Molecular Phylodynamics and Protein Modeling of Infectious Salmon Anemia Virus (ISAV)

Castro-Nallar *et al.*

BMC
Evolutionary Biology

Molecular Phylodynamics and Protein Modeling of Infectious Salmon Anemia Virus (ISAV)

Eduardo Castro-Nallar[1*], Marcelo Cortez-San Martín[2], Carolina Mascayano[3], Cristian Molina[2,3] and Keith A Crandall[1]

Abstract

Background: ISAV is a member of the *Orthomyxoviridae* family that affects salmonids with disastrous results. It was first detected in 1984 in Norway and from then on it has been reported in Canada, United States, Scotland and the Faroe Islands. Recently, an outbreak was recorded in Chile with negative consequences for the local fishing industry. However, few studies have examined available data to test hypotheses associated with the phylogeographic partitioning of the infecting viral population, the population dynamics, or the evolutionary rates and demographic history of ISAV. To explore these issues, we collected relevant sequences of genes coding for both surface proteins from Chile, Canada, and Norway. We addressed questions regarding their phylogenetic relationships, evolutionary rates, and demographic history using modern phylogenetic methods.

Results: A recombination breakpoint was consistently detected in the Hemagglutinin-Esterase (*he*) gene at either side of the Highly Polymorphic Region (HPR), whereas no recombination breakpoints were detected in Fusion protein (*f*) gene. Evolutionary relationships of ISAV revealed the 2007 Chilean outbreak group as a monophyletic clade for *f* that has a sister relationship to the Norwegian isolates. Their tMRCA is consistent with epidemiological data and demographic history was successfully recovered showing a profound bottleneck with further population expansion. Finally, selection analyses detected ongoing diversifying selection in *f* and *he* codons associated with protease processing and the HPR region, respectively.

Conclusions: Our results are consistent with the Norwegian origin hypothesis for the Chilean outbreak clade. In particular, ISAV HPR0 genotype is not the ancestor of all ISAV strains, although SK779/06 (HPR0) shares a common ancestor with the Chilean outbreak clade. Our analyses suggest that ISAV shows hallmarks typical of RNA viruses that can be exploited in epidemiological and surveillance settings. In addition, we hypothesized that genetic diversity of the HPR region is governed by recombination, probably due to template switching and that novel fusion gene proteolytic sites confer a selective advantage for the isolates that carry them. Additionally, protein modeling allowed us to relate the results of phylogenetic studies with the predicted structures. This study demonstrates that phylogenetic methods are important tools to predict future outbreaks of ISAV and other salmon pathogens.

Background

Infectious salmon anemia virus (ISAV) is a pathogen that has been associated with high fish mortality in the aquaculture industry [1]. The cumulative mortality associated with each outbreak of ISAV in Norway and other countries is very high, reaching up to 100% of fish stock in some cases [2-6]. ISAV has been classified as the only member of Isavirus genus belonging to the *Orthomyxoviridae* family, which includes Influenza viruses [7].

Virions consist mainly of a membranous envelope with two surface glycoproteins and a matrix protein surrounding a ribonucleoprotein complex. Association of nucleoprotein, three polymerase subunits and the genomic RNA, in turn, forms the ribonucleoprotein complex. Virion morphology varies from spherical to pleiomorphic. The enveloped virus particles are of 45-140 nm in diameter, however, highly pleiomorphic particles up to 700 nm in the longest dimension are occasionally observed. The genome organization is consistent with other members of the *Orthomyxoviridae*; comprising of 8 segments of single stranded negative sense RNA ((-) ssRNA), in which each segment encodes one protein

* Correspondence: castronallar@gmail.com
[1]Department of Biology, 401 Widtsoe Building, Brigham Young University, Provo, UT 84602-5181, USA
Full list of author information is available at the end of the article

except segments 7 and 8, which encode two (European strains)/three (Canadian strains) and two proteins, respectively [8]. Four major structural proteins have been identified, including a 68 kDa nucleoprotein, a 22 kDa matrix protein, a 42 kDa surface glycoprotein named haemagglutinin-esterase protein with receptor-binding and receptor-destroying activity, and a 50 kDa surface glycoprotein with fusion activity, coded by genome segments 3, 8, 6, and 5, respectively. Segments 1, 2, and 4 encode the putative viral polymerase subunits PB1, PB2, and PA. The ORF1 of segment 7 encodes a nonstructural protein with interferon antagonistic properties, while ORF2 encodes a nuclear export protein. The smaller ORF1 of segment 8 encodes the matrix protein, while the larger ORF2 encodes an RNA-binding structural protein also with interferon antagonistic properties [4,9-11].

Regarding the genes coding for surface proteins, *he* genes exhibit a length polymorphism close to the 3' end called highly polymorphic region (HPR) that has been related to virulence, although this is not entirely clear [12-14]. In turn, *f* genes seem to be more conserved in length, with the exception of the region next to the putative protease-processing site (PPPS) [15,16]. The fusion protein in ISAV is synthesized as a precursor protein designated as F0. In order to exert its biological function, F0 must be cleaved by cellular proteases to generate F1 and F2 [17]. Several insertions of 10-12 amino acids (IN1-4), identical to sequences from other genomic segments, have been suggested to confer novel protease cleavage sites [15], and thus they could be subjected to positive selection.

