~1 µg of genomic DNA) were sheared using a sonicator (Covaris (Woburn, MA, USA), S2/LE220) to fragment sizes in the range of 200–700 bp.

Table 1 *R. salmoninarum* isolates used in the study

Isolate	*Geographic origin*	*Year*	*Source*[a]	*Alternative ID*	*GenBank/ EBI accession number*
Rs 10	New Brunswick, Canada	2009	*Salmo salar* (sw)		ERR327945
Rs 2	New Brunswick, Canada	2005	*S. salar* (sw)		ERR327951
Rs 3	New Brunswick, Canada	2005	*S. salar* (fw)		ERR327947
Rs 4	New Brunswick, Canada	2006	*S. salar* (sw)		ERR327946
Rs 5	New Brunswick, Canada	2007	*S. salar* (sw)		ERR327950
Rs 6	New Brunswick, Canada	2007	*S. salar* (sw)		ERR327953
Rs 8	New Brunswick, Canada	2008	*S. salar* (sw)		ERR327944
BPS 91	Nanaimo, BC, Canada	1991	*Oncorhynchus gorbuscha*		ERR327952
BQ96 91-1	Nanaimo, BC, Canada	1996	*Oncorhynchus kisutch*		ERR327963
DR143	Alberta, Canada	1972	*Salvelinus fontinalis (*fw)[b]		ERR327954
5006	Bella Bella, BC, Canada	1996	*O. kisutch* (sw)	960046	ERR327942
5223	Kvinnherad, Hordaland, Norway	2005	*S. salar* (sw)	2005-50-579	ERR327964
6553	Hemne, Sør-Trøndelag, Norway	2008	*S. salar* (sw)	2008-09-495	ERR327955
6642	Hemne, Sør-Trøndelag, Norway	2008	*S. salar*	2008-06-633	ERR327956
6694	Hemne, Sør-Trøndelag, Norway	2008	*Oncorhynchus mykiss* (sw)		ERR327962
6695	Hemne, Sør-Trøndelag, Norway	2008	*O. mykiss* (sw)	2008-06-631	ERR327968
6863	Osterøy, Hordaland, Norway	2009	*O. mykiss* (sw)		ERR327965
7439	Sognefjorden, Sogn og Fjordane, Norway	1984	*S. salar*	1984-40.992	ERR327971
7441	Storfjord, Møre og Romsdal, Norway	1985	*S. salar*	1985-09-667	ERR327966
7448	Stranda, Møre og Romsdal, Norway	1986	*S. salar*	1986-09-4366	ERR327970
7449	Skjervøy, Troms, Norway	1987	*S. salar*	1987-09-932	ERR327969
7450	Askøy, Hordaland, Norway	1987	*S. salar*	1987-09-1185	ERR327967
684	Aurland, Sognefjorden, Norway	1987	*S. trutta* (fw)		ERR327958
1205	UK	2001	*O. mykiss*	3104-67	ERR327930
5007	Scotland	2005	*O. mykiss*	0180-18	ERR327923
7105	UK	2007	*O. mykiss* (fw)	P0416 T83 10-3 2	ERR327932
9025	Yorkshire, England, UK	2009	*O. mykiss* (fw)	16251-1	ERR327912
96071	England, Hampshire, site Z, UK	1996	*O. mykiss* (fw)		ERR327927
99326	Wales, site Y, UK	1999	*O. mykiss* (fw)	2119-8	ERR327938
99327	UK	1997	*O. mykiss* (fw)	970313-2	ERR327931
99329	Wales, site X, UK	1998	*O. mykiss* (fw)	980036-125	ERR327937
99332	Wales, site Y, UK	1999	*O. mykiss* (fw)	2119-3	ERR327943
99333	Wales, site X, UK	1998	*O. mykiss* (fw)	980036-102	ERR327921
99341	Hampshire, site Z, England, UK	1998	*O. mykiss* (fw)	980109-20	ERR327949
99344	Hampshire, England, UK	1998	*O. mykiss* (fw)	980106-1.1.5	ERR327940
99345	Wales, site X	1998	*O. mykiss* (fw)	980070-18	ERR327948
NCIMB 1114	River Dee, Scotland, UK	1962	*S. salar* (fw)[b]	5005	ERR327908
NCIMB 1116	River Dee, Scotland, UK	1962	*S. salar* (fw)[b]	96056	ERR327907
MT239	Scotland, UK	1988	*S. salar*		ERR327913
MT1363	Strathclyde, Scotland, UK	1993	*O. mykiss* (sw)		ERR327920
MT3277	Dumfries and Galloway Site A, Scotland,UK	2008	*O. mykiss* (fw)		ERR327926
MT3313	Central, Scotland, UK	2008	*O. mykiss* (fw)		ERR327925
MT3315	Strathclyde Site B, Scotland, UK	2008	*O. mykiss* (fw)		ERR327928
MT1262	Highlands, Scotland, UK	1992	*S. salar* (fw)		ERR327922
MT1351	Highlands, Scotland, UK	1993	*S. salar* (sw)		ERR327904
MT1470	Tayside, Scotland, UK	1994	*O. mykiss* (fw)		ERR327910
MT1511	Strathclyde Site B, Scotland, UK	1994	*O. mykiss* (fw)		ERR327914
MT1880	Strathclyde, Scotland, UK	1996	*S. salar* (sw)		ERR327909
MT2622	Strathclyde, Scotland, UK	2002	*O. mykiss* (sw)		ERR327929
MT2943	Highlands, Scotland, UK	2005	*S. salar* (sw)		ERR327936
MT2979	Highlands, Scotland, UK	2005	*O. mykiss* (fw)		ERR327935
MT3106	Strathclyde, Scotland, UK	2006	*O. mykiss* (fw)		ERR327939
MT3479	Orkney, Scotland, UK	2008	*S. salar* (sw)		ERR327933
MT3482	Strathclyde, Scotland, UK	2009	*S. salar* (sw)		ERR327934
MT3483	Strathclyde, Scotland, UK	2009	*S. salar* (sw)		ERR327941
MT444	Western Isles, Scotland, UK	1988	*S. salar* (sw)		ERR327916
MT452	Dumfries and Galloway Site A, Scotland,UK	1988	*O. mykiss* (fw)		ERR327918
MT839	Highlands, Scotland, UK	1990	*S. salar* (sw)		ERR327917
MT861	Scotland	1990	*S. salar* (sw)		ERR327919
Car 96	Washington State, USA	1996	*O. tshawytscha*		ERR327957
D6	Oregon, USA	1982	*O. tshawytscha*		ERR327961
GR5	Montana, USA	1997	*T. thymallus* (fw)[b]	980036-87	ERR327959
WR99 c2	Washington State, USA	1999	*O. kisutch*		ERR327960
NCIMB 2235	Oregon, USA	1974	*O. tshawytscha* (sw)	ATCC33209	ERR327911
05372K	Grande Ronde Basin, Oregon, USA	2005	*O. tshawytscha* (sw)		ERR327906
Carson 5b	Confluence Tyee Creek & Wind River, WA, USA	1994	*O. tshawytscha* (fw)		ERR327905
Cow-chs-94	Cowlitz River, Washington	1994	*O. tshawytscha* (fw)		ERR327915
ATCC 33209[c]	Oregon, USA	1974	*O. tshawytscha* (sw)		NC_010168.1
NCIMB 1111[d]	—	Not known	Not known	5004	ERR327924

Abbreviations: fw, fresh water; sw, sea water.

The complete history for some of the isolates is not known.

[a] Isolates recovered from fw or sw, where known.

[b] Isolate recovered from a wild fish species, all other isolates were recovered from farmed fish or not known.

[c] Used as a reference in this study. Sequence data downloaded from Genbank.

[d] NCIMB 1111 was deposited in the NCIMB culture collection in Aberdeen after 1960 by I Smith, who also deposited isolates NCIMB 1114 and NCIMB 1116 recorded as being isolated by her from wild *S. salar* from the River Dee in the early 1960s. NCIMB 1111 was reportedly isolated by Ken Wolf. Ken Wolf was a highly active US fish disease researcher who worked on US and Canadian strains of the pathogen in the 1950s and 1960s. It thus appears most likely that this isolate was a North American strain provided by Ken Wolf to I Smith to assist her with her studies on Dee disease (although this cannot be proven).

The ends of the DNA were then repaired and a single A base added to each 3′ end of the DNA fragment to which an indexed adapter binds. A gel size selection method was used (Invitrogen E-Gel, 2% agarose) to select the appropriate sized library that was then enriched by PCR, quantified using an Agilent (Santa Clara, CA, USA) DNA 1000 chip on the Agilent Bioanalyzer 2100, pooled with up to 12 other libraries and sequenced. Library preparation was automated using a Perkin Elmer (Waltham, MA, USA) (formally Caliper) Sciclone NGS Workstation (sonication, size selection, enrichment and pooling are not performed on the Sciclone).

Data processing, genome alignment and assembly

Close to 1 000 000 000 pairs of reads, each of length 100 bp, were created for the project in total. The raw data have been deposited in the database of the European Bioinformatics Institute, and is available at http://www.ebi.ac.uk. Accession numbers are listed in Table 1. Reads were pre-filtered through the *eliminate_singletons.py*, *eliminate_n_paired.py* and *filter_reads.py* scripts from the BioPython (version 1.60) package (Cock *et al.*, 2009) before assembly, excluding reads that would not properly pair, reads with ambiguous (‘N’) base calls and reads with an average PHRED quality score below 20.

We used MAQ v.0.7.1 (Li *et al.*, 2008) to align raw reads to ATCC33209, the reference genome published in GenBank (http://www.ncbi.nlm.nih.gov/genbank. NCBI accession number: NC_010168.1). MAQs *sol2sanger* script was used to transform PHRED scores to the PHRED + 33 style. The vast majority of the reads mapped onto the reference genome, providing an average read depth across non-gap regions of 862.0 (Inter-isolates range: 57.2–2012.5). Statistics related to read output, read depth and read mapping is presented in Supplementary File SF3. Reads that mapped equally well to multiple areas of the reference genome, such as reads representing IS994 and ISRs2 sequences, were randomly assigned to one of the possible mapping locations. Variant calls from these areas were considered unreliable and were excluded from further analyses. Non-mapping reads were sorted into a separate file for separate *de novo* assembly in Velvet v.1.2.03 (Zerbino and Birney, 2008), using a k-mer length of 31. However, unmapped reads were all from Bacteriophage phi X 174, which was used as a positive control in the sequencing stage.

To validate these results, we also did *de novo* assembly using a combination of outputs from the DBG assemblers Velvet and ABySS (Simpson *et al.*, 2009), as well as comparative assembly using the AMOScmp-shortreads tool from the AMOS package (Treangen *et al.*, 2011). These assemblies were then merged using the minimus2 tool. This is an abridged version of the pipeline conceived and described by Ji *et al.*, 2011. Comparative evaluation of the output from these two different pipelines was done in Hawkeye (Schatz *et al.*, 2007). MAUVE (Darling *et al.*, 2004) was used to screen for evidence of genomic rearrangements.

Variant calling and phylogeny reconstruction

SNP calling was done with the default settings in MAQs cns2snp and maq.pl SNPfilter scripts. Furthermore, SNPs with PHRED quality scores of less than 255 were removed from analysis, as did SNPs with a high strand-bias. For a limited number of isolates, MAQ would occasionally produce ambiguous character SNP calls, even though PHRED qualities were consistently high. Closer inspection of the alignment revealed that the issue was caused by a low frequency (<2%) of reads representing an alternative genotype. Although other plausible scenarios could explain this, slight DNA contamination is the most likely. We used the variants represented in the overwhelming majority of reads in these cases. Using SplitsTree4 (Graham *et al.*, 2005), we could find no trace of recombination in our isolates. SNP sequences were then loaded into R (www.r-project.org) and processed using the *ape* (Paradis *et al.*, 2004) and *phangorn* (Schliep, 2011) packages, in conjunction with the maximum-likelihood estimator PhyML (Guindon *et al.*, 2010), to infer the optimal phylogenetic substitution model for the data. Several models were suggested as an outcome of this analysis, but the generalized time-reversible model (Tavaré, 1986) without rate variation among sites obtained the lowest AIC score (Akaike, 1974), and was therefore chosen. Trees for both the full SNP alignment and pseudogene SNPs only were created with the MrBayes (Ronquist and Huelsenbeck, 2003) plugin in Geneious (version 6.0.3), using a generalized time-reversible-substitution model with a uniform site distribution. The Markov Chain Monte Carlo settings were set to include 10 000 000 generations with subsampling done every 2000th step. Burn-in was set to 1 000 000 generations, and the unconstrained branch length option was used. The tree was subsequently annotated in FigTree (http://www.tree.bio.ed.ac.uk/software/figtree/).

Bayesian Markov Chain Monte Carlo analyses

A dated phylogeny was constructed by means of BEAST (Drummond and Rambaut, 2007; Drummond *et al.*, 2012), based on the multiple alignment of all the SNPs that were not located in paralogous genes, using an uncorrelated lognormal relaxed clock with the generalized time-reversible model. Date of collection was used to estimate the divergence times of isolates. Two analyses were run for 600 000 000 generations, sampling every 30 000 generations. We combined the results from the two independent runs through LogCombiner (excluding the first 10% generations from each analysis). TRACER was used to evaluate the convergence of the combined analysis, the first 10% generations were discarded as burn-in; we corroborated that the effective sample size of all the parameters were greater than 200, and ensured that the trace plots of the likelihood scores randomly oscillated within a stable range. The mean rate of the molecular clock for the whole genome was calculated by multiplying the mean rate of the uncorrelated lognormal clock for the SNP collection by the SNP density in the genome.

Ancestral character state reconstruction

Mesquite (Maddison and Maddison, 2011) was used to reconstruct nodal character states based on terminal taxa values. Phylogenetic trees were constructed from pseudogene SNPs only. Both a parsimonious and a maximum-likelihood approach were used. For both approaches, geographical data were simplified to represent the country of origin in an unordered, character matrix. In the likelihood estimate, the one-parameter Markov k-state probability model (Lewis, 2001) was used.

Accession numbers

The sequence data for all the isolates were deposited in the European Bioinformatics Institute Short Read Archive under the accession numbers ERR327904 to ERR327971inclusive.

Results

Phylogenetic analysis of R. salmoninarum

A total of 3600 high-quality core-genome SNPs were identified (see Supplementary File 1), corresponding to one SNP every 876 bases. These were evenly distributed across the genome. There was no evidence of genomic rearrangement. Phylogenetic analysis resolved two major sub-lineages; lineage 1 and lineage 2 (Figure 1). Lineage 1 encompassed 90% of the studied isolates (61/68). These isolates were recovered from seven of the eight different host species, from the full range of geographical locations, and over a 50-year period (1960–2009). Despite this geographical, temporal and ecological diversity, lineage 1 isolates exhibit very low levels of

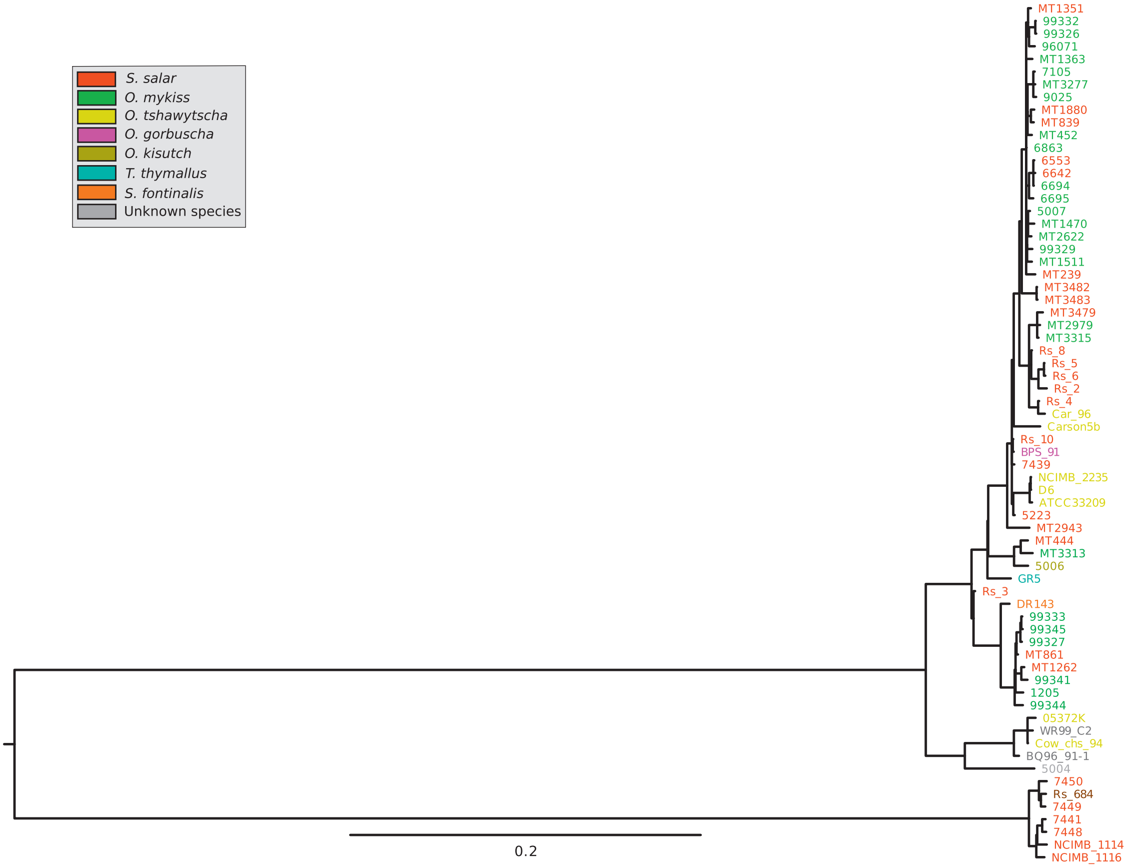


Figure 1 Phylogenetic tree of the 68 isolates of *R. salmoninarum* included in this study, showing all lineages. The evolutionary history was inferred using a Bayesian Markov Chain Monte Carlo approach, with a generalized time-reversible model (Tavaré, 1986), through the MrBayes (Ronquist and Huelsenbeck, 2003) plugin in Geneious. The consensus tree is taken to represent the evolutionary history of the taxa analyzed. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. All ambiguous positions were removed for each sequence pair. There were a total of 3600 positions in the final data set. The leftmost node represents a hypothetical most recent common ancestor. The above and bottom branches from this node represent lineages 1 and 2, respectively. Isolates are color coded according to the host: green, rainbow trout; red, Atlantic salmon; yellow, Chinook salmon; pink, pink salmon; teal, Grayling; gold, Coho salmon; orange, Eastern brook trout; brown, brown trout; gray, not known.

genetic diversity. The average pairwise nucleotide diversity (π) across the whole genome (3.15 Mb) within lineage 1 is 0.00005, corresponding to an average of 167 SNP differences between pairs of strains. This paucity of variation is indicative of a slow rate of evolution, recent common ancestry of the population with global dissemination or both. The seven isolates of lineage 2 (including the isolates recovered from the River Dee in 1960), were all isolated from the UK and Norway and are exclusively associated with the genus *Salmo*; six from *Salmo salar* (Atlantic salmon) and one from *Salmo trutta* (brown trout). The average number of SNP differences between lineage 2 isolates is 80, approximately half that of the lineage 1 isolates (π = 0.000025). The level of diversity between lineages is ~20-fold greater than the diversity within them, with an average inter-lineage difference of 2431 SNPs (π = 0.0008). The majority of the clades in the phylogeny are supported by high (>0.9) PPS (posterior probability scores) (See Figure 2; all clades from linage 2 have complete support as evidenced by PPS of 1), indicating the phylogenetic analysis has not been significantly confounded by recombination. To confirm this, we checked for the presence of recombination using the Phi test and splits decomposition as implemented in SplitsTree 4 (Graham *et al.*, 2005). We carried out this analysis on the two major lineages separately, as the relatively large distance between them would impair visual inspection of the network. The Phi test did not find significant evidence for recombination either for Lineage 1 ($P = 0.86$) or for Lineage 2 ($P = 0.1$). Although inspection of the splits decomposition networks reveals very little reticulation for Lineage 1, slight reticulation is suggested between the seven strains of Lineage 2 (Supplementary Figure S1). The evidence thus points to a near absence of recombination, although a very limited degree of recombination cannot be excluded in Lineage 2.