Unlike other members of the *Orthomyxoviridae* family that commonly infect mammals or birds, ISAV naturally infects fishes of the *Salmonidae* family. Although natural outbreaks of ISA have only been recorded in farmed Atlantic salmon (*Salmo salar*), the virus has been isolated also from rainbow trout (*Oncorhynchus mykiss*) [3] and Coho salmon (*Oncorhynchus kisutch*) in Chile [18]. The disease was first reported in Norway and the virus was associated with the disease in 1984. At present, ISAV has been reported in Canada (New Brunswick in 1996 and Nova Scotia in 2000), USA (Maine in 2001), Scotland (1998), The Faroe Islands (2000; Denmark), and recently in Chile [6,19,20]. Because of the worldwide distribution and the potential for serious economical losses, ISAV has been listed as a non-exotic disease for the European Union (EU) and is therefore monitored closely by the European Community Reference Laboratory for Fish Diseases http://www.crl-fish.eu/.

In general, phylogenetic studies of infectious diseases have been centered on human and zoonotic diseases with little attention to enzootic pathogens, although with few exceptions, for example, in plant viruses' crop-related diseases [21-24]. Given the human health impact, phylogenetic studies of *Orthomyxoviruses* have been largely devoted to influenza viruses in particular to influenza A [25,26]. To date, questions related to global distributions, molecular evolution and adaptations, sources of genetic diversity and emergence of new isolates have been extensively addressed in the influenza literature. However, little has been done with its relatives, e.g., ISAV, even when the same genetic structure, and presumably, similar evolutionary forces might be acting on ISAV populations.

In 1999 an ISAV belonging to the North American (NA) genotype was isolated from Coho salmon in Chile and the first outbreak of ISAV in marine-farmed Atlantic salmon in the Southern hemisphere occurred in the same country starting in June 2007. Given that there are no known natural hosts for ISAV in Chile, only a few studies have investigated whether the outbreak source came from North America or Europe by using basic phylogenetic methods [6,15,19,20].

In this study, we apply more sophisticated phylogenetic and population genetic methods to currently available sequence data to address questions regarding the 2007 Chilean ISAV outbreak. In particular, we asked (i) whether recombination is a relevant force acting on *f* and *he* genes, (ii) whether the age estimates of the Chilean clade agree with epidemiological data, and (iii) whether population processes leave marks in ISAV genomes that allow us to reconstruct its demographic history. Finally, we developed new improved structural homology models with which positively selected sites were mapped to relate them with functional traits.

Results and discussion

We included 70 isolates from Canada, Chile and Norway for which both genes, *f* and *he*, were available along with their sampling dates. The length of the final alignments was 1365 and 1269 bp, respectively (additional file 1). We selected TVM as the best-fit nucleotide substitution model for both genes based on Akaike Information Criterion (AIC) [27]. Rate heterogeneity was estimated as the gamma distributed shape parameter α = 0.5740 and 0.6570 for *f and he*, respectively. We also tested for whether the data were best explained by a strict or relaxed molecular clock. Bayes Factor hypothesis testing favored the relaxed model in both cases (10.44 and 21.82 log BF for *f and he*, respectively). This was further supported by the Coefficient of Variation and ucld.stdv parameters.

Genetic diversity and recombination

In order to explore the data and to see whether recombination plays a recognizable role in ISAV evolution, we estimated the population genetic parameter Theta (θ) and recombination rate (r) for each population. When

calculated under the infinite-sites assumption, the Canadian and Chilean populations were more diverse than the older Norwegian population (0.08579, 0.06247 and 0.01809, respectively). In turn, when using a coalescent sampler estimator, the Norwegian population turned out to be, as expected, more than three times more diverse 0.04774 (95% CI 0.0198-0.08448) than the young Chilean population 0.0142 (95% CI 0.008368-0.020745), and than the Canadian population 0.03523 (95% CI 0.021083-0.051501) as well. Although at first this seems contradictory, coalescent-based estimators of genetic diversity are far more reliable than those based on segregating sites [28]. By taking into account evolutionary history, coalescent-based estimates are more robust to model assumptions and allow us to determine factors influencing the observed patterns (see below). This overall result has practical implications as less diverse populations, i.e., genetically homogeneous, would be more susceptible to vaccines and antiviral treatments whereas more structured populations, i.e., individuals genetically similar within populations but different among populations, make effective vaccine and antiviral deliveries more difficult. So, population genetic inferences can be instrumental in fish health surveillance programs as demonstrated in human health settings [29-32].