Inferring patterns of global transmission

Lineage 1 contains a mixture of North American and European isolates, whilst lineage 2 is restricted to Europe. We used BEAST (Drummond and Rambaut, 2007; Drummond *et al.*, 2012) for evolutionary analysis using an uncorrelated relaxed, lognormal clock. The mean clock rate was estimated as 3.324×10^{-4} mutations/site/year for our 3600 SNPs, corresponding to an overall rate of 3.8×10^{-7}

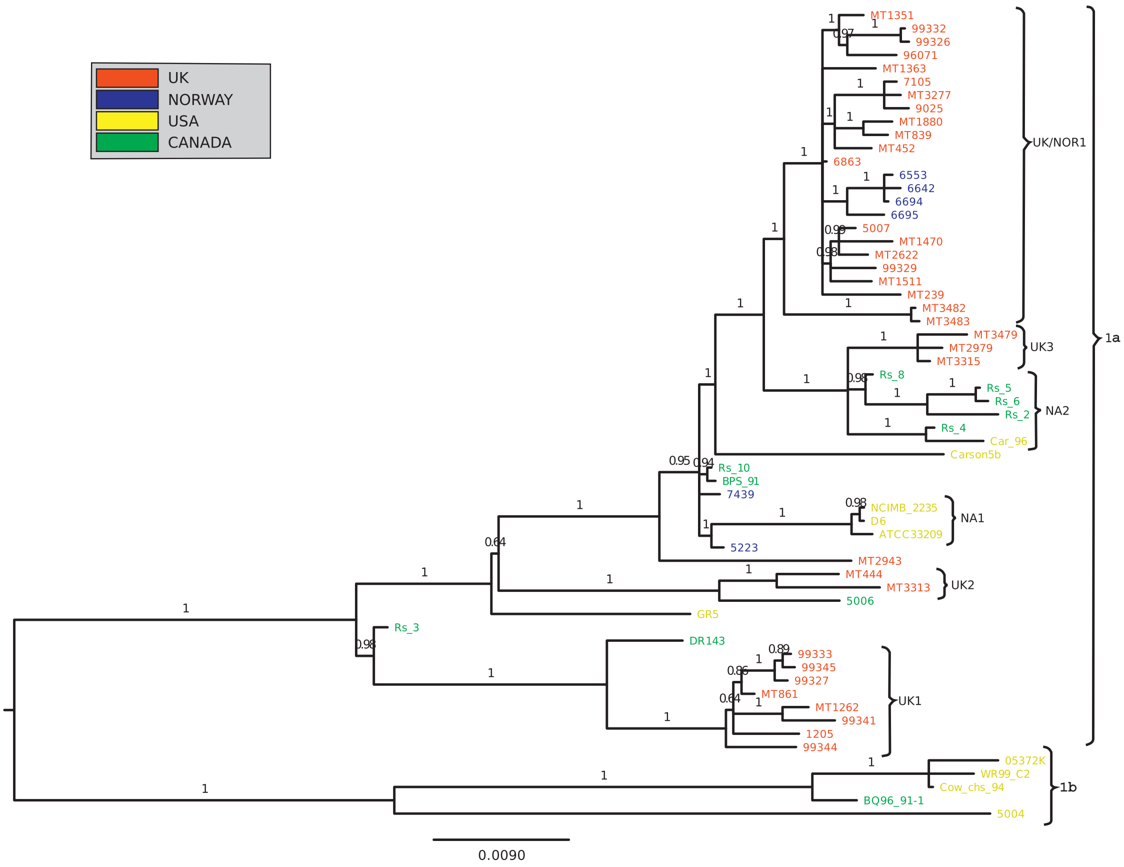


Figure 2 Detail of lineage 1. The phylogenetic tree was constructed as described under Figure 1. Posterior probability values are shown on each branch. For the detailed look at subgroup UK/NOR1, branches have been transformed so as to no longer represent evolutionary distance. Isolates are color coded according to their geographical origin: red, UK; blue, Norway; green, Canada; yellow, USA.

mutations/genome/year. This figure represents the mean rate and may not be representative of individual lineages. The time to most recent common ancestor of all samples was dated to ∼1239 years ago (95% CI 444–2720 years ago) (Supplementary Figure S2) However, it should be noted that the credible interval is rather wide leading to a not very precise point estimate; hence, caution should be exercised in considering this estimate. The lineages thus started to diverge at some point in time beyond this estimate. Lineage 1 can be further subdivided into lineage 1a and lineage 1b (Figure 2). Lineage 1b corresponds to five isolates from North America, consistent with a North American origin for this group. In contrast, Lineage 1a consists of isolates recovered from both North America and Europe. We note a general trend from the Lineage 1 tree that the most basal isolates tend to be of North American origin, supporting the view that this lineage emerged in North America and has more recently been transmitted to Europe rather than the other way around. For example, isolate Rs_3 was recovered from an Atlantic salmon in New Brunswick, Canada, and isolate DR143, which was isolated from a brook trout (*Salvelinus fontinalis*) from Alberta, Canada, both positioned in group 1a, basally to a cluster of eight isolates from the UK (labeled UK1 on Figure 2). The UK1 cluster may therefore represent a transmission event from Canada to the UK and, according to the BEAST analysis, this is most likely to have occurred ∼66 years ago (95% CI 40–120 years ago). The eight isolates corresponding to the UK1 cluster originate from fish farms in Wales, Scotland and England, indicating dissemination throughout UK farms subsequent to this transmission event that affected at least two commercial host species: Atlantic salmon (MT1262 and MT861) and rainbow trout (99 333, 99 327, 99 341, 99 344, 1205 and 99 345). The large cluster of 22 isolates from the UK and Norway (UK/NOR1) is also consistent with a single recent introduction ∼43 years ago (95% CI 30–65 years ago.), followed by rapid spread between the UK and Norway, affecting multiple host species. Finally, two small clusters of UK isolates are evident in Lineage 1a (UK2, UK3), which may also represent independent introductions and transmission across multiple host species within the UK.

The inferences above are based largely on the casual inspection of the tree. In order to investigate the origins and transmission history of the *R. salmoninarum* isolates within a more robust framework, we performed ancestral state reconstruction using Mesquite (Maddison and Maddison, 2011). We used both parsimony (Figure 3) and maximum-likelihood (Supplementary Figure S3)-based methods. In both cases, geographic origin was treated as a categorical variable with a uniform cost of switching. This parsimony analysis strongly supports the inferences discussed above in predicting a likely North American origin for the UK1, UK2, NA1, UK3 and NA2 groups. Although the immediate ancestor of the UK/NOR1 is predicted to originate from the UK (Figure 3; node A),

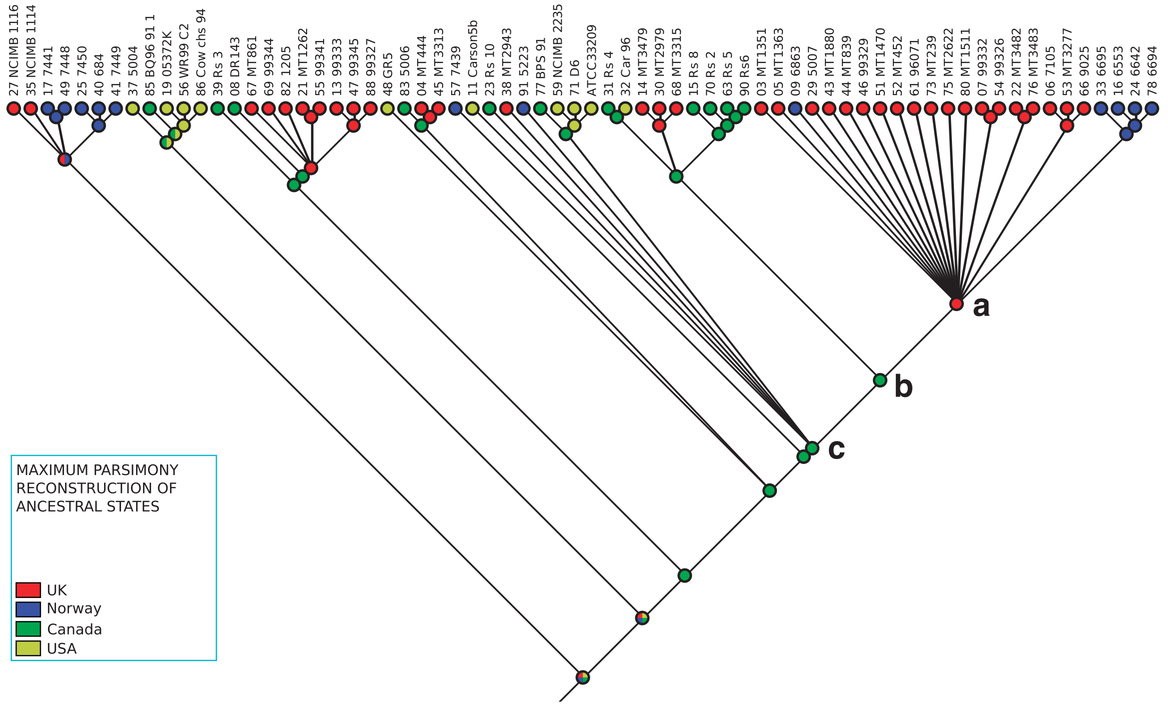


Figure 3 Ancestral state reconstruction of the geographical origin. Nodes in the trees have been estimated from a maximum parsimony evaluation of terminal values, where country of origin has been evaluated in an unordered, categorical matrix. The tree is not drawn to scale. The legend is as follows: red, UK; blue, Norway; green, Canada; yellow, USA. Nodes A, B and C are referenced in the text. The tree was calculated by using SNPs from open reading frames (ORFs) annotated as pseudogenes in the ATCC33209 genome.

the analysis points to a character switch before the emergence of this clade as the most parsimonious states at nodes B and C are of Canadian origin. This character switch thus reflects a transmission from North America to the UK.

We also considered proportional likelihoods of each state at the different internal nodes in the tree as given in Supplementary File SF2. This maximum-likelihood approach unsurprisingly confirms that the immediate ancestor of the UK/NOR1 group (node A in Figure 3) was very likely of UK origin (node 52 in SF2; proportional likelihood score for UK origin = 1.0). Moving one node back in the tree (node B in Figure 3) the proportional likelihoods of a UK and North American origin were found to be approximately equal (node 76 in SF2; proportional likelihood = 0.44 for UK and 0.48 for Canada). However, moving back one node (node C in Figure 3) the maximum-likelihood analysis provides far stronger support for a North American origin, and very weak evidence for a UK origin (node 77 in SF2; proportional likelihood = 0.05 for UK and 0.75 for Canada). This analysis therefore points to a switch (transmission) from North America to UK either between node C and node B (as suggested by maximum likelihood), or between node B and node A (as suggested by maximum parsimony) (Figure 3).

Detecting outbreaks

In addition to providing evidence concerning large-scale patterns of transmission, next-generation sequence data can also distinguish between samples of contemporaneous isolates recovered from a local setting. The technology allows the analyst to accurately assess whether or not concurrent presentations of disease form an outbreak or not. For instance, UK/NOR1 cluster isolates 6642, 6694 and 6553, recovered as part of disease outbreak investigations in the region of Sør-Trøndelag in 2008, were very closely related, differing at only five polymorphic sites. Strikingly, these three isolates also demonstrate free transmission between hosts; isolates 6642 and 6553 are from the same *S. salar* farm in Hemne, whereas 6694 and 6695 are from a nearby *O. mykiss* farm. In other cases the technology demonstrated that contemporaneous strains isolated from a single farm are not always epidemiologically linked, but may in some cases correspond to a mixture of the major lineages circulating throughout the UK and beyond. For example, isolates 99329, 99333 and 99345 were all isolated in 1998 from the same site in Wales but the latter two isolates belong to the cluster UK1, whilst 99323 belong to UK/NOR1.

Discussion

The very low density of polymorphisms and almost total conservation of gene content confirm that *R. salmoninarum* is a highly clonal pathogen. A very low level of genome diversity has been noted in other highly specialized intracellular bacterial pathogens, notably *Mycobacterium tuberculosis*, which is phylogenetically related to *R. salmoninarum* and also causes a chronic granulomatous disease. Our estimate for a mean clock rate of 3.8×10^{-7} mutations/bp/year is higher than a recent estimate for *M. tuberculosis* of 7.3×10^{-8} (Ford *et al.*, 2011. Scaled up from mutations/bp/day), and notably much higher than rates for *Yersinia pestis* that has been estimated at 8.6×10^{-9} (Morelli *et al.*, 2010). However, *R. salmoninarum* appears to mutate much more slowly than other major human pathogens such as *Staphylococcus aureus*, for which the rate was recently reported as 1.3×10^{-6} (Holden *et al.*, 2013).

Although isolates were recovered from many different species—Atlantic salmon, grayling, brook trout, rainbow trout and other *Oncorhynchus spp*.—they are almost indistinguishable genetically, indicating free interspecies transmission and broad virulence properties. Our data therefore suggest that previously observed differences in host susceptibility to BKD (Starliper *et al.*, 1997) are likely to reflect host and/or environmental factors rather than variation within the pathogen. The limited genetic diversity of UK isolates in particular may be related to the controls implemented in response to the 1937 Diseases of Fish Act, which prohibited import of live salmonids into the UK and also made it illegal to import salmonid ova and any live freshwater fish species without a license (Hill, 1996). This has likely limited associated importation of pathogens such as *R. salmoninarum,* which could be one of the reasons that a relatively limited range of subtypes of this, and another important bacterial pathogen, *Yersinia ruckeri* (Davies, 1991; Wheeler *et al.,* 2009) appear to be present within UK farmed fish. It has previously been noted that the main subtype of *Y. ruckeri* circulating in farmed UK rainbow trout also shows limited diversity and is different from the main strains circulating in Europe (Davies, 1991; Wheeler *et al.*, 2009), suggesting these controls have restricted pathogen exchange between the UK and mainland Europe aquaculture production systems.

The data reveal evidence concerning long-term geographical structuring and global transmission. The construction of a high-resolution SNP-based phylogeny robustly resolved two major lineages of *R. salmoninarum*, lineage 1 and lineage 2. Given the extremely low divergence within lineage 1, it can be speculated that its rapid geographical and ecological dispersal, through different oceans and species of salmonids, has likely been facilitated by anthropogenic means. Commercial activities over the last 150 years have provided ample opportunities for such a spread to occur. Live salmonids and ova have been traded on a global scale for aquaculture, recreational angling and fishery stock enhancement purposes since the mid-19th century (Halverson, 2010). The last 40 years in particular have witnessed increasingly intensive and vertically integrated production, initially of rainbow trout and latterly of Atlantic salmon. Eggs are typically produced in dedicated broodstock facilities before being moved to hatcheries for production of fry, which are then moved again to ongrowing sites (either seawater cages for Atlantic salmon, or larger tanks or raceways for rainbow trout). As with other intensive livestock production systems, frequent transport means that breaches in biosecurity can lead to the transmission of pathogens between premises (Plarre *et al.*, 2012). Furthermore, evidence exists that other highly clonal bacterial diseases of salmonids have spread between continents in similar ways, likely via egg and live fish movements. (Davies, 1991; Garcia *et al.*, 2000; Wheeler *et al.*, 2009).

In contrast to the global distribution of lineage 1, the absence of North American isolates in lineage 2 suggests that this lineage may represent a long-term endemic disease in European waters. Lineage 2 isolates all came from the North Sea area (including tributaries). The early outbreaks in Norway in the 1980s, all of lineage 2 origins, were reportedly in restocking hatcheries based on the capture and stripping of wild Atlantic salmon brood stock, and thus may have been caused by the pathogen present in wild populations. It is notable that the implementation of strict biosecurity measures and screening efforts in the late 1980s coincided with a significant reduction in outbreaks of BKD in Norway (Wiens and Dale, 2009; Johansen *et al.*, 2011). The origin of lineage 1 is more unclear. The estimated phylogenetic division between the two main lineages (∼1000 years ago) clearly pre-dates commercial activity. It is possible that lineage 1 emerged independently within a geographically or ecologically isolated population of Atlantic salmonid before being transferred into Pacific salmonid populations. In support of this, BKD was not reported as a problem in Pacific salmon hatcheries on the West Coast of the US (Earp *et al.*, 1953) until much later than East Coast hatcheries (Belding and Merrill, 1935). This is also consistent with the suggestion that the pathogen has co-evolved for longer with Atlantic salmonids (*Salmo* and *Salvelinus*), thus potentially explaining why Pacific salmon species are reportedly more susceptible to this pathogen (Evenden *et al.*, 1993; Starliper *et al.*, 1997). The relatively limited recovery of lineage 2 isolates as opposed to lineage 1 isolates in the UK and Norway was noteworthy. Further research is needed to establish whether any genetic advantage in either of the lineages can explain this asymmetry in distribution.

In general, the accuracy of determination of the relatedness of isolates demonstrated here was not possible with previously employed typing methods (Grayson *et al.,* 1999, 2000; Rhodes *et al.,* 2000). As an example, previous random amplified polymorphic DNA-based analysis (Grayson *et al.*, 2000) grouped the DR143 and Cow_chs_94 isolates on the same terminal branch, whereas our dated phylogenetic trees reveal that the most recent common ancestor of these isolates existed about 360 years ago. (95% CI ∼150–700 years ago). This clearly illustrates the superiority of the next-generation sequence technology for outbreak determination problems.

Our data and analysis have delineated two major European clusters within Lineage 1a (UK1 and UK/NOR1) that emerged <70 years ago, probably as a result of transmission from North America, and have spread between fish farms in the UK and Norway infecting both rainbow trout and Atlantic salmon. Support for this comes from the observation that the North American lineage 1 isolates tend to be more diverse (and basal) than the European lineage 1 isolates, suggesting the former reflect the ‘ancestral’ population and the latter consist of independent introductions (bottlenecks). Our reconstruction of ancestral states analysis also points to the North American origin of most sub-lineages in Lineage 1. In summary, this points to transatlantic commercial activity as a likely factor in the European emergence and spread of lineage 1. However, it should be emphasized that all this refers only to the major European lineage 1 clades that we have sampled, and it is not possible to extrapolate from this to general conclusions about rates of transmission in either direction, and significant transmission from Europe to North America through commercial activity may also have taken place.

Conclusion

The application of whole-genome SNP-based comparisons has offered a range of insights into the likely microevolutionary relationships of this important fish pathogen that hitherto would not have been possible. The analysis reveals an unexpected deep phylogenetic division in the population, hinting at historical allopatry, evidence for transatlantic transmission and spread over the scale of decades and even proof of principle that the approach can be used to identify single outbreak strains on a very local scale. It is recommended that this methodology should also be applied for studies of other veterinary pathogens and environmental microorganisms, particularly those with very limited genetic intraspecies variation.

Acknowledgements

The UK’s Department for Environment Food and Rural Affairs provided funding through Projects FB002 and an interagency grant DP258. Joel Caren, Inger Lise Engen and Georgina Rimmer provided valuable technical support. We thank Dr Bill Keleher for providing isolates from New Brunswick, Canada. Dr Chris Watkins from The Genome Analysis Centre (TGAC), Norwich, UK provided expert assistance with sequencing the isolates. We are grateful to the reviewers for their helpful comments.

References

Akaike H. (1974). A new look at the statistical model identification. *IEEE Trans Auto Contrl* **19**: 716–723.

Alexander SM, Grayson TH, Chambers EM, Cooper LF, Barker GA, Gilpin ML. (2001). Variation in the spacer regions separating tRNA genes in *Renibacterium salmoninarum* distinguishes recent clinical isolates from the same location. *J Clin Microbiol* **39**: 119–128.

Arkoosh MR, Clemons E, Kagley AN, Stafford C, Glass AC, Jacobson K *et al.* (2004). Survey of pathogens in juvenile salmon *Oncorhynchus* Spp. migrating through Pacific Northwest estuaries. *J Aquat Anim Health* **16**: 186–196.

Banner CR, Long JJ, Fryer JL, Rohovec JS. (1986). Occurrence of salmonid fish infected with *Renibacterium salmoninarum* in the Pacific Ocean. *J Fish Dis* **9**: 273–275.

Belding DL, Merrill B. (1935). A preliminary report upon a hatchery disease of the salmonidae. *Trans Am Fish Soc* **65**: 76–84.

Chambers E, Gardiner R, Peeler EJ. (2008). An investigation into the prevalence of *Renibacterium salmoninarum* in farmed rainbow trout, *Oncorhynchus mykiss* (Walbaum), and wild fish populations in selected river catchments in England and Wales between 1998 and 2000. *J Fish Dis* **31**: 89–96.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A *et al.* (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**: 1422–1423.

Darling ACE, Mau B, Blattner FR, Perna NT. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* **14**: 1394–1403.

Davies RL. (1991). Clonal analysis of *Yersinia ruckeri* based on biotypes, serotypes and outer membrane protein-types. *J Fish Dis* **14**: 221–228.

Drummond A, Rambaut A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Bio* **7**: 214.

Drummond AJ, Suchard MA, Xie D, Rambaut A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* **29**: 1969–1973.

Earp BJ, Ellis CH, Ordal EJ. (1953). Kidney disease in young salmon (No. 1). Special Report Ser. No. 1, Department of Fisheries, Washington State, USA, p 73.

Elliott DG, Pascho RJ, Bullock GL. (1989). Developments in the control of bacterial kidney disease of salmonid fishes. *Dis Aquat Org* **6**: 201–215.

Evelyn TPT. (1977). An improved growth medium for the kidney disease bacterium and some notes on using the medium. *Bull Off Int Epizoot* **87**: 511–513.

Evelyn TPT. (1993). Bacterial kidney disease—BKD. In: Inglis V, Roberts RJ, Bromage NR (eds) *Bacterial Diseases of Fish*. Blackwell Scientific Publications: Oxford, UK, pp 177–195.

Evelyn TPT, Prosperi-Porta L, Ketcheson JE. (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 Aquat Org* **1**: 197–202.

Evenden AJ, Grayson TH, Gilpin ML, Munn CB. (1993). *Renibacterium salmoninarum* and bacterial kidney disease—the unfinished jigsaw. *Ann Rev Fish Dis* **3**: 87–104.

Ford CB, Lin PL, Chase MR, Shah RR, Iartchouk O, Galagan J *et al.* (2011). Use of whole-genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. *Nat Genet* **43**: 482–486.

Fryer JL, Lannan CN. (1993). The history and current status of *Renibacterium salmoninarum*, the causative agent of bacterial kidney disease in Pacific salmon. *Fish Res* **17**: 15–33.

Fryer JL, Sanders JE. (1981). Bacterial kidney disease of salmonid fish. *Ann Rev Microbiol* **35**: 273–298.

Garcia JA, Larsen JL, Dalsgaard I, Pedersen K. (2000). Pulsed-field gel electrophoresis analysis of Aeromonas salmonicida ssp. salmonicida. *FEMS Microbiol Lett* **190**: 163–166.

Graham J, McNeney B, Seillier-Moiseiwitsch F. (2005). Stepwise detection of recombination breakpoints in sequence alignments. *Bioinformatics* **21**: 589–595.

Grayson TH, Atienzar FA, Alexander SM, Cooper LF, Gilpin ML. (2000). Molecular diversity of *Renibacterium salmoninarum* isolates determined by randomly amplified polymorphic DNA analysis. *Appl Environ Microbiol* **66**: 435–438.

Grayson TH, Cooper LF, Atienzar FA, Knowles MR, Gilpin ML. (1999). Molecular differentiation of *Renibacterium salmoninarum* isolates from worldwide locations. *Appl Environ Microbiol* **65**: 961–968.

Grayson TH, Cooper LF, Wrathmell AB, Roper J, Evenden AJ, Gilpin ML. (2002). Host responses to *Renibacterium salmoninarum* and specific components of the pathogen reveal the mechanisms of immune suppression and activation. *Immunology* **106**: 273–283.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**: 307–321.

Gutenberger SK, Duimstra JR, Rohovec JS, Fryer JL. (1997). Intracellular survival of *Renibacterium salmoninarum* in trout mononuclear phagocytes. *Dis Aquat Org* **28**: 93–106.

Halverson A. (2010). *An Entirely Synthetic Fish: How Rainbow Trout Beguiled America and Overran the World*. Yale University Press: New Haven, UK.

Harris SR, Feil EJ, Holden MTG, Quail MA, Nickerson EK, Chantratita N *et al.* (2010). Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**: 469–474.

Hill B. National legislation in Great Britain for the control of fish diseases (1996). *Rev Sci Tech* **15**: 633–645.

Holden MTG, Hsu LY, Kurt K, Weinert LA, Mather AE, Harris SR *et al.* (2013). A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic. *Genome Res* **23**: 653–664.

Holt KE, Baker S, Weill F-X, Holmes EC, Kitchen A, Yu J *et al.* (2012). *Shigella sonnei* genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nat Genet* **44**: 1056–1059.

Ji Y, Shi Y, Ding G, Li Y. (2011). A new strategy for better genome assembly from very short reads. *BMC Bioinformatics* **12**: 493.

Johansen LH, Jensen I, Mikkelsen H, Bjørn PA, Jansen PA, Bergh Ø. (2011). Disease interaction and pathogens exchange between wild and farmed fish populations with special reference to Norway. *Aquaculture* **315**: 167–186.

Kent ML, Traxler GS, Kieser D, Richard J, Dawe SC, Shaw RW *et al.* (1998). Survey of salmonid pathogens in ocean-caught fishes in British Columbia, Canada. *J Aquat Anim Health* **10**: 211–219.

Lewis PO. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst Biol* **50**: 913–925.

Li H, Ruan J, Durbin R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**: 1851–1858.

Maddison WP, Maddison DR. (2011). Mesquite: a modular system for evolutionary analysis. Version 2.75 http://mesquiteproject.org.