Antisense (-) ssRNA viruses are less prone to recombination than their (+) sense counterparts [33,34]. In principle this would arise from the impossibility of (-) ssRNA viruses to serve as templates for negative sense strand synthesis once the genome is uncoated, in turn rendering copy choice recombination less likely. In addition, (-) ssRNA genomes are generally protected by viral nucleoproteins that prevent genome exposure to the cytoplasm (and cell machinery) and to come in close contact with other genome copies, thus rendering copy-choice recombination also unlikely. Also, the segmented genome structure of some (-) ssRNA relaxes even more the occurrence of recombination by yielding the same outcome when shuffling segments (reassortment). In agreement with this hypothesis, our results showed a low overall recombination force (r) of (1.173 × 10^{-3} (1.90 × 10^{-4} - 3.50 × 10^{-3})) expressed as the ratio of recombination rate per site over substitution rate per site. However, recombination does happen in human [35-37] and avian [38] strains of influenza viruses as well as in some segments in ISAV, namely segments 6 and 5 [12-14]. In this sense, we tested our dataset to see whether or not it is possible to detect recombination breakpoints. Using a suite of methods [39], we consistently detected one possible recombination breakpoint in the vicinity of the *he* HPR region. This result was confirmed by substitution and phylogenetic algorithms and was further corroborated by phylogenetic topological incongruence using a genetic algorithm (GARD; Figure 1B). Since full-length HPR (HPR0) has

been detected only in natural populations (not virulent) and different HPR alleles co-occur in the same fish-farming location, these results support the hypothesis that the HPR region is the product of template switching during replication (unequal recombination), and opens the possibility of testing it in controlled experimental settings. Early indications of recombination in the *he* gene were reported in 2001, when researchers looked for topological incongruence and possible recombination patterns 'by eye' within *he* Norwegian sequences [40]. Interestingly, they found that recombination was associated with the geographic location of the isolates and that it occurred at either side of the HPR region. Our results provide statistical support for this initial observation as we detected a topological incongruence at nucleotide 900 (KH-adjusted *p*-value = 2 × 10-4; Figure 1B). We also detected recombination breakpoint signals at either side of the HPR; however, it was not possible to detect parental sequences with confidence (Rdp, maxchi, chimera and Genconv; Figure 1B). This evidence contradicts the hypothesis of Cunningham et al. [41] that the most plausible mechanism of recombination is that each HPR comes from partial deletions of precursor HPR0.

Regarding the *f* gene, we do not detect any recombination breakpoints. At first this was surprising since small non-homologous fragments have been observed at around 266-276, which might confer novel protease cleavage sites (IN1-4) and therefore have a fitness impact. We think there are at least two potential reasons for our inability to detect recombination in this gene region. First, small fragments are hard to detect with current methods because of the low information they contain. Second, methods are not designed to explicitly look for non-homologous recombination. Also, it would be hard to detect if the INs are evolutionarily linked to the *f* gene and/or have similar substitution rates.