Meyers TR, Korn D, Glass K, Burton T, Short S, Lipson K *et al.* (2003). Retrospective analysis of antigen prevalences of *Renibacterium salmoninarum* (Rs) detected by enzyme-linked immunosorbent assay in Alaskan Pacific salmon and trout from 1988 to 2000 and management of Rs in hatchery chinook and coho Salmon. *J Aquat Anim Health* **15**: 101–110.

Morelli G, Song Y, Mazzoni CJ, Eppinger M, Philippe R, Wagner DM *et al.* (2010). Phylogenetic diversity and historical patterns of pandemic spread of *Yersinia pestis*. *Nat Genet* **42**: 1140–1143.

Murray AG, Hall M, Munro LA, Wallace IS. (2011). Modelling management strategies for a disease including undetected sub-clinical infection: bacterial kidney disease in Scottish salmon and trout farms. *Epidemics* **3**: 171–182.

Murray AG, Munro LA, Wallace IS, Allan CET, Peeler EJ, Thrush MA. (2012). Epidemiology of *Renibacterium salmoninarum* in Scotland and the potential for compartmentalised management of salmon and trout farming areas. *Aquaculture* **324–325**: 1–13.

Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S *et al.* (2011). Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **477**: 462–465.

Paradis E, Claude J, Strimmer K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**: 289–290.

Paterson WD, Gallant C, Desautels D, Marshall L. (1979). Detection of bacterial kidney disease in wild salmonids in the Margaree river system and adjacent waters using an indirect fluorescent antibody technique. *J Fish Res Board Can* **36**: 1464–1468.

Plarre H, Nylund A, Karlsen M, Brevik Ø, Sæther PA, Vike S. (2012). Evolution of infectious salmon anaemia virus (ISA virus). *Arch Virol* **157**: 2309–2326.

Rhodes LD, Grayson TH, Alexander SM, Strom MS. (2000). Description and characterization of IS*994*, a putative IS*3* family insertion sequence from the salmon pathogen *Renibacterium salmoninarum*. *Gene* **244**: 97–107.

Richards R. (2011). A strategy for the control of Bacterial Kidney Disease (BKD) in Great Britainhttp://www. scotland.gov.uk/Resource/Doc/1062/0114801.pdf.

Ronquist F, Huelsenbeck JP. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572–1574.

Schatz M, Phillippy A, Shneiderman B, Salzberg S. (2007). Hawkeye: an interactive visual analytics tool for genome assemblies. *Genome Biol* 8: R34.

Schliep KP. (2011). phangorn: phylogenetic analysis in R. *Bioinformatics* **27**: 592–593.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol İ. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Res* **19**: 1117–1123.

Smith IW. (1964). *The Occurrence and Pathology of Dee Disease*. Freshwater and Salmon Fisheries Research, 34. Her Majesty's Stationery OfficeEdinburgh, UK.

Starliper CE, Smith DR, Shatzer T. (1997). Virulence of *Renibacterium salmoninarum* to salmonids. *J Aquat Anim Health* **9**: 1–7.

Tavaré S. (1986). Some probabilistic and statistical problems in the analysis of DNA Sequences. *Lect Math Life Sci* **17**: 57–86.

Treangen TJ, Sommer DD, Angly FE, Koren S, Pop M. (2011). Next generation sequence assembly with AMOS. *Curr Protoc Bioinform* **33**: 11.8.1–11.8.18.

Wheeler RW, Davies RL, Dalsgaard I, Garcia J, Welch TJ, Wagley S *et al.* (2009). *Yersinia ruckeri* biotype 2 isolates from mainland Europe and the UK likely represent different clonal groups. *Dis Aquat Org* **84**: 25–33.

Wiens GD. (2011). Bacterial Kidney Disease (Renibacterium salmoninarum). *Fish Diseases and Disorders: Volume 3: Viral, Bacterial and Fungal Infections 3*, 2nd ednWoo PTK, Bruno DW (eds) 338–374.

Wiens GD, Dale OB. (2009). *Renibacterium salmoninarum* p57 antigenic variation is restricted in geographic distribution and correlated with genomic markers. *Dis Aquat Org* **83**: 123.

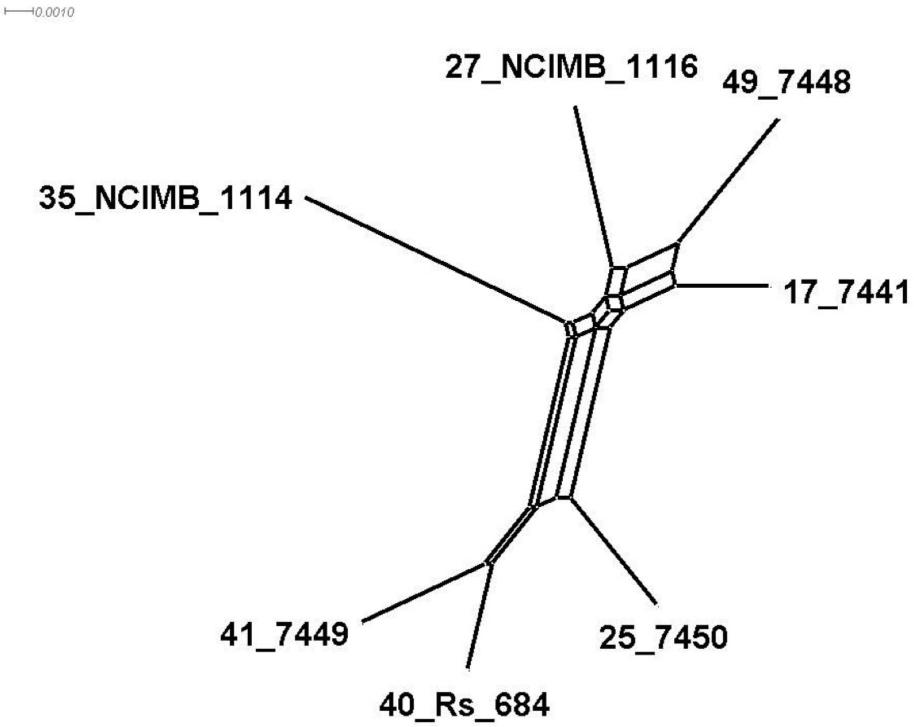
Wiens GD, Pascho R, Winton JR. (2002). A single Ala139-to-Glu substitution in the *Renibacterium salmoninarum* virulence-associated protein p57 results in antigenic variation and is associated with enhanced p57 binding to chinook salmon leukocytes. *Appl Environ Microbiol* **68**: 3969–3977.

Wiens GD, Rockey DD, Wu Z, Chang J, Levy R, Crane S *et al.* (2008). Genome sequence of the fish pathogen *Renibacterium salmoninarum* suggests reductive evolution away from an environmental Arthrobacter ancestor. *J Bacteriol* **190**: 6970–6982.

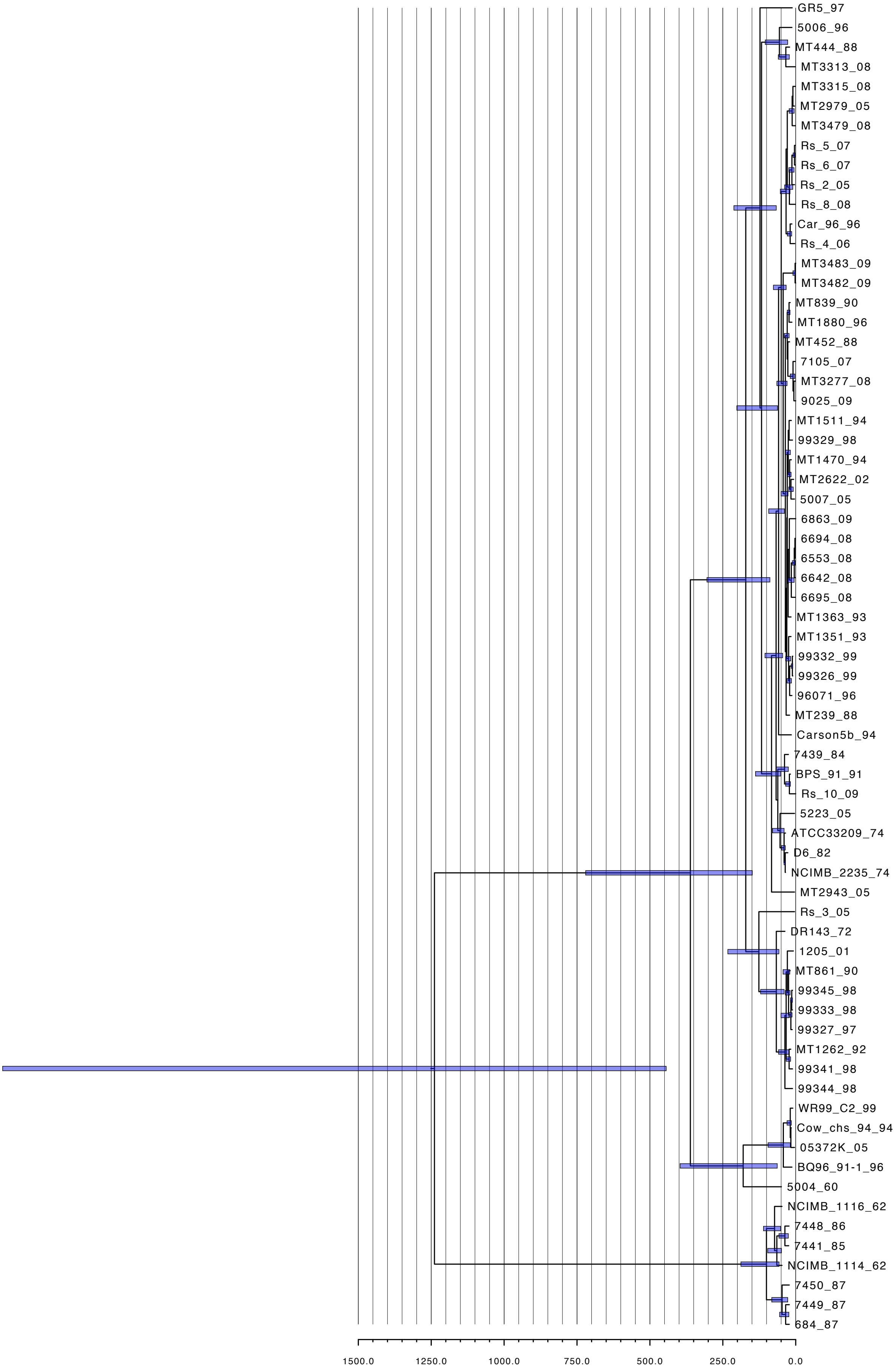
Young CL, Chapman GB. (1978). Ultrastructural Aspects of the Causative Agent and Renal Histopathology of Bacterial Kidney Disease in Brook Trout (*Salvelinus fontinalis*). *J Fish Res Board Can* **35**: 1234–1248.

Zerbino DR, Birney E. (2008). Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res* **18**: 821–829.

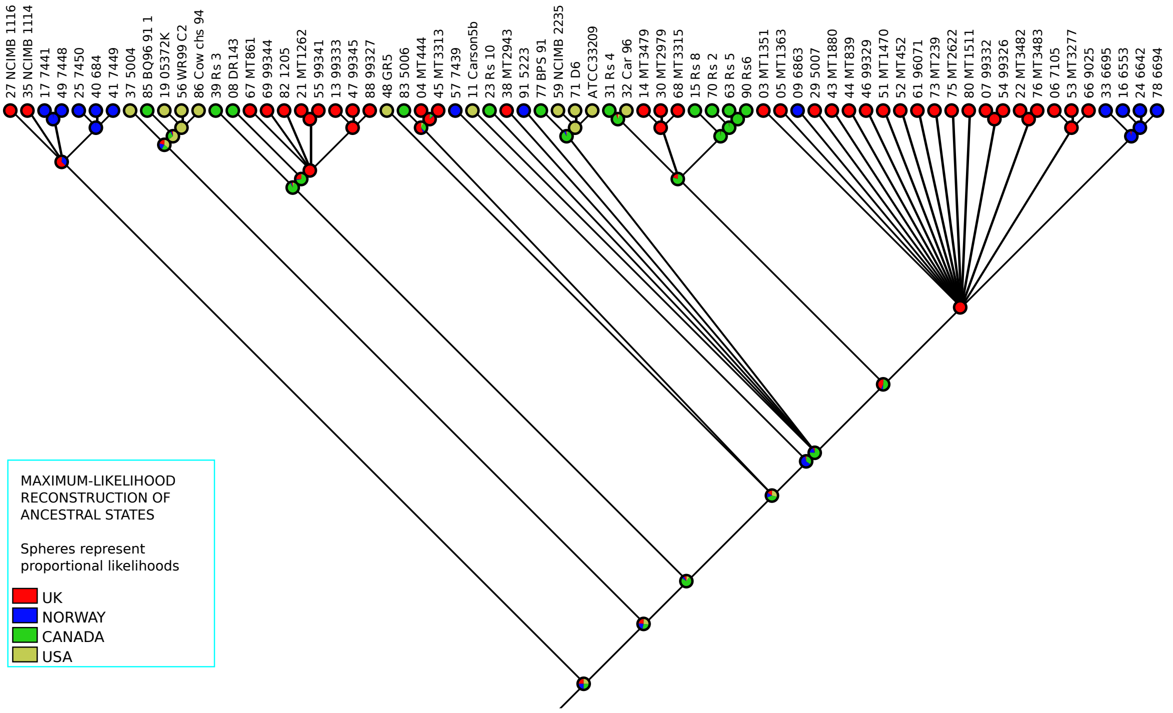
Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)



Supplementary figure 1: Recombination analysis of lineage 2, showing slight reticulation



Supplementary figure 2: Dated Bayesian phylogeny constructed via BEAST. An uncorrelated lognormal relaxed clock model was used to construct the tree based on the concatenated alignment of all the SNPs. The horizontal bars show the 95% HPD intervals for the divergence time estimates.



Supplementary figure 3: Ancestral state reconstruction of node values in the pseudogene-SNP-based tree, using a maximum-likelihood approach under the one-parameter Markov k-state model. Nodes are drawn to represent the proportional likelihoods.

Supplementary tables 1 and 2 are large files and have not been included in this thesis.

They are freely available online (10.1038/ismej.2013.186)

Supplementary Table 3: Statistics related to read output, average read depth and percentage of total reads mapped to the reference genome for each individual isolate.

Isolate	Number of reads (paired)	Avg read depth	% of total reads mapped *
MT1351	21,013,832	1307.7	98.2%
MT444	13,140,953	792.3	95.1%
MT1363	9,232,914	430.9	73.6%
7105	14,094,137	871.5	97.6%
99332	24,689,490	1526.3	97.5%
DR143	20,465,983	1255.9	96.8%
6863	21,077,399	912.6	68.3%
Carson 5b	2,544,483	116.1	72.0%
99333	17,492,167	1072.7	96.7%
MT3479	11,093,489	688.3	97.9%
Rs 8	21,427,970	1163	85.6%
6553	6,224,393	381.8	96.8%
7441	23,649,886	1444.7	96.4%
05372K	8,561,496	512.9	94.5%
MT1262	11,278,599	697.4	97.6%
MT3482	10,658,806	662.7	98.1%
Rs 10	955,147	57.2	94.5%
6642	12,203,536	627.8	81.2%
7450	14,443,651	867	94.7%
NCIMB 1116	14,052,993	780.7	87.6%
5007	15,881,479	978.7	97.2%
MT2979	19,390,144	1207.3	98.2%
Rs 4	678,134	31.7	73.7%
Car 96	11,857,707	723.9	96.3%
6695	12,610,961	475.4	59.5%
NCIMB 1114	11,695,333	717.5	96.8%
5004	22,304,012	1371.9	97.0%
MT2943	15,591,688	965.8	97.7%
Rs 3	24,449,576	1359.9	87.7%
684	18,119,160	1099.3	95.7%
7449	20,814,378	1243.9	94.3%
MT1880	9,043,182	522.3	91.1%
MT839	10,324,148	590.8	90.3%
MT3313	16,012,176	992.2	97.8%
99329	29,748,569	1841.2	97.6%
99345	18,041,657	1051.8	92.0%
GR5	17,064,552	1033.6	95.6%
7448	19,658,933	785.9	63.1%
MT1470	12,553,080	737.9	92.7%
MT452	6,281,721	379.7	95.4%
MT3277	13,770,986	694.8	79.6%
99326	15,899,635	982.1	97.4%
99341	16,163,349	1001.5	97.8%
Rs WR99 c2	22,883,403	1380.2	95.2%
7439	6,518,723	391	94.6%
NCIMB 2235	11,563,672	710.8	97.0%
96071	5,516,999	335.9	96.1%
Rs 5	8,016,378	404.2	79.5%
9025	17,616,289	1090	97.6%
MT861	12,740,758	712.2	88.2%
MT3315	15,768,868	948.6	94.9%
99344	23,133,452	1430	97.5%
Rs 2	8,522,244	494.4	91.5%
Rs D6	11,344,343	683.5	95.1%
MT239	11,954,353	730.6	96.4%
MT2622	22,714,850	908.2	63.1%
MT3483	32,367,967	2012.5	98.1%
BPS 91	26,070,672	1590.7	96.3%
6694	9,045,870	553.7	96.6%
MT1511	10,442,994	650.7	98.3%
5006	22,209,296	1181.4	83.9%
BQ96 91-1	14,882,427	570.9	60.5%
cow-chs-94	6,027,401	361.8	94.7%
99327	9,585,455	567.1	93.3%
Rs 6	12,074,705	543.8	71.1%
5223	19,465,930	1178.9	95.5%
Average	14,577,954	862	93.3%

* - This is the percentage of the total reads that mapped after quality filtering

II

Bioinformatics Advance Access published March 11, 2015

Bioinformatics, 2015, 1–8
doi: 10.1093/bioinformatics/btv070
Advance Access Publication Date: 1 February 2015
Original Paper

Genome analysis

CNOGpro: detection and quantification of CNVs in prokaryotic whole-genome sequencing data

Ola Brynildsrud[1,]*, Lars-Gustav Snipen[2] and Jon Bohlin[3]

[1]Section for Biostatistics and Epidemiology, Norwegian University of Life Sciences (NMBU), Oslo, [2]Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences (NMBU), Ås and [3]Norwegian Institute of Public Health, Division of Epidemiology, 0403 Oslo, Norway

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on April 22, 2014; revised on January 7, 2015; accepted on January 28, 2015

Abstract

Motivation: The explosion of whole-genome sequencing (WGS) as a tool in the mapping and understanding of genomes has been accompanied by an equally massive report of tools and pipelines for the analysis of DNA copy number variation (CNV). Most currently available tools are designed specifically for human genomes, with comparatively little literature devoted to CNVs in prokaryotic organisms. However, there are several idiosyncrasies in prokaryotic WGS data. This work proposes a step-by-step approach for detection and quantification of copy number variants specifically aimed at prokaryotes.
Results: After aligning WGS reads to a reference genome, we count the individual reads in a sliding window and normalize these counts for bias introduced by differences in GC content. We then investigate the coverage in two fundamentally different ways: (i) Employing a Hidden Markov Model and (ii) by repeated sampling with replacement (bootstrapping) on each individual gene. The latter bypasses the complex problem of breakpoint determination. To demonstrate our method, we apply it to real and simulated WGS data and benchmark it against two popular methods for CNV detection. The proposed methodology will in some cases represent a significant jump in accuracy from other current methods.
Availability and implementation: *CNOGpro* is written entirely in the R programming language and is available from the CRAN repository (http://cran.r-project.org) under the GNU General Public License.
Contact: ola.brynildsrud@nmbu.no
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Copy number variation (CNV) is a type of structural variation that refers to any abnormality in the frequency at which a DNA sequence occurs in a genome. It is a critical component of the genetic variability of organisms (Alkan *et al.*, 2011). CNVs can include duplications (sometimes referred to as amplifications) or deletions of a particular stretch of sequence. Considerable research has been carried out on the relatively easy problem of determining whether a strain contains a gene or not, but less work has been devoted to the more complex problem of measuring non-zero variation in the gene copy number. The phenotypical implications of CNV are also less clear than those resulting from a functional deletion, although CNVs are known to be a source of important genetic variation in both humans (Stranger *et al.*, 2007) and bacteria (Riehle *et al.*, 2001). Research in humans has typically focused on the role of CNVs in cancer and inherited disease (Hastings *et al.*, 2009), and there is also evidence of adaptive duplications (Cooper *et al.*, 2007). Less work has been devoted to exploring the role of CNVs in prokaryotes. However, bacteria display substantial variation in gene copy numbers, and the extra cost of maintaining a redundant gene comes with the payoff of a selective advantage under certain environmental conditions (Klappenbach *et al.*, 2000; Kondrashov, 2012).

The most common way of testing the hypothesis of no CNV is by examining the read counts (RCs) along the chromosome after alignment of the sequenced reads to a reference genome or de novo assembly. The argument is that the RC in any non-overlapping, equally sized bin can be considered as a stochastic variable with a particular probability distribution and that there is an inherent proportionality between the expected value (mean) of this variable and the underlying copy number (Alkan *et al.*, 2009; Campbell *et al.*, 2008; Medvedev *et al.*, 2009), analogous to principles of CNV estimation in array Comparative Genome Hybridization technology. However, RCs are affected by biases that must be corrected for before any valid conclusions can be made.

There are currently at least 48 available tools for the discovery of CNVs from next-generation sequencing data (for a detailed runthrough, see Zhao *et al.*, 2013). Of these, few attempt to quantify the number of copies of any particular chromosomal segment (Klambauer *et al.*, 2012). Furthermore, most of these tools are designed with a diploid, human setting in mind. This presents a number of problems when applying these tools to prokaryotes. Prokaryotes are organisms of indefinite ploidy, and there are some additional allowances on the valid copy number outcomes when compared with diploid genomes. Although for human genomes a copy number of 1.5 would mean that the segment in question had two copies on one chromosome but one on the other, in bacteria the interpretation would vary from species to species depending on the number of sets of chromosomes, both complete and incomplete, it maintains. Complicating matters even further, the copy number result may vary due to bacterial growth mechanics. In fact, any nonnegative decimal copy number could make sense for bacteria, because bacteria growing under exponential growth conditions are able to replicate their chromosomes faster than they can divide (Pecoraro *et al.*, 2011). This is reflected in DNA sequencing data, with copy numbers representing a mixture of the discrete number of chromosomes that the bacterium maintains at the stationary, nondividing phase and a fractional number that stems from the fact that cells are in different phases of binary fission cycle.

This article presents a tool for the discovery of CNVs from prokaryote-origin whole-genome sequencing (WGS) data. We have developed an R package called *CNOGpro*, which is an acronym for ‘Copy Numbers of Genes in prokaryotes.’ The main purpose of the tool is to quickly estimate the number of copies of any gene or intergenic segment in a resequencing experiment. *CNOGpro* supports rapid calling of copy number using a hidden Markov model and additionally allows for the construction of confidence intervals around copy number estimates by bootstrapping. Although several publicly available CNV analysis tools designed for work on human-origin data can, with varying amounts of tinkering, also accept prokaryote data, to our knowledge no existing tool for CNV analysis focuses specifically on prokaryote data.