Phylogenetic inference and past population dynamics

In order to estimate phylogenetic relationships, we used Bayesian phylogenetic methods to test the Chilean outbreak isolates for monophyly, to estimate rates of molecular evolution across the genomes, and estimate divergence times for clades of particular interest (e.g., geographically isolated outbreaks). Recombination has detrimental effects on phylogeny estimation and phylogeny-based analyses [34,42,43] and thus the recombinant portion of the *he* gene was not included in these analyses. Segmented viruses have the potential to re-assort their segments in novel combinations, which in phylogenetic terms has the same effect as recombination. Therefore, we estimated phylogenies and molecular evolution parameters separately for each gene since concatenating both datasets would render inaccurate inferences as ISAV might exhibit reassortment between segments.
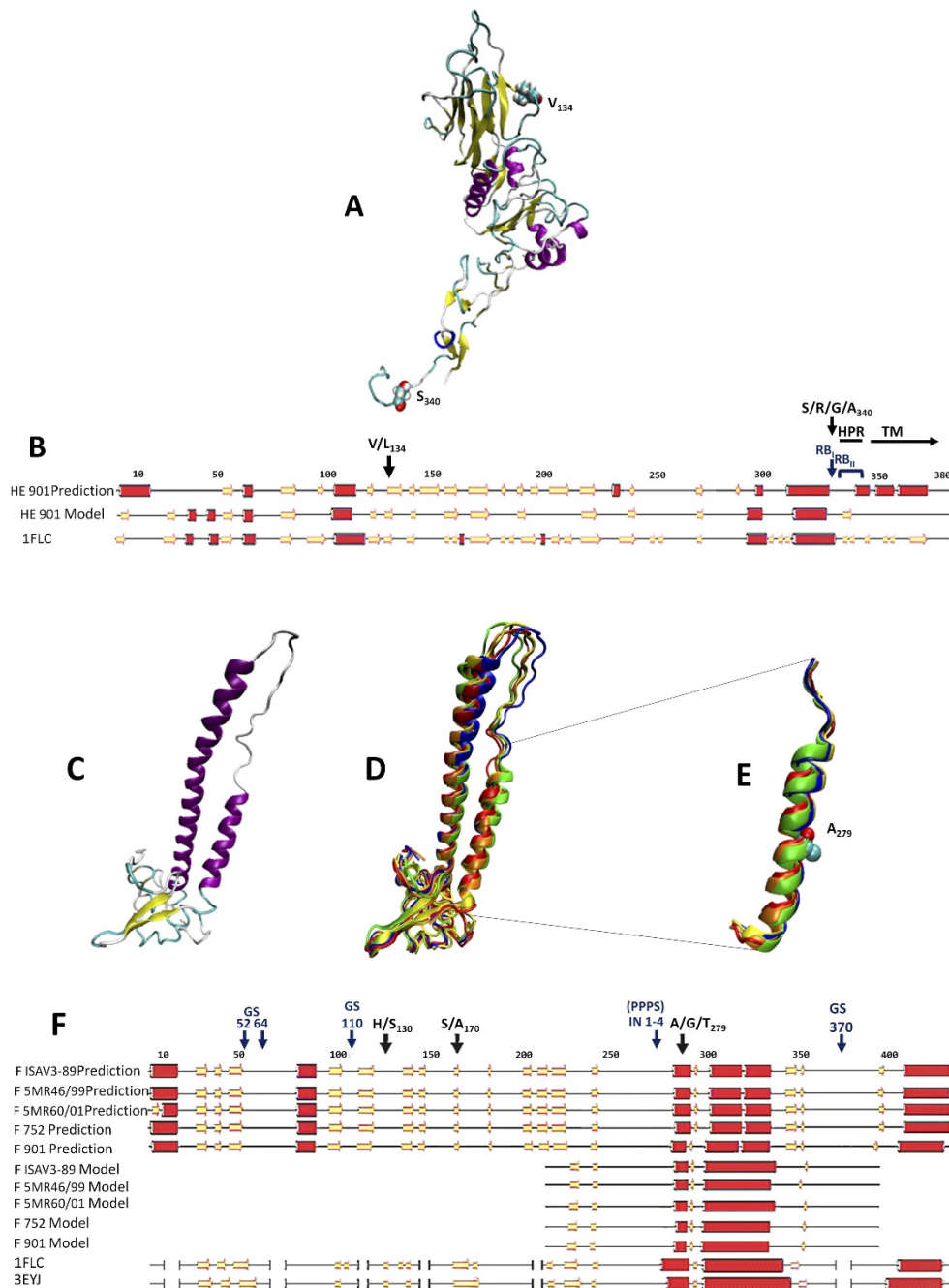
Figure 1 Schematic representation mapping of positively selected sites and tertiary structure models for both ISAV surface proteins. **(A)** Structural model of the HE protein from ISAV by homology with the HEF protein from influenza C virus (PDB ID 1FLC) with V134 and S340 residues in Van der Waals representation. **(B)** Alignment of secondary structural prediction (using PSIPRED) for sequence and model of ISAV HE against HEF protein. The alpha helix structural prediction is shown as red cylinders and Beta strands in yellow arrow representation. **(C)** Structural model of the Fusion protein from ISAV 901 (without insertion) by homology with HEF protein from influenza C virus (PDB ID 1FLC) and HE from Influenza A virus (PDB ID 3EYJ). **(D)** Superposition of models of the ISAV fusion protein with and without different insertions. **(E)** Zoomed image of alpha helix, which contains putative fusion peptides and different insertions with A279 residue in Van der Waals representation. **(F)** Alignment of secondary structural prediction (using PSIPRED) for different sequences and models of ISAV Fusions against HEF protein. The alpha helix structure prediction is shown as red cylinders and Beta strands as yellow arrows. Positively selected sites were deemed as such when the dN/dS rate ratio was greater than 1 and with a *p*-value less than 0.1. Amino acid numbering according to ISAV752 09 (ADF36500 for Hemagglutinin and ADF36499 for Fusion). Black arrows denote positively selected sites. Blue labels indicate landmarks in the genes, i.e., GS, glycosylation site; IN, insertion; RB, recombination breakpoint; PPPS, putative protease-processing site; HPR, Highly Polymorphic Region; TM, Transmembrane domain.

In the case of the *f* gene, the Chilean outbreak clade was inferred as monophyletic suggesting that ISAV was introduced once in Chilean fish farming (Figure 2). However, the *he* gene tree (additional file 2) shows that outbreak isolates spread in two clades, i.e., paraphyletic; one composed of most of the outbreak isolates and the other composed of three isolates that are also closely related to isolates from Norway (VT11282007-38/-39 and 2006B13364 in additional file 2). For both genes, posterior probability support was close or equal to 1, either for *f* monophyletic outbreak clade or *he* paraphyletic outbreak clades. We speculate that this might be the consequence of reassortment as *he* genes from Chile exhibit two evolutionary pathways. Interestingly, the HPR0 allele (SK779/06HPR0 in Figure 2) is suggested to be the precursor of all HPR alleles by differential deletion of the polymorphic region [12]. However, in both gene trees, SK779/06HPR0 (not virulent) does not share a most recent common ancestor with all the alleles, but it does for the Chilean outbreak alleles. This result suggests that the ancestor of the HPR0 allele is not necessarily the precursor of all modern HPRs. Nevertheless, this result agrees with the hypothesis that the ISAV that finally led to the Chilean outbreak was not virulent [20]. Overall, these results are in rough agreement with previously published phylogenies using other methods. However, our analysis fully resolved relationships in crown nodes and allowed us to estimate molecular evolutionary parameters taking into account phylogenetic and parameter uncertainty. We focused in two events in the evolutionary history of the virus, the time in which Norwegian and Chilean isolates diverged, and the time in which the Chilean epidemic clade expanded. Whether it was in Norway or in Chile, the most recent common ancestor (MRCA) between Norwegian and Chilean isolates was estimated to have occurred in 1995 [1992-1996 95% Highest Probability Density (HPD)] and the Chilean clade expansion in 2003 (2001-2005 95% HPD) for the *f* gene. For the *he* gene, in turn, the MRCA was 1989 (1982-1994 95% HPD) and 2001 (1996-2004 95% HPD). This is in agreement with reported distance-based analyses for the *f* gene but not for the *he* gene [20]. However, those methodologies do not provide confidence estimates in a probabilistic framework, which, in turn, could allow for meaningful comparisons. For instance, our estimates show that both segments have different evolutionary histories since they coalesce at significantly different points back in time (0.83 and 0.94 posterior probabilities for *he* and *f* Chilean clades, respectively). Together with the paraphyly of the *he* outbreak clade, the results suggest that the ancestor between Norwegian and Chilean isolates underwent reassortment of at least both surface protein genes. Ancestral State Reconstruction (ASR) clearly shows that

Chilean isolates came from Norway as previously hypothesized and this result is consistent across both genes (color-coded in Figure 2 and additional file 2).

As other (-) ssRNA viruses, ISAV exhibited extremely high substitution rates. Similar to influenza viruses, ISAV replication machinery lacks proofreading activity, which has been proposed as the increased source of variability in RNA viruses [33,44]. Accordingly, substitution rates were estimated at 7.83×10^{-4} substitutions per site per year (s/s/y) (4.31×10^{-4} - 1.18×10^{-3} 95% HPD) for the *f* gene and 1.039×10^{-3} (s/s/y) (3.87×10^{-4} - 1.86×10^{-3} 95% HPD) for the *he* gene (Figure 2).

An interesting aspect of phylogenetic analyses is to infer processes from patterns in the time coordinate. We hypothesized that if ISAV is such a fast-evolving entity, its genome must contain information that could be used to infer ecological and/or population processes. For ISAV, we inferred a population level process of a bottleneck, i.e., a population going through one or more generations of small size followed by subsequent population growth, perhaps associated with the outbreak it recently experienced around 2007 when it was first officially reported in Chile http://www.sernapesca.cl. We reconstructed generation time-scaled effective population size ($N_{ef}*\tau$) through time to see whether or not it is possible to recover that bottleneck using the Bayesian Skyline non-parametric model (BSP) [45] (Figure 3). Interestingly, the analysis shows a strong reduction in population size around 2007 followed by a nearly exponential growth. In particular, the greatest reduction in population size was inferred from the *f* gene and occurred in 2006.42 and 2006.02 for the *he* gene (Figure 3). One assumption of analyses based on the coalescent and thus the BSP is that the dataset comprises one non-structured population. Although not explored in depth