2 Materials and methods

2.1 Data preparation

The first step of RC-based CNV discovery consists of quality control of the sequencing data, including filtering of poor and uninformative data, thereafter mapping the reads onto the backbone of some related genome. [If the reference and the test organisms are too distantly related, we are introducing bias (Nijkamp *et al.*, 2012).] The filtering should be informed by such parameters as total coverage, average quality of reads, frequency of ambiguous characters (e.g. ‘N’) and quality distribution according to read length [Easily checked with programs such as FASTQC (Babraham) Bioinformatics. http://bioinformatics.babraham.ac.uk/projects/fastqc]. It is especially important to remove polymerase chain reaction (PCR) duplicates introduced in the sequencing, easily performed by the rmdup command in samtools (Li *et al.*, 2009) or the DupRecover script of Zhou *et al.* (2014). As for aligning reads from the resequencing experiment to a reference sequence, Bowtie (Langmead *et al.*, 2009) or Maq (http://maq.sourceforge.net) are both reliable and recommended third-party software solutions. In this article, we use a complete genome, but *CNOGpro* should also accept draft genomes as reference, as long as they adhere to the GenBank flat file format and are parsed contig-wise into the pipeline.

Next, one needs to apply some counting scheme on the coverage metrics. For independent observations in a number series, we can only count each read once. One way of doing this is by counting each read at its leftmost end, i.e. the lowest chromosomal coordinate to which the read maps. This information is available in the default output for the SAMtools binary alignment/map (.bam) format and can be used to create best-hit read location files. (Details about this procedure can be found in *CNOGpro’s* manual, provided in the R package.) Figure 1 shows the workflow of *CNOGpro*’s methods.

2.2 Counting reads

RCs are then made in neighboring, non-overlapping windows. (Referred to as RCs, observations or coverage interchangeably in the following chapters.) Each count represents the number of reads that have start points within that respective window. The average coverage and the desired sensitivity/specificity ratio should determine the window size. An average number of reads equal to 20–30 in each window is appropriate (Abyzov *et al.*, 2011). We have noted only a very slight drop in specificity when the average coverage is lowered from 100× to 20× and a more moderate drop when coverage is lowered to 10× (respective numbers: 99.96%, 99.7% and 99.0%; see Supplementary Table ST3). Long windows will make sure few false-positive CNVs are called, but one may also miss local coverage variation, which could be suggestive of small CNVs. The reverse is true for short windows. As a rule of thumb, shortening window length improves sensitivity at only marginal specificity cost and improving average coverage increases specificity and may increase sensitivity

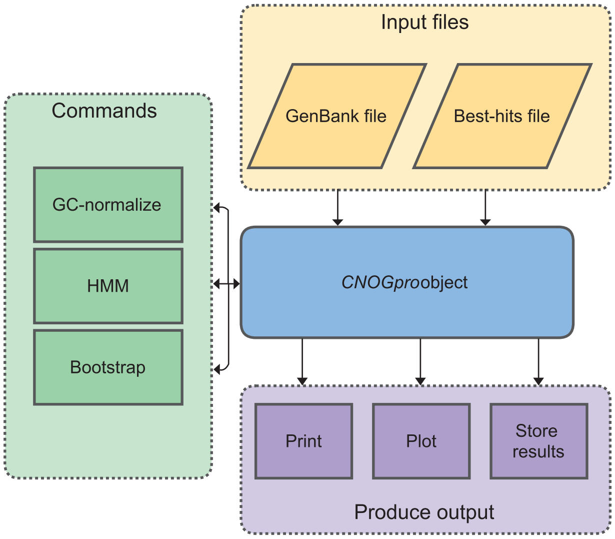


Fig. 1. Workflow diagram of *CNOGpro*

slightly. This is at least valid within window ranges 30–200 and coverage ranges 10×–1000×.

2.3 Biases and normalization

If the null hypothesis of no CNV between the reference organism and the data was true and there were no sources of bias, the expected RC in each window throughout the chromosome could be considered as independent and identically distributed random variables. Because of a number of known and unknown biases that influence the coverage, this is not the case. Therefore, before statistical inference is attempted, normalization must be performed on the RC metric. With flawless normalization, only the underlying CNV number of a genomic segment determines the expected RC value in a window. [The term segment is hereafter used to describe any continuous stretch of DNA with associated RC observations, parsed according to genes (including RNA genes) and intergenic stretches in the reference organism.]

Significant sources of bias in RC observations are local GC content of the chromosome (Dohm *et al*., 2008), GC content of the sequencing probes (Diskin *et al*., 2008), genomic mappability (Derrien *et al*., 2012; Lee and Schatz, 2012) and copy number bias stemming from the fact that cells are in different stages of the replication cycle (Skovgaard *et al*., 2011; Zomer *et al*., 2012). On top of this, one should expect to see batch-specific effects from different platforms, runs and laboratories (Khrameeva and Gelfand, 2012).

GC content has a major effect on coverage on all current sequencing platforms (Aird *et al*., 2011; Dohm *et al*., 2008). The bias is thought to be a PCR-related sequencing artifact, arising especially during the library preparation step. *CNOGpro* uses Charif and Lobry's R package SeqinR (Charif and Lobry, 2007) to calculate the local GC content in each window. Normalization is done in accordance with the method proposed by Yoon *et al*. (2009), who suggested calculating median RCs for windows with GC content $(0, 1, 2, \ldots, 100\%)$ and weighting the individual RC of each window i with the overall median RC (mRC) divided by the median RC corresponding to that window's GC content $\mathrm{mRC}_{\mathrm{GC}i}$.

$$\mathrm{RC}_i^{\mathrm{corr}} = \mathrm{RC}_i \times \frac{\mathrm{mRC}}{\mathrm{mRC}_{\mathrm{GCi}}} \tag{1}$$

2.4 CNV detection and quantification

Using graphical techniques, it is simple to detect areas of the chromosome where the median normalized RCs deviates from the juxtapositional ones. CNVs are identified as significantly increased or decreased RCs over multiple consecutive windows. Other methods recommend the application of a segmentation algorithm to determine the breakpoints at where the coverage changes at this stage. However, this approach discards important information such as the starts and stops of individual genes. Typically in re-sequencing experiments, this information is available from the reference genome. We therefore propose a method wherein the coverage of each individual gene (or intergenic segment) is modeled separately. This has a number of advantages: first, the problem of breakpoint determination disappears. (An assumption here is that duplications and deletions work at the gene level.) Second, genomic wave patterns will not be called as CNVs, because these signals will cancel each other out, and the mean coverage will be an unbiased estimator of the true copy number (assuming that there are a certain number of independent and identically distributed observations within each gene). Third, we are not as concerned with outliers and observations taking extreme values just by chance. The RC observation should behave as a stochastic variable following a distribution in the Poisson family, usually with a certain degree of overdispersion (Sepúlveda *et al*., 2013; Smith *et al*., 2008) due to uncorrected biases. In segmentation algorithms, extreme observations could influence the breakpoint estimation, but in our method such observations are attributed to random variation of the stochastic variable.

2.4.1 Bootstrapping approach to gene-wise copy number estimates

Having assigned each RC observation to a gene or intergenic region, we attempt to estimate the copy number of each individual segment. According to our assumptions, observations in the same segment represent samples from a single underlying count distribution belonging to the Poisson family. Regardless of overdispersion, the expected value of any such distribution is equal to a rate parameter λ, also known as the mean. The mean of the observed RCs is representative of the underlying distribution from which the observations are drawn. *CNOGpro* computes this parameter for each gene or intergenic region by simply taking the mean of that region's associated observations. Confidence intervals around the estimate are constructed by repeated sampling with replacement (bootstrapping) of the associated observations. The means directly reveal the underlying copy number if they are subsequently normalized to be expressed as a multiple of what we know is the 'true' mean for non-CNV regions. This leaves us with one problem: How do we determine which regions are non-CNV prior to any copy number inferences? We will suggest two possible approaches: (i) using a priori information on gene copy numbers or (ii) using the method implemented in *CNOGpro*'s hidden Markov model method.

2.4.2 Hidden Markov Model

CNOGpro also implements the Viterbi algorithm (Viterbi, 1967) for detecting the most probable sequence of copy number states in the input (normalized) RC data. In brief, an emission matrix is created wherein the log probabilities of each possible RC token in each possible copy number state are stored. The probabilities in each state are taken from negative binomial probability distributions (commonly used for overdispersed count data) whose parameters are indirectly inferred by sampling from the input RC data. First, RCs are sampled from the input data to establish the mean and variance. The most common results are averaged to represent the true mean and variance of the counts of segments with copy number equal to one, building upon the assumption that most genomic segments in a resequencing assembly to a reference organism will not be duplicated or deleted. The sampling distribution is parameterized as a negative binomial distribution with parameters p and r by solving the following equations, inherent to the negative binomial distribution, with respect to p and r:

$$\mathrm{mean} = \frac{\mathrm{pr}}{(1-\mathrm{p})} \tag{2}$$

$$\mathrm{var} = \frac{\mathrm{pr}}{(1-\mathrm{p})^2} \tag{3}$$

The probability distribution of each possible state over the various output tokens (represented as $k = \{0, 1, 2 \ldots\}$ in the following) are calculated with the following formula:

$$\Pr(X=k) = \binom{k+r-1}{k} p^k (1-p)^r \tag{4}$$

To prevent arithmetic overflow in normal computer systems, the binomial coefficient in (4) needs to be calculated using the alternative formulation using the gamma function:

$$\Pr(X=k) = \frac{\Gamma(k+r)}{\Gamma(k+1)\Gamma(r)} p^k (1-p)^r \tag{5}$$

which can be further modified by exponentiation of the logarithmic expression:

$$\Pr(X = k) = \mathrm{e}^{(\ln\Gamma(r+k) - \ln\Gamma(k+1) - \ln\Gamma(r) + x\ln(p) + r log(1-p))} \tag{6}$$

CNOGpro assumes that the mean and variance scales linearly with copy number. For the special case of a deletion, i.e. where the copy number is zero, the probabilities are taken from a geometric distribution with the parameter p representing the rate of erroneously mapped reads:

$$\Pr(X = k) = (1-p)^k p \tag{7}$$

The second parameter in the Hidden Markov model (HMM), the transition matrix, holds log-probabilities of switching between the different possible states in the chain. *CNOGpro* currently only accepts a single subparameter as input to the transition matrix, namely the probability q of switching states. The probability of remaining in the same state is calculated as $1-q$, whereas all remaining transitions are considered equally probable and share the probability q between them. A Viterbi path is then calculated as the most likely sequence of states in the chain, each step being only dependent on the one immediately before it as well as the emission and transition matrices.

The user is allowed control of the most important parameters such as the number of states to include, the probability of changing states and the fraction of erroneously mapped reads. Figure 2 presents a closer look at the relation between the HMM and bootstrap methods.

3 Results

3.1 Application to WGS data

To demonstrate its utilities, we tested *CNOGpro* on WGS data from *Staphylococcus aureus* TW20, the details of which can be found in the Supplementary File SD1. Results can be seen in Supplementary Table ST1 and isolate data in Supplementary Table ST2.

3.2 Considerations on sensitivity and specificity

To validate our protocol, we used the ART sequencing simulator of Huang *et al.* (2011), to create simulated datasets of the Illumina paired-end protocol. The sequence data were created with FN433596 as a template. We set the parameters of the simulator to best approximate the real sequence data we had used; average coverage was set to 100 and read length to 76 × 2 nt with an insert size of 500 and a standard deviation of 100. We discarded the first and final 500 nucleotides in the assembly, because to our knowledge, ART does not support circular chromosomes, leading to no representation of reads across the origin. To test sensitivity (ratio of the number of true positives to the number of positives), we introduced 30 CNVs by deleting or duplicating randomly chosen ORFs or intergenic segments in FN433596, as shown in Supplementary Table ST3. A random number generator chose the segments and corresponding CNV levels. To test the specificity (ratio of the number of true negatives to the number of negatives), we also used FN433596 directly (i.e. with no introduced CNVs). Non-mapping and

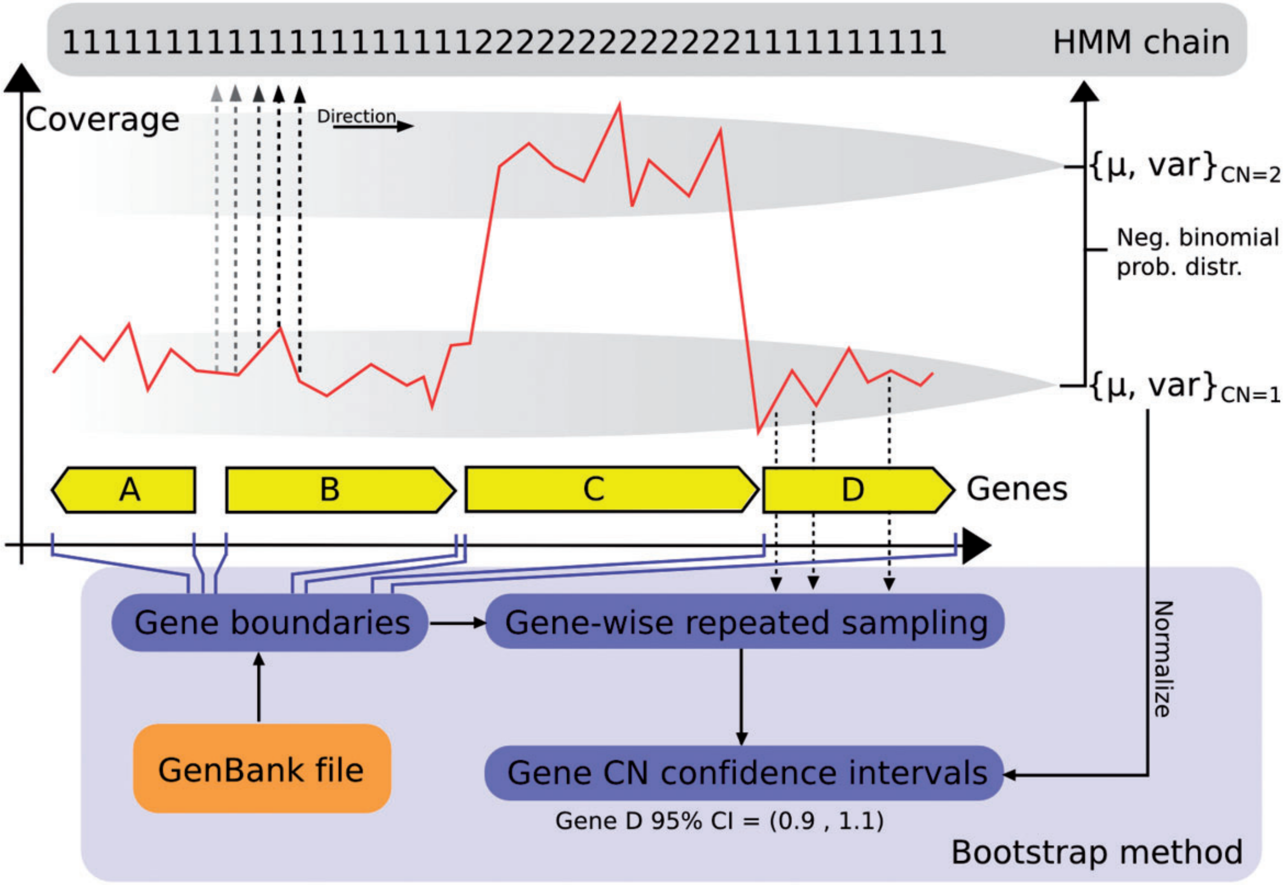


Fig. 2. The relation between *CNOGpro*'s principal methods of CNV quantification. Hidden Markov modeling and gene-wise bootstrapping, as well as their relation to the boundaries of genomic features needed in the analysis

low-quality reads (average PHRED score < 20) were filtered out before alignment, and reads were aligned with Maq (Version 0.7.1. Available at http://maq.sourceforge.net) against the reference sequence. Only reads that mapped in pairs were kept, and in cases where reads mapped to multiple possible sites, one was chosen at random.

Our algorithm called no CNVs for the non-altered sequence data, indicating a specificity of 100%. For the data with CNVs introduced, we encountered two false-positive calls: one in an IS256 region between coordinates 1 955 731 and 1 956 099, a repeat region with many occurrences in the chromosome, which would suggest that it was called because of a mapping problem and the other in a 230 bp intergenic region between coordinates 2 613 756 and 2 613 985. With a total of 5437 true-negative segments, this points to a specificity of 99.96%.

As expected, our algorithm performed poorly for CNV regions with a length of less than 100 bp, our window length for this analysis, calling only 2 out of 10 regions. All CNVs with a length >100 bp were detected, and of these, 11 out of 20 were called with the correct copy number using bootstrapping (true copy number represented in a 99% confidence interval), whereas the HMM called five CNVs correctly and unambiguously and 14 partly correct, meaning that the correct copy number were among the suggested solutions. (Because the HMM calculates breakpoints independently of gene starts and stops, gene-wise results from the HMM tend to reflect mixtures of copy numbers.) There was a strong correlation between CNV length and the accuracy of the copy number estimate. The results point to an overall sensitivity of 73% but significantly higher for longer CNV regions. Our quantification algorithms are less accurate, with 5/30 and 11/30 correctly called CNVs for the HMM and bootstrap approaches, respectively. Altogether, 19/30 regions were quantified with the correct or partly correct copy number. However, we note that simulated data do not suffer from identical biases as those introduced in real sequencing data.

We also benchmarked our algorithm against those of cnv-seq (Xie and Tammi, 2009) and cn.MOPS (Klambauer *et al*., 2012), two current standards in CNV analysis. Similar to our algorithm, neither cnv-seq nor cn.MOPS called any false positives. They did, however, perform considerably worse than our algorithm when it comes to sensitivity. Cnv-seq detected 14 out of the 30 CNV regions, which would indicate a sensitivity of 46% for this dataset. cn.MOPS correctly detected seven CNV regions and indicated the correct copy number for four of these, which would result in a sensitivity of 23% for this dataset. It must be noted here that cn.MOPS is designed to accept more than two samples as an input, and so the results might differ if we had included additional samples. Results from these analyses can be found in Supplementary Table ST3 and are visualized in Figure 3.

We additionally redid the above analyses in *CNOGpro* while letting the window length vary between 30 and 200 nt. At an average coverage of 100×, there was a trend of increased sensitivity with the lower window lengths (17/30 = 57% for window size 30 versus 12/30 = 40% for window size 200) without any apparent loss in specificity. Dropping the coverage to 20× did not impact sensitivity at all but led to a slightly higher rate of false-positive CNV calls (16/5437, Sp = 99.7%). At an average coverage of 10, sensitivity is 43% (13/30) and specificity 99.0% (5386/5437). In summary, it is possible to call CNVs even from runs with an average coverage as low as 10×, and in fact, there is only a moderate drop in sensitivity when moving from 100× to 10× coverage.

4 Discussion

4.1 CNV calling in prokaryotes versus in humans

The departures from CNV calling problems on human data are multifold. First, the levels of non-coding and repetitive DNA are much lower in prokaryotes. Consequently, genomic mappability (Lee and Schatz, 2012, Derrien *et al*., 2012) is of diminished importance as a source of bias when compared with eukaryote data. The bias does not seem prominent for the Illumina GA platform, and normalizing may in fact introduce *more* bias to the RCs than what was already there [Evident from the results of Magi *et al.* (2011).] We, therefore, suggest that correcting for this bias is unnecessary for most prokaryotic sequencing experiments, at least those sequenced on Illumina GA machines. Second, the sizes of prokaryotic genomes are a fraction of eukaryotic ones. This relation typically translates to CNV regions as well, with regions being up to several kilobases long. This is the lower threshold of the range of known human CNVs, which can be many megabases long (Redon *et al*., 2006). *CNOGpro* investigates coverage on a gene-by-gene basis, providing a higher sensitivity to detect short CNV regions. Third, although human-origin data are nearly always from diploid cell types, prokaryotes can have a varying number of copies of each chromosome, as well as plasmids, as well as partial copies and chimeras. This affects copy number quantification attempts by algorithms designed with a diploid setting in mind. Note, however, that the nature of cell replication and division in exponentially growing prokaryotes allows many copies of the chromosome to pile up in a single cell (Pecoraro *et al*., 2011). Origin-proximal regions are often the most amplified in this growing phase, with a downward sloping gradient towards the more distant parts of the genome (Chao *et al*., 2013; Gallagher *et al*., 2011; Skovgaard *et al*., 2011; Zomer *et al*., 2012). The result is that one will often find non-integer copy numbers of a gene, which actually represents a mixture of the base copy number and the consensus of the replication cycles of sequenced cells.

4.2 Strengths and weaknesses

The primary strength of *CNOGpro* lies in its ability to inform the user of copy number called through two fundamentally different approaches: through gene-wise bootstrapping of RCs and by cycling through RCs chromosome-wise in an HMM. The former allows high sensitivity, whereas the latter is faster and only returns integer results. Both methods have weaknesses that can be alleviated by the other. For example, consider the fact that genomic coverage usually follows wave-like patterns, the causes and significance of which are not known, although it is known that the pattern correlates with GC content, not only of the genomic region but also of the probes used in the sequencing (Diskin *et al*., 2008). This wave pattern confound methods that use segmentation algorithms to try to properly assign breakpoints, such as HMM-based approaches, but when each gene's coverage is investigated individually, this becomes less of a problem because the tips of the waves cancel each other out while the median, on average, remains more or less the same. We tested this in our simulated TW20 data multiplying observed coverage by a sine function with a 1000-bp period (roughly the size of an average gene), allowing the amplitude to randomly vary between periods, but with a maximum of 30% of the mean coverage. This changed the copy number calls for two segments out of the tested 5466, indicating that it had little overall impact in the analysis.

Using the bootstrapping method, it is also possible to predict CNV in genes that are present in multiple copies in the reference. If, for example, the reference has two copies of a gene and we find that our test organism presents with a bootstrap copy number result of

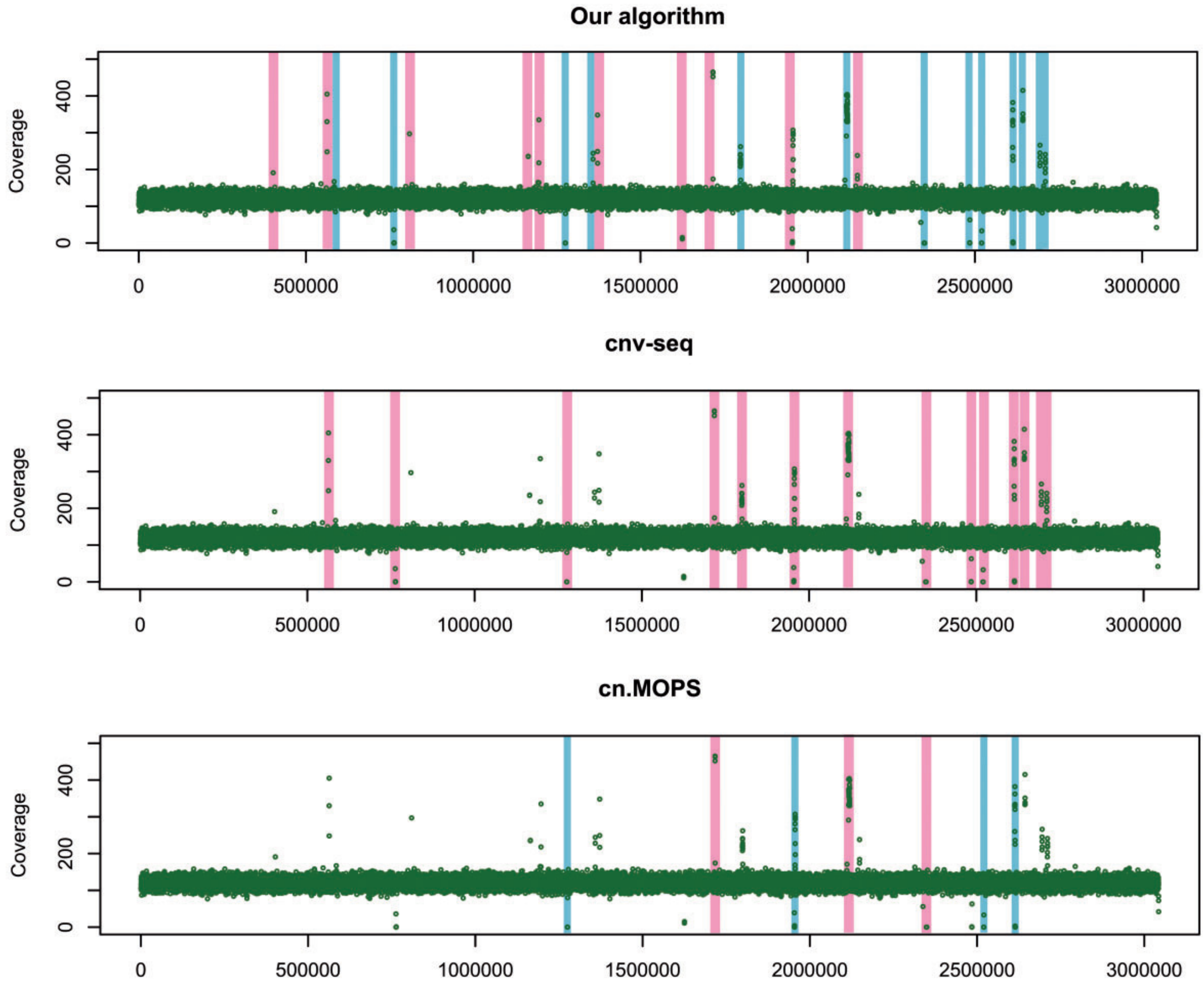


Fig. 3. Comparison of the present algorithm, cnv-seq and cn.MOPS in detecting and quantifying CNVs in simulated data with artificially introduced CNVs. Red color indicates that the algorithm correctly detected a CNV region, but that it was either not quantified (as in cnv-seq) or it was quantified incorrectly. Segments that were quantified correctly are highlighted in blue. The *x*-axes correspond to genomic position

1.5, we would expect the test organism to host three copies of the gene in question. Conversely, the expected bootstrap copy number result in a situation of 'de-duplication' where the reference has three copies and the test organism 2 would be 0.67. It is helpful to have knowledge of multi-copy occurrences of genes in the reference, because such information may help distinguishing genuine CNV from noise. In any case, the sensitivity towards CNV in multi-copy genes will be somewhat lower than what we have estimated in the previous section because such variation have lower signal intensities than variation in single-copy genes.

CNOGpro only compares coverage internally in a chromosome. This is helpful because significant biases have been demonstrated between identical isolates sequenced at different laboratories, on different machines and even with slightly differing protocols (Aird *et al.*, 2011; Khrameeva and Gelfand, 2012). [There are other potential causes of bias that we have not accounted for that may also play a part. For example, supercoiling of the genome has been shown to affect transcription under *in vivo* conditions (Pruss and Drlica, 1989).]

A few weaknesses with the gene-by-gene method of analyzing CNVs must also be noted: First, we will not discover intragenic duplications or deletions such as, for example, the duplication of some intragenic sequence motif. (However, the HMM model might still discover it if the duplication signal is strong. In this case, multiple possible copy numbers will be suggested for the gene in question.) Second, for very small genes, we may not have the required amount of RC observations to reject the null hypothesis of no variation in copy number, even if the gene is in fact present in more or less copies than in the reference sample. These weaknesses underscore the need for careful inspection of the data and the generous application of graphical methods for control.

4.3 Conclusion

We have presented a simple, quick and effective tool for detecting and quantifying CNVs in WGS data of prokaryotic organisms. When comparing similar prokaryotic genomes where details about the genomic layout in the reference are available, it represents a considerable jump in accuracy over other methods. It additionally has functions for creating high-quality informative plots and figures, an example of which can be seen in Figure 4. Our method starts with WGS data in the SAM format. Data are easily accessible through the Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/Traces/sra), from

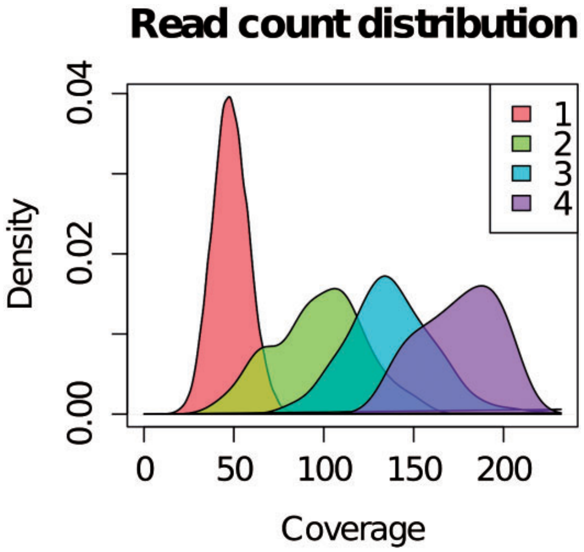


Fig. 4. Output of *CNOGpro*'s plot method, showing density curves of read count observations partitioned according to assigned copy number state

where one can also download the SRA toolkit and create SAM-format files from sequencing experiments. There are probably datasets from thousands of sequencing experiments freely available in the different sequence banks, just waiting for someone to analyze the clues that have been hidden in the frequencies with which each sequence occurs. *CNOGpro* is written entirely in the R programming language and is freely available under the GNU public license GPL-2. We believe it will be a valuable addition to the toolbox of every researcher conducting resequencing experiments to study copy number variance.

Acknowledgements

We thank the Computational Life Science initiative at the University of Oslo for helpful comments related to the analysis. We would also like to thank the reviewers of this article for suggestions to improvements and in particular the encouragement to publish this algorithm as an R package.

Funding

This work was funded by the Norwegian University of Life Sciences (NMBU).

Conflict of Interest: none declared.

References

Abyzov,A. *et al.* (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.*, **21**, 974–984.

Aird,D. *et al.* (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.*, **12**, R18.

Alkan,C. *et al.* (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat. Genet.*, **41**, 1061–1067.

Alkan,C. *et al.* (2011) Genome structural variation discovery and genotyping. *Nat. Rev. Genet.*, **12**, 363–376.

Campbell,P.J. *et al.* (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.*, **40**, 722–729.

Chao,M.C. *et al.* (2013) High-resolution definition of the Vibrio cholerae essential gene set with hidden Markov model-based analyses of transposon-insertion sequencing data. *Nucleic Acids Res*, 1–16.

Charif,D. and Lobry,J.R. (2007) SeqinR 1.0-2: a contributed package to the r project for statistical computing devoted to biological sequences retrieval and analysis. In: D.U.,Bastolla *et al* (eds.) *Structural Approaches to Sequence Evolution, Biological and Medical Physics, Biomedical Engineering*. Springer-Verlag, Berlin, Germany, pp. 207–232.

Cooper,G.M. *et al.* (2007) Mutational and selective effects on copy-number variants in the human genome. *Nat. Genet.*, **39**, S22–S29.

Derrien,T. *et al.* (2012) Fast computation and applications of genome mappability. *PLoS One*, **7**, e30377.

Diskin,S.J. *et al.* (2008) Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res.*, **36**, e126.

Dohm,J.C. *et al.* (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, **36**, e105.

Gallagher,L.A. *et al.* (2011) Genome-scale identification of resistance functions in *Pseudomonas aeruginosa* using Tn-seq. *mBio*, **2**, e00315–e00310.

Hastings,P.J. *et al.* (2009) Mechanisms of change in gene copy number. *Nat. Rev. Genet.*, **10**, 551–564.

Huang,W. *et al.* (2011) ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**, 593–594.

Khrameeva,E.E. and Gelfand,M.S. (2012) Biases in read coverage demonstrated by interlaboratory and interplatform comparison of 117 mRNA and genome sequencing experiments. *BMC Bioinformatics*, **13**, S4.

Klambauer,G. *et al.* (2012) cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.*, **40**, e69.

Klappenbach,J.A. *et al.* (2000) rRNA operon copy number reflects ecological strategies of bacteria. *Appl. Environ. Microbiol.*, **66**, 1328–1333.

Kondrashov,F.A. (2012) Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc. R. Soc. B Biol. Sci.*, **279**, 5048–5057.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Lee,H. and Schatz,M.C. (2012) Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. *Bioinformatics*, **28**, 2097–2105.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Magi,A. *et al.* (2011) Read count approach for DNA copy number variants detection. *Bioinformatics*, **28**, 470–478.

Medvedev,P. *et al.* (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods*, **6**, S13–S20.

Nijkamp,J.F. *et al.* (2012) De novo detection of copy number variation by co-assembly. *Bioinformatics*, **28**, 3195–3202.

Pecoraro,V. *et al.* (2011) Quantification of ploidy in proteobacteria revealed the existence of monoploid, (mero-)oligoploid and polyploid species. *PLoS One*, **6**, e16392.

Pruss,G.J. and Drlica,K. (1989) DNA supercoiling and prokaryotic transcription. *Cell*, **56**, 521–523.

Redon,R. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.

Riehle,M.M. *et al.* (2001) Genetic architecture of thermal adaptation in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **98**, 525–530.

Sepúlveda,N. *et al.* (2013) A Poisson hierarchical modelling approach to detecting copy number variation in sequence coverage data. *BMC Genomics*, **14**, 128.

Skovgaard,O. *et al.* (2011) Genome-wide detection of chromosomal rearrangements, indels, and mutations in circular chromosomes by short read sequencing. *Genome Res.*, **21**, 1388–1393.

Smith,D.R. *et al.* (2008) Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res.*, **18**, 1638–1642.

Stranger,B.E. *et al.* (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, **315**, 848–853.

Viterbi,A. (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory*, **13**, 260–269.

Xie,C. and Tammi,M.T. (2009) CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*, **10**, 80.

Yoon,S. *et al.* (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.*, **19**, 1586–1592.

Zhao,M. *et al.* (2013) Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*, **14**, S1.

Zhou,W. *et al.* (2014) Bias from removing read duplication in ultra-deep sequencing experiments. *Bioinformatics*, 1073–1080.

Zomer,A. *et al.* (2012) ESSENTIALS: software for rapid analysis of high throughput transposon insertion sequencing data. *PLoS One*, **7**, e43012.

Supplementary results:

WGS data from *Staphylococcus aureus* TW20

To test our algorithm we downloaded six Illumina runs on different isolates of *Staphylococcus aureus* subsp. *aureus* TW20, all paired-end 2x76bp (2x75 for ERR142616), from the DNA Data Bank of Japan (*http://www.ddbj.nig.ac.jp*), the details of which can be found in Table S1. Isolates that met eligibility criteria of appearing to originate from the same isolate, submitter (The Sanger Center), instrument (GA II) and sequencing protocol, were chosen by a convenience approach by taking the top 6 samples from a DDBJ search that matched these inclusion criteria. Alignments to the reference sequence of *S. aureus* subsp. TW20, available from GenBank (*http://www.ncbi.nlm.nih.gov/genbank*) under accession number FN433596, were extracted using the SRA toolkit (http://www.ncbi.nlm.nih.gov/Traces/sra). The completed sequence of FN433596 consists of a single-chromosome of length 3,043,210 bp, with 2,957 genes and a GC content of 32.7%. For statistical properties, we made sure that the median coverage of each isolate was at least 20. None of the runs displayed significant positional bias, but all counts were corrected for GC content. (See figure S1)

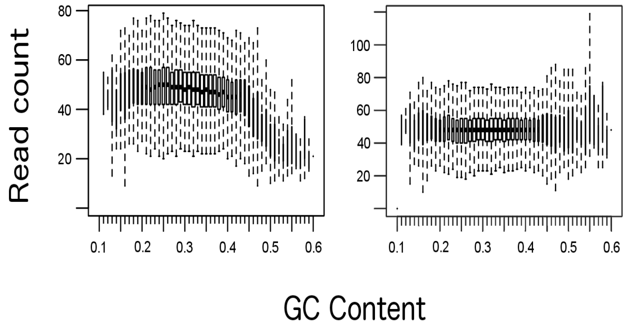


Figure S1: The read counts of the ERR043375 isolate plotted by local GC content. A convex shape of the counts is apparent before the correction. (Left). After adjustment (right), the counts are independent of the GC content.

As would be expected, the vast majority of the genomes had a coverage that was most consistent with a copy number of 1, i.e. the DNA sequence was present at the same number of copies as in the reference organism. All CNVs had a duplication number of 2, 3 or 4; there were no segments that had been more than fourfold duplicated, nor any full deletions that could be detected in any of the investigated isolates, although some isolates revealed a coverage profile that might indicate a reduction in some genes. For example, investigation of the coverage of the rRNA genes (5S, 16S and 23S) in ERR043375 reveals a mean coverage of 0.67, 0.83 and 0.74, respectively. Since the reference genome contains respectively 6, 5 and 5 copies of these genes, coverage in ERR043375 points to a likely deletion of 2 copies of the 5S gene, 1 copy of the 16S gene and 1 or 2 copies of the 23S gene. Alternatively, these findings may indicate paucity in growth and reproduction (Klappenbach *et al.*, 2000). There is also the possibility that the coverage is lower because of poor alignment, however, from careful inspection of our alignments we can find no reasons why this should be the case.

All 6 isolates displayed a relatively similar CNV pattern. For example, the ~8,500 bp system of consecutive Open Reading Frames (ORFs) between IS431-1 and IS431-2 (FN433596 coordinates 59,286 to 66,405) was amplified in all isolates. This genomic region is part of the staphylococcal cassette chromosome mercury (*SCCmercury*) and contains the *merRTAB* genes, which are involved in the handling and resistance against mercuric chloride, an obsolete disinfectant (Chongtrakool *et al.*, 2006). These genes are likely part of a single functional operon. It is curious to note that the ends of the duplicated segment contain transposase elements, suggesting a cut-and-paste type of duplication. The coverage of the *merR* gene suggested a copy number of 3 in all our isolates, while the other genes in this region had a coverage that was most likely for a copy number of 2 (ERR043367, ERR043371, ERR043375, ERR043379) or 3 (ERR142616, ERR316404).

Also amplified in all isolates was the *cadA* gene, involved in resistance towards cadmium (Nies, 1992). Results indicated that the copy number of this gene was 2 in all investigated isolates. Interestingly, the juxtapositional ORFs *cadC* and cadmium resistance transporter did not show any sign of duplication, which would suggest that only the *cadA* gene was duplicated, not the entire cadmium resistance complex. The third and final duplication that was seen in all isolates was located between coordinates 2,873,916 and 2,877,632, a segment containing the *tnpR* resolvase gene, transposase and SATW20_27070. The most likely copy number for this segment was 3, although in ERR043367 and ERR043371 the data indicated that *tnpR* resolvase might be present at a copy number of 4.

Table S1. Details of *Staphylococcus aureus* TW20 sequencing runs used in this study.

DDBJ acc. no.	Number of reads	%	Mean (range) coverage	CNVs
ERR043367	7,208,896	93.3	168 (16-2804)	37
ERR043371	2,571,438	95.3	61.2 (8-961)	37
ERR043375	1,511,462	97	35.7 (0-194)	25
ERR043379	3,974,320	96.4	93 (0-494)	27
ERR142616	46,283,212	96.5	732.2 (189-3919)	27
ERR316404	1,388,562	97	33.6 (0-185)	26

% = Percentage of reads mapped to reference. CNVs = Number of segments called as CNV regions (out of the total number of 5466 segments tested.)

The fact that the same regions were duplicated in all the 6 investigated isolates could mean one of several things: i) They are all more closely related to each other than to FN433596, *i.e.* all investigated isolates have an ancestor that acquired these duplications, and this ancestor may be a descendant of FN433596. It seems unlikely that each isolate separately acquired this specific CNV pattern. Alternatively, ii) the duplicated areas could be mutational hotspots. However, this seems less likely since isolates displayed similar (but not identical) copy number coefficients for these segments. iii) A final possibility (although least likely, in our opinion) is that FN433596 is incorrectly assembled, and that these duplications actually exist in the reference genome as well.

An exhaustive list of all segments with copy numbers is presented in Supplementary table 2.

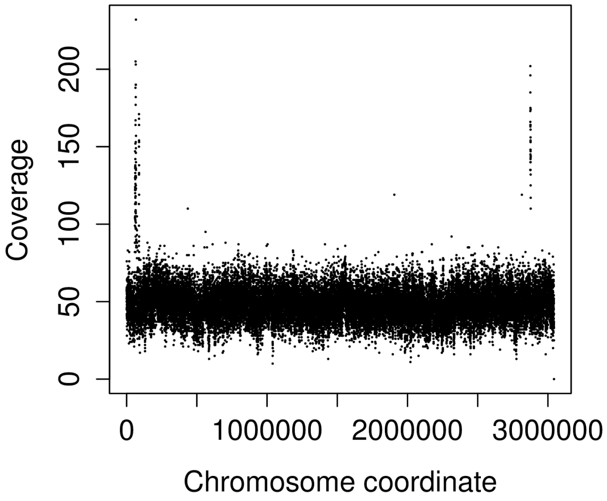


Figure S2: GC-corrected read count observations (y-axis) plotted by chromosome coordinate (x-axis) for the ERR043375 run.

REFERENCES

Chongtrakool,P. *et al.* (2006) Staphylococcal Cassette Chromosome mec (SCCmec) Typing of Methicillin-Resistant Staphylococcus aureus Strains Isolated in 11 Asian Countries: a Proposal for a New Nomenclature for SCCmec Elements. *Antimicrob. Agents Chemother.*, **50**, 1001–1012.

Klappenbach,J.A. *et al.* (2000) rRNA operon copy number reflects ecological strategies of bacteria. *Appl. Environ. Microbiol.*, **66**, 1328–1333.

Nies,D.H. (1992) Resistance to cadmium, cobalt, zinc, and nickel in microbes. *Plasmid*, **27**, 17–28.

Supplementary Table ST3: Comparison of CNOGproto cnv-seq and cn.MOPS on simulated data using varying window length and coverage paramters

Segment coordinates	Segment length	True CNV number	Detected	CNV number(s) HMM	Quantified correctly HMM	CNV point estimate, w=100	Rounded result, w=100	CNV interval Bootstrap (0.01 - 0.99)	Quantified correctly Bootstrap	Point estimate w=30	Rounded result, w=30	Point estimate w=200	Rounded result, w=200	Point estimate w=100, cov=20	Rounded result, w=100, cov=20	Point estimate w=100, cov=10	Rounded result, w=100, cov=10	Detected by cnv-seq	Detected by cn.MOPS	Quantified correctly by cn.MOPS
(330012,330015]	4 bp	0		1			1		No		1									
(344821,344843]	23 bp	2		1			1		No		1									
(388478,388485]	8 bp	0		1			1		No		1									
(402047,402109]	63 bp	4	X	1, 2	No		1	1.09 - 1.58	No	1.58	2									
(562944,563213]	270 bp	3	X	1, 2, 3	Partly		2	1.24 - 3.19	Yes	2.71	3	1.83	2	1.76	2	2.44	2	X		
(762780,763607]	828 bp	0	X	0, 1	Partly	0.02	0	0 - 0.36	Yes	0.02	0	0.29	0	0.17	0	0.15	0	X		
(809546,809626]	80 bp	4	X	1, 3	No		2	0.93 - 2.47	No	2.17	2	1.35	1	1.83	2	1.76	2			
(1164412,1164687]	276 bp	2	X	1, 2	Partly	1.7	2	1.19 - 1.97	No	1.7	2	1.53	2	1.77	2	1.98	2			
(1196369,1196578]	210 bp	3	X	1, 2, 3	Partly	1.98	2	1.35 - 2.79	No	2.27	2	1.74	2	2.26	2	2.34	2			
(1275253,1275864]	612 bp	0	X	0, 1	Partly	0.1	0	0 - 0.30	Yes	0.05	0	0.21	0	0.15	0	0.13	0	X	X	Yes
(1358393,1358649]	248 bp	2	X	1, 2	Partly	1.57	2	1.08 - 2.00	Yes	1.73	2	1.4	1	1.5	2	1.94	2			
(1371773,1371988]	216 bp	3	X	2, 3	Partly	2.25	2	1.81 - 2.89	No	2.53	3	1.89	2	2.3	2	2.48	2			
(1625170,1625303]	134 bp	0	X	1	No	0.41	0	0.09 - 1-02	No	0.13	0	0.56	1	0.58	1	0.47	0			
(1716617,1716970]	354 bp	4	X	2, 4	Partly	3.21	3	2.01 - 3.84	No	3.26	3	3.21	3	3.12	3	3.65	4	X	X	No
(1798287,1799588]	1302 bp	2	X	2	Yes	1.82	2	1.68 - 1.95	No	1.86	2	1.83	2	1.98	2	2.04	2	X		
(1953890,1954411]	522 bp	0	X	0, 1, 2	Partly	0.79	1	0.46 - 0.99	No	0.79	1	0.67	1	0.67	1	0.84	1	X	X	Yes
(2115756,2119531]	3776 bp	3	X	2, 3	Partly	2.91	3	2.73 - 3.02	Yes	2.94	3	2.87	3	2.96	3	2.98	3	X	X	No
(2148564,2148827]	264 bp	2	X	1, 2	Partly	1.49	1	1.1 - 1.87	No	1.65	2	1.29	1	1.47	1	1.28	1			
(2186039,2186055]	17 bp	0		1			1		No											
(2195082,2195085]	4 bp	2		1			1		No											
(2219052,2219068]	17 bp	0		1			1		No											
(2338314,2338339]	26 bp	0		1			1		No											
(2348148,2349824]	1677 bp	0	X	0	Yes	0.05	0	0 - 0.21	Yes	0.02	0	0.13	0	0.06	0	0.09	0	X	X	No
(2484353,2484868]	494 bp	0	X	0	Yes	0.09	0	0 - 0.01	Yes	0.03	0	0.33	0	0.12	0	0.17	0	X		
(2519733,2520386]	654 bp	0	X	0	Yes	0.05	0	0 - 0.18	Yes	0.03	0	0.15	0	0.1	0	0.06	0	X	X	Yes
(2612863,2613755]	893 bp	2	X	2, 3	Partly	2.3	2	1.67 - 2.76	Yes	2.4	2	2.3	2	2.47	2	2.67	2	X	X	Yes
(2643228,2643743]	515 bp	3	X	1, 3	Partly	2.6	3	1.89 - 3.14	Yes	2.79	3	2.61	3	2.76	3	2.44	3	X		
(2645385,2645421]	37 bp	3		1			1		No		1									
(2694225,2694725]	501 bp	2	X	1, 2	Partly	1.78	2	1.31 - 2.06	Yes	1.88	2	1.77	2	1.51	2	1.68	2	X		
(2710239,2710898]	660 bp	2	X	2	Yes	1.74	2	1.74 - 1.87	No	1.75	2	1.63	2	2.3	2	1.72	2	X		

The algorithm presented in this article was benchmarked against those of cnv-seq and cn.MOPS using simulated *Staphylococcus aureus* TW20 WGS data with 30 introduced CNV regions. Segments of a lower length than 100 bp are highlighted in yellow. Correctly quantified segments are highlighted in light green. A segment was regarded as correctly quantified if the point estimate rounded to nearest integer corresponded to the true copy number. For the bootstrap interval, a segment was regarded as quantified correctly if the true copy number was inside the 99% confidence interval. A partial match (highlighted in purple) means that the true copy number were among those suggested by the HMM, although because of inaccurate breakpoint determination multiple possible copy numbers were suggested for these segments.