in the literature, population structure, in principle, could lead to biased estimates of substitution rates and associated parameters. This, in turn, can produce patterns in the BSP that reflect changes in the degree of structure instead of changes in the overall population size, especially when using a single marker [46,47]. To address this potential problem, we calculated Parsimony Score (PS) and Association Index (AI) for both genes under study. These statistics aim to assert the degree of association of a discrete character, geographic location in this case, with a posterior distribution of trees. Both PS and AI statistics were significant (*p*-value < 0.05), suggesting that it is possible to reject the null hypothesis of one non-structured population. Considering that two independent genes showed similar patterns, and that the BSP correlates well with surveillance data, plus the agreement of substitution rate estimates with others [20,48], we think is safe to assume that these results reflect a population bottleneck. Certainly, simulation studies will aid to assert how robust BSP estimates are to structured populations.
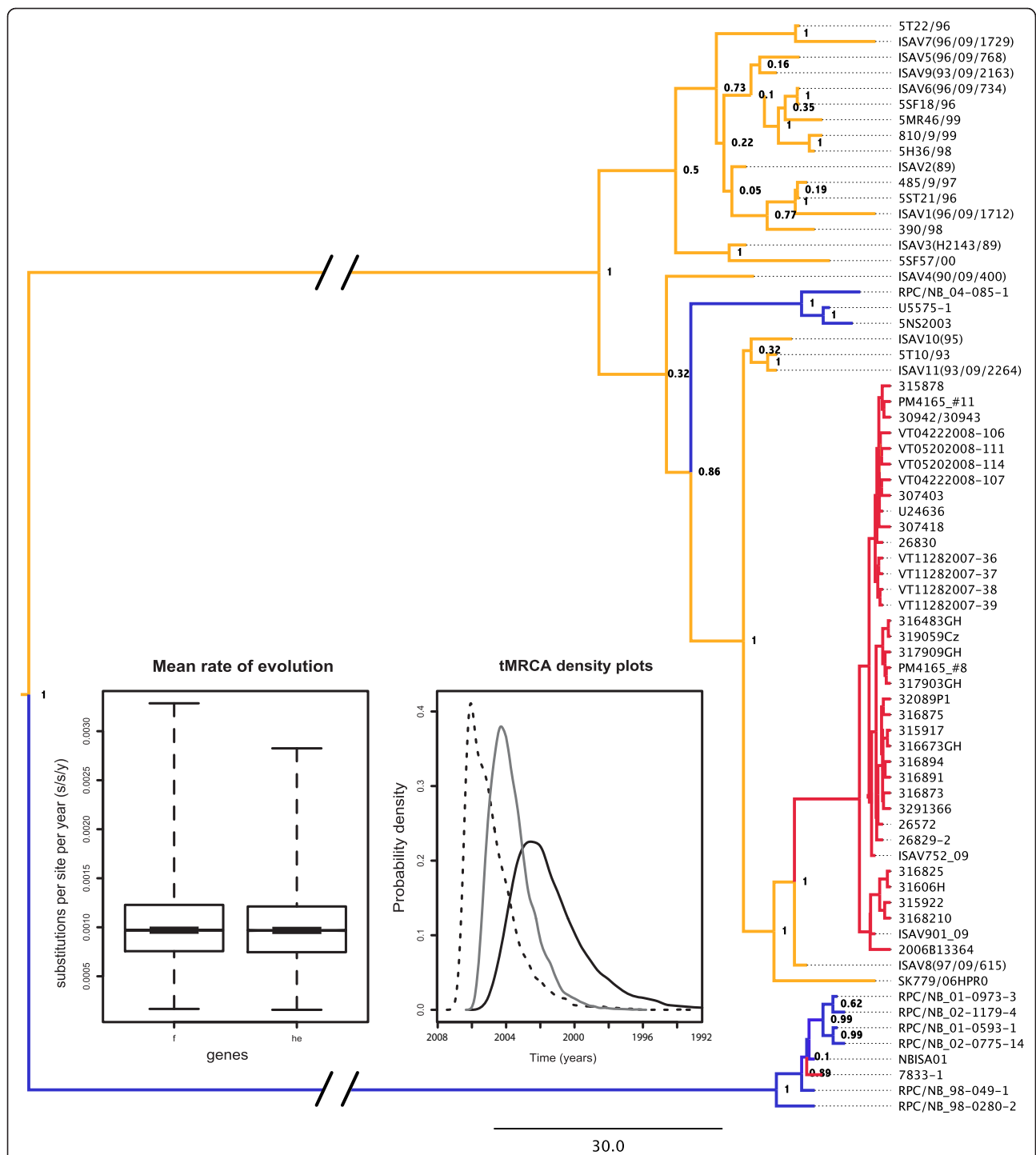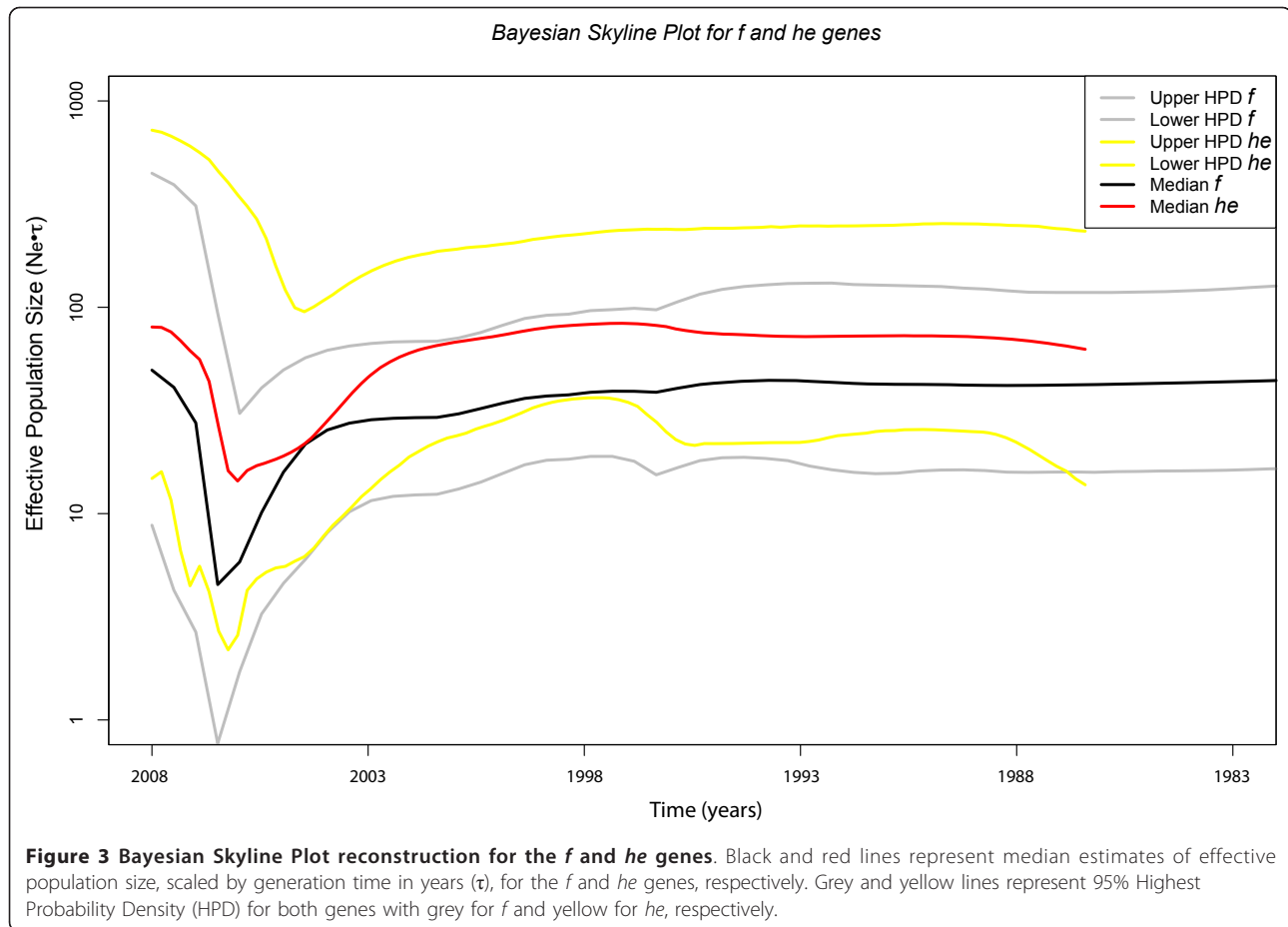
Figure 2 Bayesian Phylogenetic Inference for ISAV *f* gene. Maximum clade credibility (MCC) phylogeny depicting the evolutionary relationships of isolates from Europe, Canada and Chile. Colors indicate ancestral locations as inferred from Ancestral State Reconstruction (ASR). Red = Chile, Blue = Canada and Yellow = Europe. Support is expressed as posterior probabilities for major nodes. Inset = boxplots of evolutionary rates for the *f* and *he* genes and density plots for node age estimates (black dashed line = Chilean *he* minor clade; black line = Chilean *he* major clade; grey line = Chilean *f* clade). Isolate NBISA01 belongs to the Canadian clade and it was reported in Chile in 2001.