Supplementary table 1 is a large file and has not been included in this thesis. It is available online (10.1093/bioinformatics/btv070)

III

Identifying copy number variation of the dominant virulence factor *msa* within genomes of the fish pathogen *Renibacterium salmoninarum*

O. Brynildsrud[a]#, S. Gulla[b], E. J. Feil[c], S. F. Nørstebø[a], L. D. Rhodes[d].

Department of Food Safety and Infection Biology, Norwegian University of Life Sciences (NMBU), Oslo, Norway[a]; Department of Bacteriology - Aquatic and Terrestrial Animals, Norwegian Veterinary Institute (NVI), Oslo, Norway[b]; Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath, United Kingdom[c]; Northwest Fisheries Science Center, National Marine Fisheries Service, Seattle, Washington USA[d].

Running head: CNV in the virulence factors of *R. salmoninarum*

#Address correspondence to Ola Brynildsrud, ola.brynildsrud@nmbu.no

Renibacterium salmoninarum is the causative agent of bacterial kidney disease (BKD), an important disease of farmed and wild salmonid fish worldwide. In spite of the wide spatiotemporal distribution for this disease and habitat pressures ranging from natural environment to aquaculture and rivers to marine environments, little variation has been observed in the *R. salmoninarum* genome. Here we use the coverage depth from genomic sequencing and real-time quantitative PCR to detect copy number variation (CNV) among the genes of *R. salmoninarum*. CNV was limited to the known dominant virulence factors *msa* and p22. Among 68 isolates representing the United Kingdom, Norway, and North America, the *msa* gene ranged from two to five identical copies and the *p22* gene ranged from one to five copies. CNV for these two genes co-occurred, suggesting they may be functionally linked. Isolates carrying CNV were phylogenetically restricted, and originated predominantly from sites in North America, rather than the United Kingdom or Norway. Although both phylogenetic relationship and geographic origin were found to correlate with CNV status, geographic origin was a much stronger predictor than phylogeny, suggesting a role for local selection pressures in the repeated emergence and maintenance of this trait.

Introduction

Renibacterium salmoninarum is the causative agent of Bacterial Kidney Disease (BKD) in cultured and wild salmonid fish. BKD can result in acute morbidity or mortality, or it can be a slowly progressive disease causing an often-dramatic decline in growth. BKD is economically important in aquaculture, where it can spread horizontally throughout sea pens of juvenile and subadult Atlantic salmon (*Salmo salar*) (1) or vertically through transferred broodstock or eggs (2). It is also a concern for conservation and restoration efforts for endangered fish stocks because infections are prevalent among more susceptible free-ranging Pacific salmon in river and marine systems (3–5).

Although the pathogenicity of *R. salmoninarum* is incompletely understood, several antigenic determinants have been described, including capsular synthesis, heme acquisition operons, hemolysins and an immunosuppressive 22 kDa surface protein provisionally named *p22* (6). However, the dominant immunogenic protein produced by this organism is an abundant heat-stable 57-kDa extracellular protein known as Major Soluble Antigen (MSA) (7, 8). The MSA protein makes up 60 - 70% of all surface protein in *R. salmoninarum (6, 9)*, and it is involved in immunosuppression(6, 7, 10), agglutination (8, 11, 12) and virulence (11, 13, 14).

The genome of the type strain of *R. salmoninarum*, ATCC33209, contains two identical transcriptionally active copies of the MSA-encoding gene; *msa1* and *msa2 (14, 15)*. Both genes are essential for the development of clinical disease and mortality (13). Whilst it seems certain that a single copy was originally acquired through horizontal gene transfer and subsequently duplicated within the bacterial genome (16), the origin of this gene is unclear, as no homolog to the *msa* gene has ever been found in any other sequenced genome. Both *msa* loci are flanked by insertion sequences and transposases, and *msa2* is additionally flanked by several degraded genes related to conjugation (including *traA* relaxase, type IV secretion protein and site-specific recombinase resolvase). Because multiple copies of identical genes are unusual in bacterial genomes, O'Farrell and Strom suggested that multiple *msa* copies might confer a selective advantage (14). Subsequently, Rhodes *et al.* demonstrated the presence of a third copy in some isolates, and provided evidence for a positive correlation between *msa* copy number and mortality at lower infection doses (17). It is not presently known whether the third *msa* locus (*msa3*) resides within an amplified chromosomal region or within extrachromosomal DNA. Plasmids have not been detected in *R. salmoninarum* (18), and a report of a bacteriophage by electron microscopy (19) has not yet been confirmed.

The horizontal acquisition of genes is a very rare event in this species; whole-genome sequencing data from phylogenetically diverse isolates of *R. salmoninarum* sampled over the last 50 years failed to find a single gene that was not represented in the genome of the type strain (20). However, the findings of Rhodes *et al.* (17), suggest that copy number variation (CNV) in the *msa* genes of *R. salmoninarum* has high phenotypic relevance. Gene duplication has been shown to be adaptive in bacteria (21), and CNV is known to be an important mechanism for dose variation of specific proteins under appropriate environmental conditions (22). One relevant example is that of *opa* virulence genes in *Neisseria meningitidis*. This bacterium alters its surface antigen structure by expressing none, one, or more Opa proteins, and each isolate can harbor up to ten similar *opa* copies that have arisen through horizontal transfer (23, 24). A recent study demonstrated that some strains of *Mycobacterium tuberculosis* harbored a large, tandem gene duplication and noted greater expression of an anaerobic survival regulon that is contained within the duplication (25).

The aim of this study was to screen a diverse collection of *Renibacterium salmoninarum* isolates for evidence of copy number variation in any of the core genes and, if found, to investigate phylogenetic and spatial patterns of the distribution of genetic variants. This work can provide a better understanding of *Renibacterium* evolution and may shed light on mechanisms for differential disease manifestation in different populations.

Materials and methods

Computational analyses

Sixty-eight isolates whose spatial and temporal origins varied widely were sequenced on the Illumina GAII platform at TGAC, Norwich, as part of a previous effort by the authors, and are available at the Sequence Read Archive of the National Center for Biotechnology Information (NCBI) under accession numbers listed in Table 1. Non-pairing reads, reads containing ambiguous characters, and reads with an average PHRED score of < 20 were discarded before alignment to reference genome ATCC33209 (available from NCBI GenBank under accession number NC010168) with Geneious v7.1 (Biomatters, Auckland, New Zealand), using the option to randomly map reads with multiple best hits.

CNVs were discovered using the R (R development core team, 2012) package *CNOGpro* (26) with the following parameters: Coverage counted in sliding windows of length 50 bp, prior probability of changing states (for each read count observation) was set to $p=1.0 \cdot 10^{-10}$, and the error-rate parameter was set to 0.01. The *runHMM* method was used to call CNV regions and copy numbers were considered correct if they agreed with credible intervals (percentiles 1 through 99) from the *runBootstrap* method. When evaluating results we discarded IS*994* tallies (unless they were co-occurring with CNVs in neighboring genes), as 69 copies (69 *orfA* and 67 *orfB*) of this element are known to exist in the reference genome (16), making it impossible to evaluate copy number variation with our method. We also considered standalone CNV calls in segments shorter than 300 bp unreliable, as such calls would not be completely unexpected from chance alone. When quantifying total *msa* enrichment, the signal from *msa1* and *msa2* were added together. The number of *p12* duplications was considered equal to type I *msa* duplications, meaning the frequencies of type II duplication could be found by subtracting type I from the total (Supplementary figure S1).

Putative alien sequence prediction

Pursuing the idea that the CNV regions discovered in this paper constituted genomic islands, we used the Sigi-HMM tool from Colombo v 3.8 (27) for prediction of genomic islands in prokaryotic genomes with default options activated to screen the ATCC33209 genome for genomic islands. Sigi-HMM uses Hidden Markov Models to detect aberrant synonymous codon usage gene-wise, and determines the most likely donor of putative alien genes from prokaryotic cluster groups.

Corroboration of results with real-time quantitative PCR

Through the real-time quantitative PCR (qPCR) cycle threshold (Ct) ratio we sought to supplement the results derived from *CNOGpro* analysis. The rationale behind the Ct detection ratio approach is as follows: When comparing strains varying in gene copy number, the ratio of the Ct-values can quantify the relative multiplicity of a CNV gene. This would however require simultaneous detection of a ubiquitous non-CNV reference gene to minimize the influence of within-strain variation. Furthermore all primers would have to be equally efficient (and target conserved gene regions) in all target strains. The method is explained in further detail at the end of this section.

The *Renibacterium salmoninarum* strains selected for assessment by this method were 5223 (an isolate whose coverage data indicated copy number variation) and ATCC33209 (cryopreserved at the Norwegian Veterinary Institute). They were taken from frozen stocks kept at -80 °C, plated onto SKDM2 (28) and cultivated at 15 °C for 20 days. Genomic DNA was extracted using a Gentra Puregene Yeast/Bact. Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. Concentration and purity of the DNA extracts were analyzed using a NanoDrop ND-1000 spectrophotometer (Thermo Scientific, Waltham, MA, USA).

Primer pairs targeting three genes, representing each major duplication cluster, were designed: Both *msa* genes (the ORFs of *msa1* and *msa2* are identical), the RSal33209_3334 gene which encodes the p22 protein, and the *lepA* gene, a ubiquitous and highly conserved housekeeping gene whose copy number is not thought to vary (29). The primers (Supplementary Table ST1) were designed in Geneious v7.1 (Biomatters, Auckland, New Zealand) using sequences from strain ATCC33209 as templates. A chief priority in the primer design was similar melting temperatures of all pairs.

Each qPCR reaction volume consisted of 25 µl Power SYBR Green PCR Master Mix (Applied Biosystems, Waltham, MA, USA.), 0.3 µM of both forward- and reverse primers (Invitrogen), 10 µl DNA template and a final addition of *Milli-Q* water to reach a total reaction volume of 50 µl. Subsequent qPCR was conducted on a Stratagene Mx3005P thermal cycler (Stratagene, La Jolla, CA, USA.), with thermal cycles involving i) 95°C for 10 min, ii) 45 cycles of 95°C for 15 sec; 60°C for 20 sec; 72°C for 15 sec, and iii) 95°C for 30 sec; 55°C for 30 sec; 95°C (gradual heating) for 30 sec. Fluorescence data was collected through the SYBR channel towards the end of the annealing steps (for amplification plots) and successively through the last, gradual heating step (for dissociation curves). Data analysis was performed using MxPro v4.10 software (Stratagene, La Jolla, CA, USA.).

When comparing Ct-values between strains and primer pairs, it is essential that the primer efficiencies are relatively similar both across genes and strains. This was verified by creating standard curves for each primer-and-strain pair, using four DNA-concentrations (10x dilutions). In order to minimize the influence of factors unrelated to primers and/or templates, only Ct-values from the same qPCR-setup/-run were compared in this way. We visually inspected the degree of parallelism between the standard curves, and additionally verified that the sample coefficient of variation was low (<5 %). Under an equal efficiencies assumption, it is possible to compare Ct-values of the constant term of the standard curve directly. We enforced parallelism of the standard curves by first finding the regression line that minimized sum-of-squared residuals for the pool of all concentration-Ct-value data points, and then proceeding with linear regression for individual isolate-gene-pairs while constraining the slope term to be equal to that found for the pooled data. This procedure finds the best-fitting parallel lines and allows direct comparison of Ct-values. Utilizing a reference gene *ref* whose copy number is (nearly) always one (such as *lepA*) to normalize within-strain variation, the copy number ratio of a gene of interest *X* between a *test* strain and a *control* strain (here; 5223 and ATCC33209, respectively), wherein the copy numbers are known *a priori*, can be found by using the following formula:

$$\frac{X_{test}}{X_{control}} = 2^{\Delta\Delta Ct} = 2^{\left(Ct_X - Ct_{ref}\right)_{control} - \left(Ct_X - Ct_{ref}\right)_{test}}$$

Here Ct_X and Ct_{ref} refer to the cycle threshold of the gene of interest and reference gene, and *test* and *control* refers to the strain of interest and control strain, respectively.

This qPCR 'comparative Ct-value' method will presumably be subject to random error and likely also unknown sources of bias. It should thus not be considered as entirely accurate with regards to detection of copy number variation, but rather used to corroborate results acquired e.g. through *CNOGpro* analysis.

Regression analysis

In the following analyses, the presence or absence of *msa* copy number variation in an isolate was considered a binary trait. Associations between this trait and year of isolation, host species, and saltwater/freshwater habitat were investigated by univariable logistic regression using R.

Cluster analysis through matrix correlation:

Phylogenetic trees were created from SNP alignments with the program MrBayes (30) as specified in Brynildsrud *et al.* (20).Pairwise patristic distances between isolates were calculated as the sum of branch lengths between leaf pairs of the consensus tree. Pairwise geodesic distance between isolate geographical origins were calculated by solving for central angle in the spherical law of cosines and multiplying by the radius of Earth. The latitude-longitude coordinates were rounded to the closest degree. In some cases the exact sample origin was not known, in which case the coordinate pair was set to represent geographical midpoints for the sub-national region. In order to test for phylogenetic and spatial clustering of CNV presence/absence, we created a binary matrix where equal CNV statuses of isolate pairs were coded as 1 and unequal as 0. In this analysis we regarded isolates listed in Table 2 (excluding 5006) as positive for the duplication in question and the remaining isolates as negative. We then adopted a Mantel test-like approach by performing the Mann-Whitney U test of equal distributions between groups defined by CNV-status on patristic/geodesic distance data. This test estimator was subsequently compared to those obtained from 10.000 random permutations of the CNV status matrix. The trait was considered to be phylogenetically or spatially clustered if the test estimator fell below the lower 1-percentile limit in the distribution of permuted data set estimators.

Results

Overall, very little copy number variation was seen in our isolates. In fact, the coverage data of most isolates (57/68) indicated no variation at all. This finding is consistent with previous reports of a high degree of sequence conservation in the *R. salmoninarum* genome. Furthermore, it confirms the supposition that the minimum copy number of *msa* genes is two, as no isolate presented with read coverage that was suggestive of only a single copy. Nevertheless, copy number variation was found in eleven isolates, shown in Table 2.

In total, there were nine distinct CNV regions. Four of these were unique to the Carson5b isolate and two to isolate 5006. The Carson5b isolate contained two copies of the following genes: LacI family transcriptional regulator (Rsal33209_0109), NADH-dependent flavin oxidoreductase (Rsal33209_1458), ferredoxin NADH-reductase (Rsal33209_2607) and the hypothetical protein-encoding Rsal33209_3193. The 5006 isolate contained large 1->2 duplications of the segments that in the reference genome ATCC33209 can be found between coordinates 2,974,628 to 3,084,569 and 3,088,016 to 3,100,482.

The remaining three CNVs were non-unique and co-occurring in all eleven CNV isolates except isolate 5006. They are described in the following section and illustrated schematically in Figure 1. To more clearly categorize *msa* duplication types, we provisionally introduce the nomenclature "type I" and "type II."

Type I *msa* duplication included the *msa* gene, the 12 kDa hypothetical protein Rsal33209_0132, the transposase-encoding Rsal33209_0133 and the inactivated IS sequence IS*Rs3*, including all intergenic segments and flanking inverted IS*994* sequences. Type I *msa* duplication very closely resembles the genomic region roughly between coordinates 110.000-115.000 in ATCC33209.

Type II *msa* duplication included the *msa* gene with the intergenic sequence from the terminus of the gene and roughly 800 bp downstream.

The total number of *msa* copies in CNV-positive isolates ranged from three to five. Note that although the *msa1* and *msa2* genes differ very slightly at upstream and downstream sites, the ORFs themselves are identical, and we could therefore not conclusively determine whether the extra copies were duplications of *msa1*, *msa2* or a mixture. Complicating the *msa* copy number quantification was our discovery that the *msa* gene and downstream sequence contains several large (130-180 bp) inverted and direct repeats and one 91 bp perfect palindrome, confounding read mapping (Figure 1). Using terminator prediction tool ARNold (31), we discovered that the palindrome at the 3' of the *msa* ORF contained a predicted rho-independent terminator at both the *msa* loci, although with "G" as the central loop nucleotide for *msa1* and "C" for *msa2*.

The third non-unique CNV region represented the region between coordinates 2,965,759 and 2,967,751 in ATCC33209. This region contains a single ORF; a 22 kDa hypothetical protein (hereafter referred to as p22) labeled RSal33209_3334. Also part of the duplication unit was the intergenic segments on both sides of this ORF as well as the flanking inverted IS*994* sequences. The total number of *p22* copies in CNV-positive isolates ranged from one (in 5006) to five (in Carson5b).

A complete list of all CNVs discovered among the 68 isolates in this study can be found in Supplementary dataset S1.

Trait clustering

The *msa*-*p22*-duplication trait was found to be highly clustered within phylogenetically and spatially defined groups (Figure 2). For phylogenetic data, Mann-Whitney's U was computed as 419,090, which is lower than the full range of all permuted-matrix values (424,489 - 525,215) ($p < 1.0 \cdot 10^{-4}$ (calculated as in Diniz-Filho *et al.* (32)). However, since the distribution of U values follow a near-perfect normal distribution (as calculated by the Anderson-Darling test of normality), a parametric p-value estimation of $p = 7.4 \cdot 10^{-5}$ can be used (Supplementary figure S2). This translates to a strong, negative correlation between identical trait status and patristic distance. In other words, when two isolates are more closely related, they are both more likely to either possess or not possess the duplication trait. However, geodesic distance between isolate origins was even more highly correlated to trait identity. Here, the Mann-Whitney U estimator took the value of 255,344, compared to the full range 432,002 - 515531 from the permuted dataset ($p < 1 \cdot 10^{-4}$ by conservative (32) estimation; $p < 1.0 \cdot 10^{-50}$ by Gaussian parameterization). There were no statistically significant associations between CNV and isolation year, host species or freshwater/saltwater habitat (Tested by logistic regression, both bivariate and multivariate with interaction terms).

Corroboration by real-time qPCR

The *in silico* CNV analysis results were experimentally corroborated using real-time qPCR. This approach examines the copy number ratio of genes between strains by comparison of Ct-values: Two CNV-genes of interest (*msa* and *p22*) and one non-CNV reference gene (*lepA*) were assessed simultaneously for two strains (ATCC33209 and 5223). Given overall equivalent primer efficiencies (as validated: Supplementary Figure S3), we could infer the relative multiplicity of the CNV-genes between the two strains (see Materials & Methods for details). These ratios were then compared to those resulting from *CNOGpro* analysis. In the 5223 isolate, our qPCR results indicated a per-*msa* signal strength of 1.82x baseline, or a total of 3.64 copies. The equivalent most parsimonious result of the *in silico* analysis was 4 copies (obviously the true numbers are integers). For the *p22* gene, qPCR results indicated a copy number of 3.03 in 5223, while *in silico* results were most compatible with a copy number of 4. The coverage of the two ~700 bp intergenic segments surrounding the *p22* gene and flanked by IS*994* sequences was most consistent with a copy number of 3, suggesting that this may in fact be the true copy number, and that the HMM overestimated by one copy.

Putative alien prediction

Twenty-two genes clustered in nine genomic segments were annotated as putative aliens, shown in Supplementary Table S2. This list included both *msa* genes and their neighboring genes, but not the gene encoding the p22 protein. It was notable that five of the genes had a codon usage bias (CUB) reminiscent of *Oceanobacillus iheyensis*, which is found in deep-sea sediments. Four other CUB profiles suggested a *Streptomyces*/*Micromonospora* origin, typically found in soil and water, one suggested *Vibrio mimicus*, and one *Aeromonas salmonicida*. Taken together these suggested donors implicate other marine environment microbes as being important in the acquisition of extraneous genetic material. The origin of the *msa* gene remains a mystery however as its CUB profile did not match with any putative donor.

Discussion

With the exception of genes encoding ribosomal and transfer RNA subunits, it is unusual for multiple identical genes to occur in a bacterial genome (33). In this paper we have found that some *R. salmoninarum* isolates contain up to five identical copies of genes known to be important virulence factors. Given the high cost of maintaining redundant gene copies, it is tempting to conjecture that the observed copy number mutations must be adaptive or opportunistic, at least over shorter time frames. The immediate benefit of duplications could be through modulation of protein dosage under variable environmental conditions, while the long-term advantage is that the extra copies can, over time, accumulate mutations and evolve new functions (34–36). In favor of a selectionist explanation is the observation that these duplications do not seem to be simply random mutations that are quickly removed from the population, but rather a trait that is shared and in some cases maintained by closely related isolates. Isolates 05372K and Cow-chs-94, for example, are very closely related despite being from separate river systems and isolated 11 years apart. They both have multiple duplications of both the *msa*, *p12* and *p22* genes, although there are variations in the exact numbers of each gene.

In a previous publication, we inferred *R. salmoninarum* phylogeny as divided into two major lineages; lineage 1, containing a mixture of European and North American isolates, and lineage 2, which exclusively contained European isolates. Lineage 1 could further be subdivided into 1A, with isolates from both Europe and North America, and 1B, with exclusively North American isolates (20). In the following, we will discuss the present CNV findings in the context of this phylogeny.

No CNV was found in any of the lineage 2 isolates. Across lineage 1, CNV was found in three major genes: *msa*, *p12* and *p22*. It was notable that copy number variation occurred in all three genes, or none - No isolates were found to have CNV in one gene but not the others, suggesting that these genes have a functional interaction relationship where increased copies of either are not valuable without concomitant copy number increases of the other, or that the genes are somehow duplicated together due to linkage. The latter possibility is perhaps somewhat marginalized by the known genomic distance between *msa* and *p22*, which in ATCC33209 is around 300.000 bp between *msa1* and *p22*, going through the origin. However, it is possible that these genes are more closely located in strains other than ATCC33209.

Copy number variation in *msa* has been documented previously (17), and the current study extends CNV to *p12* and *p22*. The *p22* gene encodes a poorly described loosely associated surface protein (37) that has been implicated in suppression of antibody production and a stronger agglutination of leucocytes than that which is seen for the MSA protein (6). Little is known about the hypothetical *p12* gene. We do not know its function or even whether it is interesting in itself, but at least it might be used to predict the relative fraction of type I and type II *msa* duplications (Supplementary Figure S1).

The presence of additional *msa*, *p12* and *p22* copies was seen in several phylogenetic clusters from both lineage 1A and 1B, but without any clear phylogenetic distribution pattern. Within these clusters, the patristic distances are short, but distances between them are larger. Ten of 68 isolates (~15%) displayed an increased copy number *msa*, *p12* and *p22* genotype, which is in stark contrast to the 19 of 26 isolates (~73%) that Rhodes *et al.* (17) found to be *msa3*-positive. In that paper, every isolate except two (MT239 and GL64, which are *msa3*-negative) were from the Pacific Northwest geographical region of the U.S., suggesting that the strains circulating in that particular region have a higher frequency of multiple-copy *msa* genotypes. This conjecture is supported by the present study: Among the 10 isolates containing additional-copy *msa* genes, 4 are from the USA, 5 from Canada, and 1 from Norway, corresponding to 44%, 45% and 8% of the total investigated isolates from each respective country. Notably, out of the 36 UK isolates, not a single one displayed CNV. The North American CNV isolates were sampled from both the West coast (British Columbia, Washington, Oregon) and

East coast (New Brunswick) of North America, as well as the interior US (Montana). In the present study we found that isolates with closer geographic origins were more likely to have identical *msa-p22* duplication trait status. Interestingly, the geographic origin was a much stronger predictor of CNV status than the phylogeny itself. In other words, the trait is present in phylogenetic tree leaf nodes that are of close geographic origin but less close in terms of phylogenetic distance. Isolates 05372K and Cow-Chs-94 for example are of lineage 1B origin, and thus thought to have diverged from lineage 1A isolates such as Carson5b between 100-700 years ago (20); In spite of this these isolates all have CNV in the *msa*, *p12* and *p22* genes. Note however that they are all isolated from a relatively confined geographical area of the U.S. state of Washington. The fact that we observe this pattern of low intra-cluster but high inter-cluster patristic distances and that isolates originate from multiple geographic locations across North America (and, in a single case, Norway), sampled over a 19 year period from 5 different species of salmon from both freshwater and saltwater habitats, strongly suggests multiple independent introductions of the trait rather than simple inheritance through binary fission. Furthermore, there seems to be a strong geographic component to the emergence of CNV. Ignoring the possibility that our isolate selection suffers from a high degree of bias, this raises multiple interesting questions: Is the duplication an adaptation to local selection pressure from the environment or are we seeing the effects of an unknown dispersal mode for this particular genotype? One possibility is that these duplications spontaneously and independently arise from within the genome and are selectively maintained due to fitness increase. Alternatively, the putative alien origin of these regions might suggest that the duplications are in fact local expansions of horizontally transferred pathogenicity islands (PAIs), and the relatively high frequency of this PAI in

North American waters as compared to European may indicate locally prevalent gene transfer vectors in that area.

A predominantly North American CNV distribution is also consistent with the findings of Wiens and Dale, who observed the *msa3* gene in North American but not in European isolates (38). The observed higher frequency of the *msa*- and *p22*-duplicate genotypes circulating in North America could thus be an important factor contributing to the usually much higher mortalities seen in BKD outbreaks in North America than in Europe (39), although this may also be attributed to a higher degree of innate resistance against the infection in Atlantic salmon and particularly rainbow trout. Wiens and Dale suggest a plasmid context of *msa3*, based on variable hybridization intensity in Southern blots, but an equally possible scenario could be the association of *msa3* with a prophage. It would be very interesting to see whether the phage reported by Fryer and Lannan in 1993 (19) included *msa* and *p22* genes. It is intriguing that both *msa1* and *msa2* are flanked by inverted IS sequences, notably IS*994*, and IS*3*-like insertion sequences as well as other ORFs with high homology to transposable elements and transposases, suggesting that they could be transferred and integrated by homologous recombination of these insertion elements.

We have found that there are (at least) two major types of *msa* duplications, which we categorize as type I and type II. Type I duplications are very similar to *msa1* with neighboring genomic regions in ATCC33209, including the *p12* gene, transposase, IS*Rs3* and flanking IS*994* elements. Type II duplications are limited to the *msa* gene and downstream non-coding sequence.

We were unable to determine if duplications were primarily of *msa1* or *msa2* origin. A key difference between *msa1* and *msa2* is the nucleotide 37 bp downstream of the ORF. In the sense direction, this is "G" for *msa1* and "C" for *msa2*. The polymorphism is located in the loop region of a predicted terminator (Colored green in Figure 1), which could potentially indicate differential regulation of the *msa* genes. Unfortunately, the location of the polymorphism in the center of a 91-bp perfect palindrome means we cannot use it to reliably establish whether duplications originate from *msa1* or *msa2*, since it would be impossible to distinguish sense-direction *msa1* reads from antisense *msa2* and vice versa. The fact that duplications include flanking IS sequences and the *msa1*-associated *p12* gene would suggest that *msa1* is the more likely candidate, in agreement with Rhodes *et al.* (17).

Although we have detected several large gene duplications, we have not been able to predict their relative orientation and distance to each other or to the rest of the chromosome. More research is needed to precisely determine the genomic context of these duplications, such as their genomic locations and whether or not they are associated with plasmids. Furthermore, it is not at all clear to what extent the duplications we have found in the present work impact on overall fitness. O'Farrell and Strom proposed a selective advantage of two *msa* gene copies, and it is tempting to speculate that the additional copies that we have found are increasingly beneficial to the bacterium. In support of this, Rhodes *et al.* (17) noted that the presence of a third *msa* copy was associated with increased mortality in a challenge experiment when the inoculum dose was beneath a certain threshold. Nevertheless, more research is needed to conclusively determine the relative fitness- and virulence relationships between different duplication-value *R. salmoninarum* isolates.

Acknowledgements:

OBB and SFN are funded through NMBU. SG is funded through the Norwegian Research Council and VaxxiNova Norway. EJF is funded by BBSRC/NERC grant WGS-AQUA (BB/M026388/1). LDR is funded through NOAA Fisheries.

References:

1. **Murray AG, Munro LA, Wallace IS, Allan CET, Peeler EJ, Thrush MA**. 2012. Epidemiology of *Renibacterium salmoninarum* in Scotland and the potential for compartmentalized management of salmon and trout farming areas. Aquaculture. **324–325**:1–13.

2. **Evelyn TPT, Prosperi-Porta L, Ketcheson JE**. 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 Aquat Organ. **1**:197–202.

3. **Sandell TA, Teel DJ, Fisher J, Beckman B, Jacobson KC**. 2015. Infections by *Renibacterium salmoninarum* and *Nanophyetus salmincola* Chapin are associated with reduced growth of juvenile Chinook salmon, *Oncorhynchus tshawytscha* (Walbaum), in the Northeast Pacific Ocean. J Fish Dis. **38**(4):365-378

4. **Rhodes LD, Rice CA, Greene CM, Teel DJ, Nance SL, Moran P, Durkin CA, Gezhegne SB**. 2011. Nearshore ecosystem predictors of a bacterial infection in juvenile Chinook salmon. Mar Ecol Prog Ser. **432**:161–172.

5. **Pascho RJ, Elliott DG, Achord S**. 1993. Monitoring of the in-river migration of s molts from two groups of spring chinook salmon, *Oncorhynchus tshawytscha* (Walbaum), with different profiles of *Renibacterium salmoninarum* infection. Aquac Res. **24**:163–169.

6. **Fredriksen Å, Endresen C, Wergeland HI**. 1997. Immunosuppressive effect of a low molecular weight surface protein from *Renibacterium salmoninarum* on lymphocytes from Atlantic salmon (*Salmo salar* L.). Fish Shellfish Immunol. **7**:273–282.

7. **Turaga P, Wiens G, Kaattari S**. 1987. Bacterial kidney disease: the potential role of soluble protein antigen(s). J Fish Biol. **31**:191–194.

8. **Wiens GD, Kaattari SL**. 1991. Monoclonal antibody characterization of a leukoagglutinin produced by *Renibacterium salmoninarum*. Infect Immun. **59**:631–637.

9. **Wood P, Kaattari S**. 1996. Enhanced immunogenicity of *Renibacterium salmoninarum* in chinook salmon after removal of the bacterial cell surface-associated 57 kDa protein. Dis Aquat Organ. **25**:71–79.

10. **Brown LL, Iwama GK, Evelyn TPT**. 1996. The effect of early exposure of Coho salmon (*Oncorhynchus kisutch*) eggs to the p57 protein of *Renibacterium salmoninarum* on the development of immunity to the pathogen. Fish Shellfish Immunol. **6**:149–165.

11. **Senson PR, Stevenson RMW**. 1999. Production of the 57 kDa major surface antigen by a non-agglutinating strain of the fish pathogen *Renibacterium salmoninarum*. Dis Aquat Organ. **38**:23–31.

12. **Wiens GD, Chien MS, Winton JR, Kaattari SL**. 1999. Antigenic and functional characterization of p57 produced by *Renibacterium salmoninarum*. Dis Aquat Organ. **37**:43–52.

13. **Coady AM, Murray AL, Elliott DG, Rhodes LD**. 2006. Both *msa* genes in *Renibacterium salmoninarum* are needed for full virulence in Bacterial Kidney Disease. Appl Environ Microbiol. **72**:2672–2678.

14. **O'Farrell C, Strom M**. 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 Aquat Organ. **38**:115–123.

15. **Rhodes LD, Coady AM, Strom MS**. 2002. Expression of duplicate *msa* genes in the salmonid pathogen *Renibacterium salmoninarum*. Appl Environ Microbiol. **68**:5480–5487.

16. **Wiens GD, Rockey DD, Wu Z, Chang J, Levy R, Crane S, Chen DS, Capri GR, Burnett JR, Sudheesh PS, Schipma MJ, Burd H, Bhattacharyya A, Rhodes LD, Kaul R, Strom MS.** 2008. Genome sequence of the fish pathogen *Renibacterium salmoninarum* suggests reductive evolution away from an environmental *Arthrobacter* ancestor. J Bacteriol. **190**:6970–6982.

17. **Rhodes LD, Coady AM, Deinhard RK**. 2004. Identification of a third *msa* gene in *Renibacterium salmoninarum* and the associated virulence phenotype. Appl Environ Microbiol. **70**:6488–6494.

18. **Toranzo AE, Barja JL, Colwell RR, Hetrick FM.** 1983. Characterization of plasmids in bacterial fish pathogen. Infect Immun. **39**:184–192.

19. **Fryer JL, Lannan CN**. 1993. The history and current status of *Renibacterium salmoninarum*, the causative agent of bacterial kidney disease in Pacific salmon. Fish Res. **17**:15–33.

20. **Brynildsrud O, Feil EJ, Bohlin J, Castillo-Ramirez S, Colquhoun D, McCarthy**

U, Matejusova IM, Rhodes LD, Wiens GD, Verner-Jeffreys DW. 2014. Microevolution of *Renibacterium salmoninarum*: evidence for intercontinental dissemination associated with fish movements. ISME J. **8**:746–756.

21. **Riehle MM, Bennett AF, Long AD**. 2001. Genetic architecture of thermal adaptation in *Escherichia coli*. Proc Natl Acad Sci. **98**:525–530.

22. **Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, Grassi A de, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavaré S, Deloukas P, Hurles ME, Dermitzakis ET**. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science. **315**:848–853.

23. **Hobbs MM, Seiler A, Achtman M, Cannon JG**. 1994. Microevolution within a clonal population of pathogenic bacteria: recombination, gene duplication and horizontal genetic exchange in the *opa* gene family of *Neisseria meningitidis*. Mol Microbiol. **12**:171–180.

24. **Marri PR, Paniscus M, Weyand NJ, Rendón MA, Calton CM, Hernández DR, Higashi DL, Sodergren E, Weinstock GM, Rounsley SD, So M**. 2010. Genome sequencing reveals widespread virulence gene exchange among human *Neisseria* species. PLoS ONE. **5**:e11835.

25. **Domenech P, Kolly GS, Leon-Solis L, Fallow A, Reed MB**. 2010. Massive gene duplication event among clinical isolates of the *Mycobacterium tuberculosis* W/Beijing Family. J Bacteriol. **192**:4562–4570.

26. **Brynildsrud O, Snipen L-G, Bohlin J**. 2015. CNOGpro: Detection and quantification of CNVs in prokaryotic whole-genome sequencing data. Bioinforma. Oxf Engl.

27. **Waack S, Keller O, Asper R, Brodag T, Damm C, Fricke WF, Surovcik K, Meinicke P, Merkl R**. 2006. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. BMC Bioinformatics. **7**:142.

28. **Evelyn T, Prosperi-Porta L, Ketcheson JE**. 1990. Two new techniques for obtaining consistent results when growing *Renibacterium salmoninarum* on KDM2 culture medium. Dis Aquat Organ. **9**:209–212.

29. **Margus T, Remm M, Tenson T**. 2007. Phylogenetic distribution of translational GTPases in bacteria. BMC Genomics **8**:15.

30. **Ronquist F, Huelsenbeck JP**. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. **19**:1572–1574.

31. **Naville M, Ghuillot-Gaudeffroy A, Marchais A, Gautheret D**. 2011. ARNold: a web tool for the prediction of Rho-independent transcription terminators. RNA Biol. **8**:11–13.

32. **Diniz-Filho JAF, Soares TN, Lima JS, Dobrovolski R, Landeiro VL, de Campos Telles MP, Rangel TF, Bini LM**. 2013. Mantel test in population genetics. Genet Mol Biol. **36**:475–485.

33. **Riley M, Anilionis A**. 1978. Evolution of the bacterial genome. Annu Rev Microbiol. **32**:519–560.

34. **Conant GC, Wolfe KH**. 2008. Turning a hobby into a job: How duplicated genes find new functions. Nat Rev Genet. **9**:938–950.

35. **Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV**. 2002. Selection in the evolution of gene duplications. Genome Biol. **3**.

36. **Kondrashov FA**. 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc R Soc B Biol Sci. **279**:5048–5057.

37. **Fredriksen Å, Bakken V**. 1994. Identification of *Renibacterium salmoninarum* surface proteins by radioiodination. FEMS Microbiol Lett. **121**:297–301.

38. **Wiens GD, Dale OB**. 2008. Renibacterium salmoninarum p57 antigenic variation is restricted in geographic distribution and correlated with genomic markers. Dis Aquat Organ. **83**:123.

39. **Evenden AJ, Grayson TH, Gilpin ML, Munn CB**. 1993. *Renibacterium salmoninarum* and bacterial kidney disease — the unfinished jigsaw. Annu Rev Fish Dis. **3**:87–104.

Figure legends

Figure 1 - Duplications w dotplot

Panel A) Schematic view of the three major copy number variant regions discovered in the current study.

Panel B) Genome dot plot of the major (type I) and minor (type I) *msa* duplication units to itself, showing repeat regions and palindromic sequence. Solid lines represent a minimum of 85% sequence identity. DR = Direct repeat. IR = Inverted Repeat. The 91 bp palindrome encodes a predicted rho-independent terminator with a central loop polymorphism between *msa1* and *msa2*. The polymorphism is located 37 bp downstream of the *msa* ORF. We could not resolve the correct orientation of this segment in duplications, and the polymorphism is therefore labeled by the ambiguity character S (C/G).

Figure 2 - Phylogeny

Figure 2 - Phylogenetic tree revealing patterns of CNV distribution. Horizontal branches represent patristic distances, and isolates are colored according to their nation of origin. Purple stars indicate CNV in the *msa* and *p22* genes, while an olive star represents CNV in other genes (detailed in Table 1). The most probable copy number of each of the *msa*, *p12* and *p22* genes is shown on the extreme right. Adapted from Brynildsrud *et al.*, (2013)

Tables

TABLE 1 *R. salmoninarum* isolates screened for copy number variation. Legend: sw/fw - saltwater/freshwater habitat; f/w - farmed/wild fish origin.

Sample ID	Host	sw/fw	f/w	Origin	Year	Alternative ID	EBI Accession no.
MT1351	*S. salar*	sw	f	Scottish Highlands, UK	1993		ERR327904
Carson 5b	*O. tshawytscha*	fw	f	Tyee Creek / Wind River, USA	1994		ERR327905
05372K	*O. tshawytscha*	sw	f	Grande Ronde Basin, USA	2005		ERR327906
NCIMB 1116	*S. salar*	fw	w	River Dee, UK	1962	96056	ERR327907
NCIMB 1114	*S. salar*	fw	w	River Dee, UK	1962	5005	ERR327908
MT1880	*S. salar*	sw	f	Strathclyde, UK	1996		ERR327909
MT1470	*O. mykiss*	fw	f	Tayside, UK	1994		ERR327910
NCIMB 2235	*O. tshawytscha*	sw	f	Oregon, USA	1974	ATCC33209	ERR327911
9025	*O. mykiss*	fw	f	Yorkshire, UK	2009	16251-1	ERR327912
MT239	*S. salar*			Scotland, UK	1988		ERR327913
MT1511	*O. mykiss*	fw	f	Strathclyde, UK	1994		ERR327914
Cow-chs-94	*O. tshawytscha*	fw		Cowlitz River, USA	1994	GR 16	ERR327915
MT444	*S. salar*	sw	f	Western Isles, UK	1988		ERR327916
MT839	*S. salar*	sw	f	Scottish Highlands, UK	1990		ERR327917
MT452	*O. mykiss*	fw	f	Dumfries and Galloway, UK	1988		ERR327918
MT861	*S. salar*	sw	f	Scotland, UK	1990		ERR327919
MT1363	*O. mykiss*	sw	f	Strathclyde, UK	1993		ERR327920
99333	*O. mykiss*	fw	f	Wales, UK	1998	980036-102	ERR327921
MT1262	*S. salar*	fw	f	Scottish Highlands, UK	1992		ERR327922
5007	*O. mykiss*			Scotland, UK	2005	0180-18	ERR327923
MT3313	*O. mykiss*	fw	f	Central Scotland, UK	2008		ERR327925
MT3277	*O. mykiss*	fw	f	Dumfries and Galloway, UK	2008		ERR327926
96071	*O. mykiss*	fw	f	Hampshire, UK	1996	TEST VALLEY FDL	ERR327927
MT3315	*O. mykiss*	fw	f	Strathclyde, UK	2008		ERR327928
MT2622	*O. mykiss*	sw	f	Strathclyde, UK	2002		ERR327929
1205	*O. mykiss*		f	UK	2001	3104-67	ERR327930
99327	*O. mykiss*	fw	f	UK	1997	970313-2	ERR327931
7105	*O. mykiss*		f	UK	2007	P0416 T83 10-3 2	ERR327932
MT3479	*S. salar*	sw	f	Orkney, UK	2008		ERR327933
MT3482	*S. salar*	sw	f	Strathclyde, UK	2009		ERR327934
MT2979	*O. mykiss*	fw	f	Scottish Highlands, UK	2005		ERR327935
MT2943	*S. salar*	sw	f	Scottish Highlands, UK	2005		ERR327936
99329	*O. mykiss*	fw	f	Wales, UK	1998	980036-125	ERR327937
99326	*O. mykiss*	fw	f	Wales, UK	1999	2119-8	ERR327938
MT3106	*O. mykiss*	fw	f	Strathclyde, UK	2006		ERR327939
99344	*O. mykiss*	fw	f	Hampshire, UK	1998	980106-1.1.5	ERR327940
MT3483	*S. salar*	sw	f	Strathclyde, UK	2009		ERR327941
5006	*O. kisutch*	sw	f	Bella Bella, Canada	1996	960046	ERR327942
99332	*O. mykiss*	fw	f	Wales, UK	1999	2119-3	ERR327943

Rs 8	*S. salar*	sw	f	New Brunswick, Canada	2008		ERR327944
Rs 10	*S. salar*	sw	f	New Brunswick, Canada	2009		ERR327945
Rs 4	*S. salar*	sw	f	New Brunswick, Canada	2006		ERR327946
Rs 3	*S. salar*	fw	f	New Brunswick, Canada	2005		ERR327947
99345	*O. mykiss*	fw	f	Wales, UK	1998	980070-18	ERR327948
99341	*O. mykiss*	fw	f	Hampshire, UK	1998	980109-20	ERR327949
Rs 5	*S. salar*	sw	f	New Brunswick, Canada	2007		ERR327950
Rs 2	*S. salar*	sw	f	New Brunswick, Canada	2005		ERR327951
BPS 91	*O. gorbuscha*			Nanaimo, Canada	1991		ERR327952
Rs 6	*S. salar*	sw	f	New Brunswick, Canada	2007		ERR327953
DR143	*S. fontinalis*	fw	w	Alberta, Canada	1972	GR 17	ERR327954
6553	*S. salar*	sw	f	Hemne, Norway	2008	2008-09-495	ERR327955
6642	*S. salar*		f	Hemne, Norway	2008	2008-06-633	ERR327956
Car 96	*O. tshawytscha*			Washington, USA	1996		ERR327957
684	*S. trutta*	fw	f	Aurland, Norway	1987		ERR327958
GR5	*T. thymallus*	fw	w	Montana, USA	1997	980036-87	ERR327959
WR99 c2	*O. kisutch*			Washington, USA	1999		ERR327960
D6	*O. tshawytscha*			Oregon, USA	1982		ERR327961
6694	*O. mykiss*	sw	f	Hemne, Norway	2008		ERR327962
BQ96 91-1	*O. kisutch*			Nanaimo, Canada	1996		ERR327963
5223	*S. salar*	sw	f	Kvinnherad, Norway	2005	2005-50-579	ERR327964
6863	*O. mykiss*	sw	f	Osterøy, Norway	2009		ERR327965
7441	*S. salar*		f	Storfjord, Norway	1985	1985-09-667	ERR327966
7450	*S. salar*		f	Askøy, Norway	1987	1987-09-1185	ERR327967
6695	*O. mykiss*	sw	f	Hemne, Norway	2008	2008-06-631	ERR327968
7449	*S. salar*		f	Skjervøy, Norway	1987	1987-09-932	ERR327969
7448	*S. salar*		f	Stranda, Norway	1986	1986-09-4366	ERR327970
7439	*S. salar*		f	Sognefjorden, Norway	1984	1984-40.992	ERR327971
ATCC 33209[a]	*O. tshawytscha*	sw	f	Oregon, USA	1974		NC_010168.1

[a] Type strain. Sequence data downloaded from Genbank

TABLE 2 *R. salmoninarum* isolates with copy number variation in virulence factors. The number sign (#) indicates number of copies. Baseline copy numbers are 2, 1 and 1 for *msa*, *p12* and *p22*, respectively. In the last column, duplicated genes are denoted by their locus_tag stripped by their prefix "RSal33209_" with common product names in parentheses.