Figure 3 Bayesian Skyline Plot reconstruction for the *f* and *he* genes. Black and red lines represent median estimates of effective population size, scaled by generation time in years (τ), for the *f* and *he* genes, respectively. Grey and yellow lines represent 95% Highest Probability Density (HPD) for both genes with grey for *f* and yellow for *he*, respectively.

Although controversial, Vike et al. [19] suggested that the source of ISAV in the Chilean outbreak were infected eggs brought from Norway, which could be explained by the fact that until 2008 this country was the main provider of embryonated eggs http://www.sernapesca.cl. It is clear that the Chilean outbreak clade has a sister relationship with Norwegian isolates (closely related to ISAV8 (97/09/615); see Figure 2), that the basal clades and ancestral nodes in the phylogeny are from Norway, and that the tMRCA is around 1989-1996, which matches the date when egg importation grew from less than 20 million to 50 million imported units http://www.sernapesca.cl. In addition, the BSP analysis shows that the ISAV population increased exponentially since 2006, suggesting that the virus population remained constant since 2001 (tMRCA of Chilean outbreak clade) and that the outbreak was publicly reported one year after this predicted population expansion. These results constitute valuable information for fish health officials to make informed future decisions based on what contingency measures were implemented at the time of transmission from Norway to Chile. For instance, it is very interesting to note that in 2011 the main ISAV genotype detected in Chile corresponded to HPR0 http://www.sernapesca.cl. Conversely, HPR7b was nearly 80% prevalent between epidemic years 2007 and 2008 [20]. This could be attributed to the change of source material from Norway to Iceland in 2009. For a future study it would be interesting to address the origin of newly detected avirulent HPR0 strains that would come from a country other than Norway.

Selection analysis and structure models
In order to find potentially relevant sites in the primary sequence of both surface genes and to determine possible structure-function relations, we used a two-rate Fixed-Effects likelihood (two-rate FEL) approach to find positively selected sites (dN/dS rate ratio >1). In a gene-based approach, we found that both genes are evolving under purifying selection (dN/dS rate ratio <1) as expected for fast-evolving entities with relatively small genomes (0.2038 and 0.1382 for *he and f* genes, respectively).

Using a site-by-site approach, we found that the *he* gene has two sites under positive selection within codons in positions 134 and 340 (Figure 1B; numbering according to ISAV752 isolate). In the *he* gene, amino acid Val/Leu$_{134}$ is located specifically on the globular

distal domain (Figure 1A), which is related to HE protein main function, i.e., cell receptor recognition [49]. Amino acid Ser/Arg/Gly/Ala$_{340}$ lies upstream the HPR region, located in the stalk of the protein (Figure 1A), which prompted us to hypothesize that it might be involved in recombination. Interestingly, the expression of recombinant ISAV HEΔ$_{339-351}$ results in decreased expression levels, impaired salmon serum recognition, and hema-adsorption activity. In contrast, HEΔ$_{349-351}$ shows normal parameters [49], suggesting that positively selected amino acid in position 340 is due to its great importance for HE activity.