Isolate	Accession	# *msa*	# *p12*	# *p22*	Other duplications (annotation) [#]
05372K	ERR327906	5	3	4	
5223	ERR327964	4	3	4	
5006	ERR327942	2	1	1	2,974,628 to 3,084,569 [2] and 3,088,016 to 3,100,482 [2]
BPS_91	ERR327952	4	2	3	
BQS96_91-1	ERR327963	4	3	2	
Carson5b	ERR327905	4	2	5	0109 (LacI family transcriptional regulator) [2], 1458 (NADH-dependent flavin oxidoreductase) [2], 2607 (ferredoxin NADP-reductase) [2], 3193 (hypothetical) [2]
Cow_Chs_94	ERR327915	4	2	2	
GR5	ERR327959	3	2	2	
Rs_2	ERR327951	3	2	2	
Rs_6	ERR327953	5	3	4	
Rs_10	ERR327945	3	2	2	

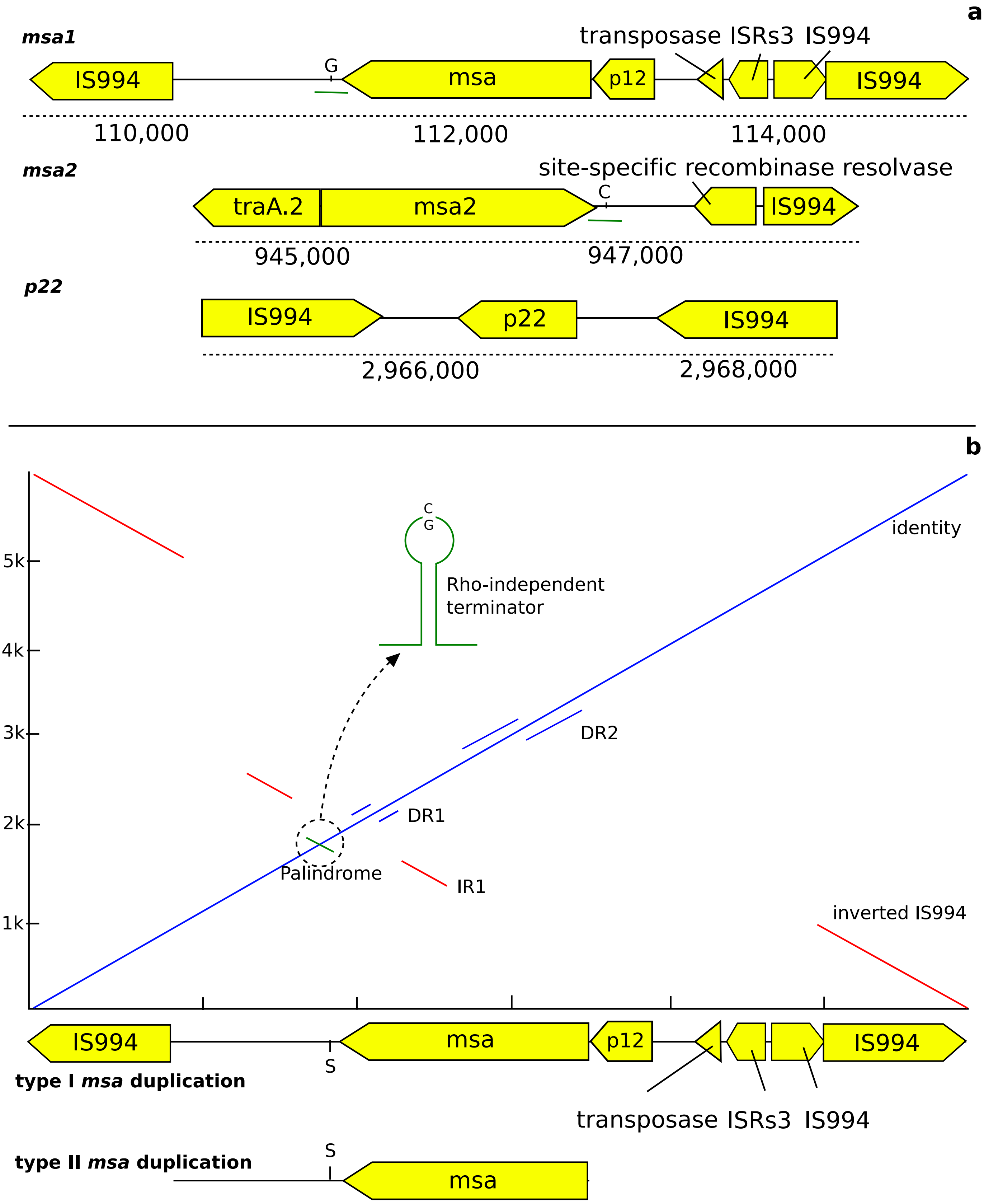


Fig 1 a) Schematic view of the three major copy number variant regions discovered in the current study. b) Genome dot plot of the major (type I) and minor (type I) msa duplication units to itself, showing repeat regions and palindromic sequence. Solid lines represent a minimum of 85% sequence identity. DR = Direct repeat. IR = Inverted Repeat. The 91 bp palindrome encodes a predicted rho-independent terminator with a central loop polymorphism between *msa1* and *msa2*. The polymorphism is located 37 bp downstream of the *msa* ORF. We could not resolve the correct orientation of this segment in duplications, and the polymorphism is therefore labelled by the ambiguity character S (C/G)

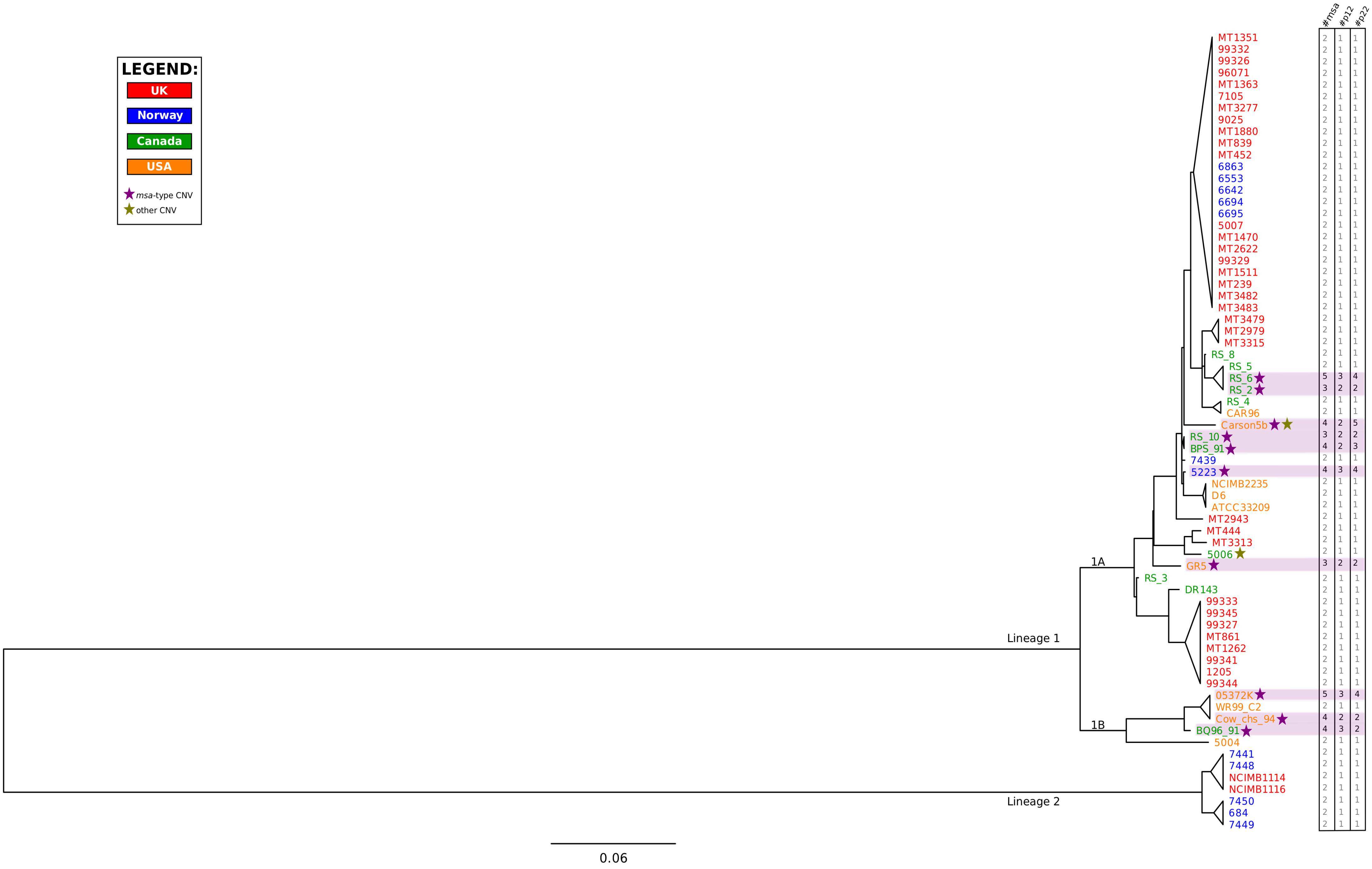


Fig 2 Phylogenetic tree revealing patterns of CNV distribution. Horizontal branches represent patristic distances, and isolates are colored according to their nation of origin. Purple stars indicate CNV in the *msa* and *p22* genes, while an olive star represents CNV in other genes (detailed in Table 1). The most probable copy number of each of the *msa*, *p12* and *p22* genes is shown on the extreme right. Adapted from Brynildsrud *et al.*, (2013)

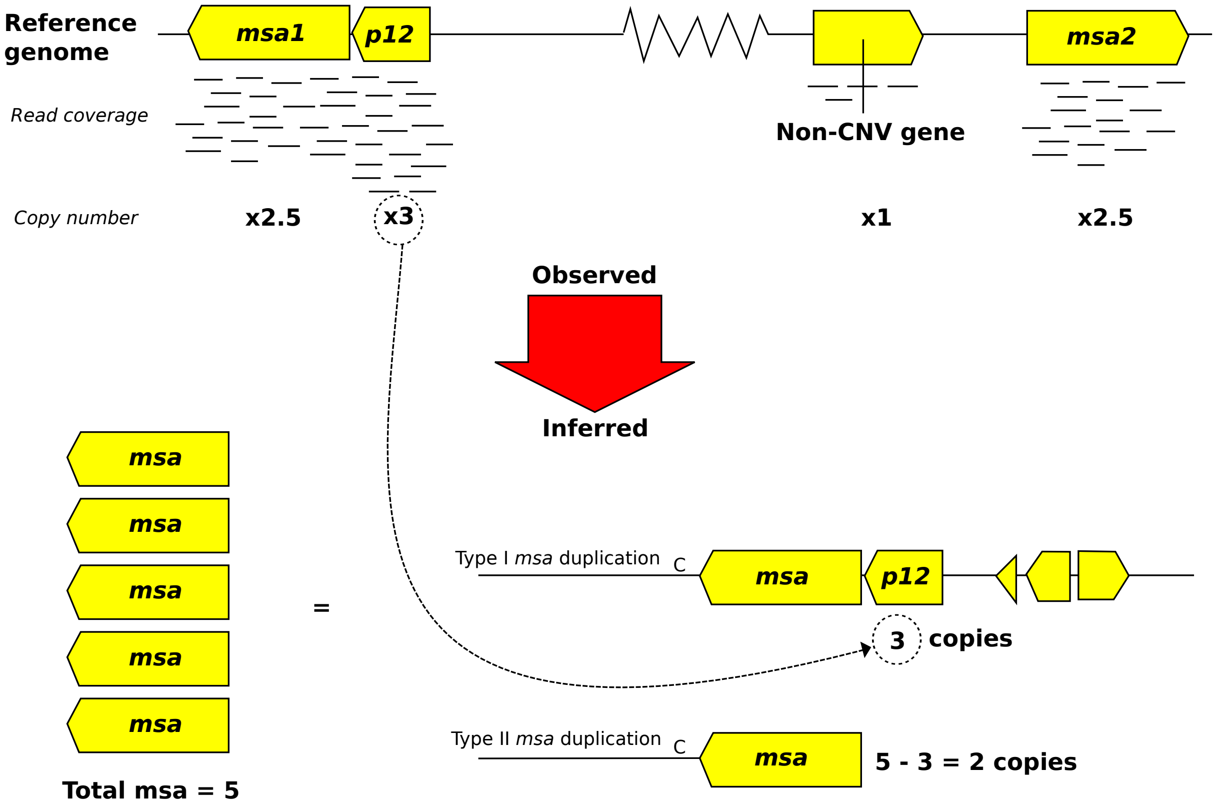


FIG S1 Detail on how the relative frequencies of type I and type II *msa* duplications were inferred from the read coverage data.

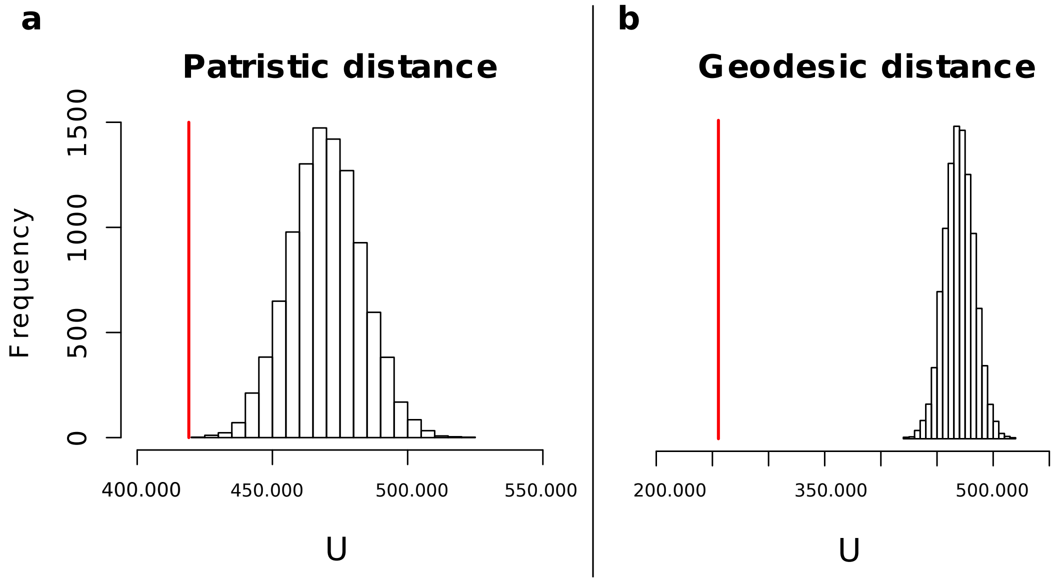


FIG S2 Mann-Whitney U test statistic distribution in the Mantel correlation analysis. Correlation is measured between pairwise patristic (panel a) and geodesic (panel b) distances to identical CNV status, measured as a binary trait. The vertical red line represents our observed statistic and the white boxes represent the histogram of the 10.000 permuted matrix-statistics. Note the gaussian distribution of Us for both the patristic (Panel a) and geodesic (Panel b) distance analyses. The increased distance between our observed U and the permuted matrix-Us in panel b indicate a more extreme correlation.

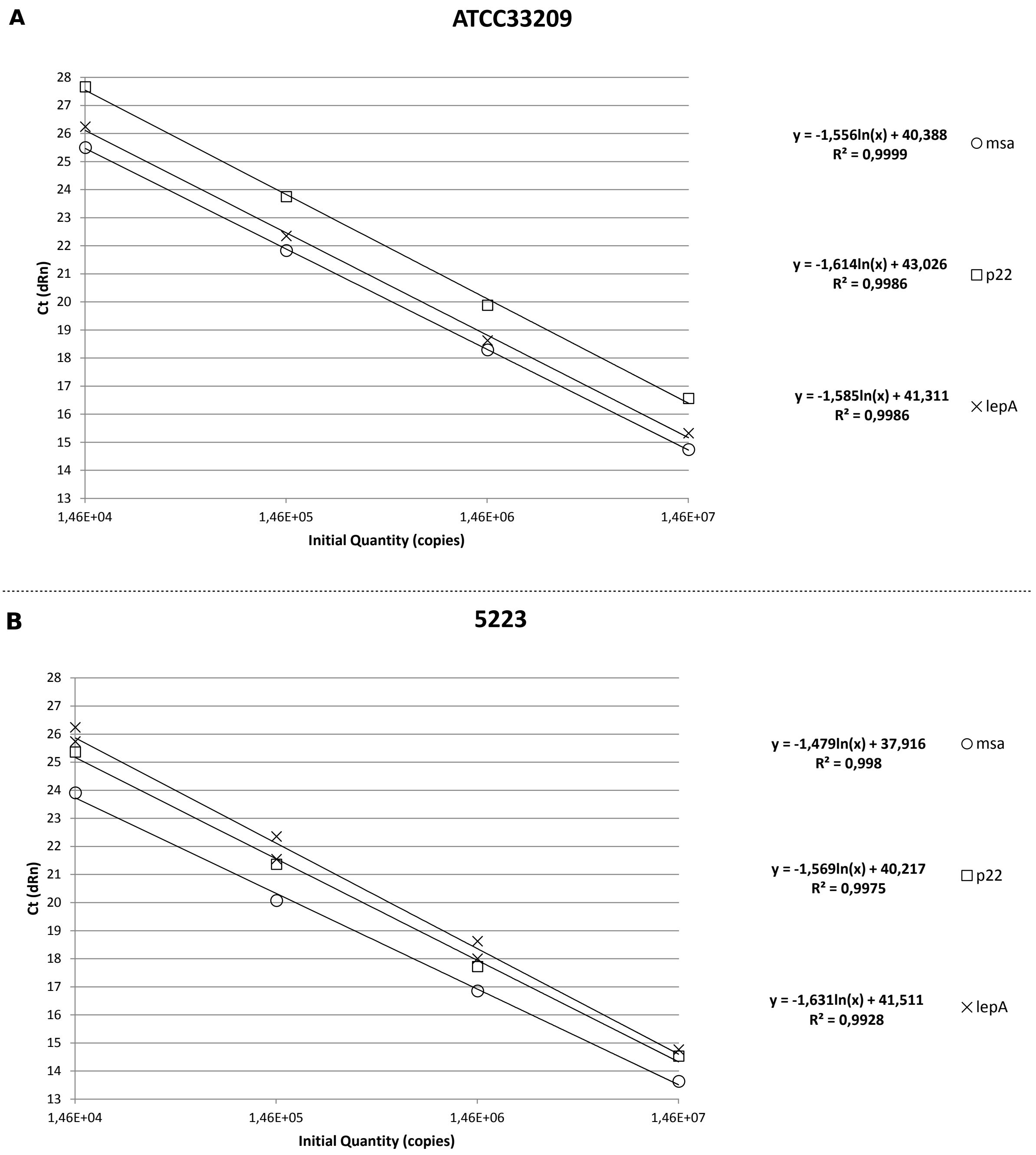


FIG S3 Standard curves of the chosen primer pairs for A) the reference strain (ATCC33209) and B) a CNV-positive strain (5223 / NVI-5223).

Table S1 Primers

Target gene	Primer Name	Sequence (5' - 3')	Length	Mol.wt.	Tm
msa	MSA1_1507F	GATGCCCAGACTGTTGCCT	19	5780.8	70
	MSA1_1611R	CTCAAAAACACCGAAACTCGTCTTA	25	7564.0	71
p22	p22_140F	AGAACACTTCTGACTTTGTGGTAGATA	27	8314.4	72
	p22_234R	GCTTGCTTGGTTGAGCGTAAA	21	6493.2	69
lepA	lepA_1490F	CGGATCTGGTCAAGGTCGATATT	23	7095.6	72
	lepA_1604R	CGCAATTTCCCCGTCATCATC	21	6278.2	71

Details of primers used in this study. Mol.wt. = molecular weight in Daltons. Tm = melting temperature.

Table S2 Putative alien ORFs

	Start	Stop	ORF	Name	Suggested donor
🟧	111,225	112,901	RSal33209_0131	*msa1*	???
🟧	112,912	113,250	RSal33209_0132	hypothetical protein	Bacilli 6
🟧	113,450	113,596	RSal33209_0133	transposase	Actinobacteria 3
🟧	113,625	113,882	RSal33209_0134	IS*Rs3*	Bacilli 8
🟧	113,959	114,324	RSal33209_0135	IS*994*	Bacilli 8
	512,680	513,201	RSal33209_0614	flavin reductase-like FMN-binding protein	Actinobacteria 5
🟦	944,415	945,059	RSal33209_1118	*traA.2*	Bacilli 4
🟦	945,077	946,777	RSal33209_1119	*msa2*	???
🟦	947,576	947,890	RSal33209_1120	site-specific recombinase resolvase	Actinobacteria 5
🟦	947,972	948,337	RSal33209_1121	IS*994*	Bacilli 8
🟪	1,318,739	1,319,110	RSal33209_1541	flavodoxin	Gammaproteobacteria 17
🟪	1,319,135	1,319,740	RSal33209_1542	acetyltransferase	Bacilli 8
🟩	1,479,328	1,480,179	RSal33209_1718	hypothetical protein	???
🟩	1,480,176	1,480,562	RSal33209_1719	hypothetical protein	Gammaproteobacteria_27
🟩	1,480,730	1,481,371	RSal33209_1720	*TetR* family transcriptional regulator	Actinobacteria_3
	1,944,020	1,944,535	RSal33209_2194	esterase	Actinobacteria 5
🟥	2,304,457	2,304,936	RSal33209_2566	hypothetical protein	Gammaproteobacteria 17
🟥	2,305,097	2,305,408	RSal33209_2567	hypothetical protein	Bacilli 8
🔵	2,305,833	2,306,492	RSal33209_2569	tetracycline repressor protein class #3	Bacilli 2
🔵	2,306,507	2,307,334	RSal33209_2570	short chain dehydrogenase	???
	2,373,965	2,374,495	RSal33209_2647	hypothetical protein	Actinobacteria 5
	2,843,772	2,843,942	RSal33209_3197	hypothetical protein	Bacteroides 1

Donor family	Members	Habitat
Bacilli 6	*Streptococcus sobrinus*	Oral cavity
Actinobacteria 3	*Saccharopolyspora spinosa*	Rum still
Bacilli 8	*Oceanobacillus iheyensis HTE831*	Deep sea
Actinobacteria 5	*Streptomyces clavuligerus*	Soil
Bacilli 4	*Tetragenococcus halophilus*	Soy sauce production
Gammaproteobacteria 17	*Vibrio mimicus*	Marine organisms
Gammaproteobacteria 27	*Aeromonas salmonicida*	Salmon, water
Bacilli 2	*Streptococcus criceti*	Oral cavity
Bacteroides 1	*Bacteroides fragilis*	Anaerobic environments

Donor family	Number of donated genes
Bacilli 8	5
Actinobacteria 5	4
Actinobacteria 3	2
Gammaproteobacteria 17	2
Bacilli 6	1
Bacilli 4	1
Gammaproteobacteria 27	1
Bacteroides 1	1
Bacilli 2	1

Top: List of all ORFs identified as putative alien sequence by the Sigi-HMM tool. Consecutive ORFs are colored identically. Middle: Type organism for donor families with typical habitat. Bottom: Frequencies of the different donor families.