We also found three sites under positive selection in the *f* gene within codons that code for amino acid residues 130, 170, and 279 (Figure 1F). It was not possible to reliably model the *f* gene protein product structure. However, we were able to map amino acid Ala/Gly/Thr$_{279}$, which lies immediately downstream of insertions (INs) [6]. This constitutes an interesting finding given that insertions in this region have been reported to confer novel PPPS [15], as outlined in Figure 1F. In addition, Ala/Gly/Thr$_{279}$ lies immediately downstream to Arg$_{278}$ (Arg$_{267}$ in ISAV fusion without insertion), a putative cleavage site [17] that has been suggested as a virulence factor [50]. Conversely, a recent study suggested that substitution of Gln/His$_{266}$ to Leu$_{266}$ (two positions upstream to our predicted site) in ISAV fusion protein without insertion is in association with an increase of virulence behavior, as seen in Chilean ISAV 901 strain [13,15]. This has epidemiological implications as isolates carrying these elements would have an advantage compared to those that do not. This finding predicts that the frequency of these alleles will increase in the populations where they are present, e.g., IN4, which is only found in Chile with almost 80% prevalence. It has been suggested that INs close to the cleavage site exposes this region to the solvent probably helping in recognition and processing. Our improved model shows that insertions lie inside the alpha helix structure that is enlarged with subsequent addition of residues, which leads to the consequent exposure of PPPSs (Figure 1D and 1E), in comparison with ISAV901 fusion protein without insertion (Figure 1C). These findings also support the notion that IN's serve as virulence determinants. For instance, the unique insertion, IN4, provides a cleavage site for Asp-N endopeptidase metalloprotease, which has been detected in the skin mucosa of salmonids [15]. Since the HPR in *he* as well as the INs in *f* are associated with virulence and are also under natural selection pressure, we think that experimental confirmation of these results will lead to a better understanding of molecular determinants of pathogenicity for both surface proteins.

Conclusions

In summary, this study investigated the evolutionary history of ISAV, an important economic infection for salmon culture worldwide. Although less exploited in non-human diseases, phylogenetic and phylodynamic analyses characterize pathogens in space and time, which is perfectly suited for surveillance programs. For instance, identification of the origin of an outbreak could aid officials in formulating management policies. Though not pursued in this study, phylogeographic analyses could help identify virus sources and venues from which pathogens are spreading into different populations or fisheries. Reconstructing past population dynamics could help monitor how an infectious agent is changing or reacting to sanitary interventions in terms of its effective population size. Also, examining molecular sequences for positive selection, all in a structural context, can identify determinants of virulence or other molecular adaptations.

In this study we found that ISAV, as its relatives, (i) exhibit a low recombination rate, though recombination is present in the *he* surface gene. We hypothesized that HPR region diversity is governed by recombination probably due to template switching. (ii) We also provided more evidence in favor of the Norwegian origin hypothesis in a coherent phylogenetic and statistical framework. Contrary to the widely accepted view [41], (ii) ISAV HPR0 genotype is not the precursor of all ISAV *he* genotypes, but appears to share a common ancestor with the Chilean outbreak clade. Finally, (iv) we identified potential determinants of fitness as identified by being under positive selection in a structural context; the protein models allowed us to relate the results of phylogenetic studies with the predicted structures, which suggest an important structural function for the predicted positively selected sites. Our analysis shows that the ISAV outbreak remained undetected in Chile since 2001 and that the population started to increase prior to the identification of the outbreak. Thus, we demonstrate the utility of evolutionary analyses in characterizing temporal, spatial, and mechanistic aspects of ISAV outbreaks, as well as their use in predicting future outbreaks of ISAV and other pathogens. Undoubtedly, collaboration between basic experimental research and phylogenetic fields will help understand the dynamics of ISAV spread in space and time as it has proven fruitful in human health and emerging infectious diseases studies.

Methods
Sequences and alignment
For this study, we focused on the two most abundant genes available, which also have been linked to virulence and used extensively for genotyping, namely the Fusion (*f*) and Hemagglutinin-esterase (*he*) genes. Sequences ranging

from 1989-2009 were obtained from GenBank according to several criteria. First, in order to implement a molecular clock, we selected only full-length sequences with reported isolation dates. Second, we only used isolates for which both target genes were sequenced. After these criteria were applied, seventy full-length sequences from each gene were retrieved from GenBank (accession numbers in additional file 3) from strains isolated in Canada, Norway, and Chile (alignment in additional file 1). To avoid alignment artifacts and to come up with an optimal statement of homology, nucleotide sequences were visualized in Seaview 4.2.7 [51] and then converted to amino acid sequences for alignment with MAFFT [52] under L-INS-i algorithm, thereby preserving the open reading frame for these protein coding sequences. Amino acid sequences were then back translated into their original nucleotide sequences for further analyses.

Genetic diversity and recombination

Genetic diversity (θ) within populations was estimated both by the traditional segregating sites approach of Watterson [53] (implemented in DnaSP5 [54]) and by a coalescent Bayesian Markov Chain Monte Carlo method (BMCMC [55]; implemented in Lamarc 2.1.3 [56]). Recombination breakpoint analysis was performed under a suit of methods looking for different types of recombination evidence (Rdp, Genconv, MaxChi, Chimaera as primary methods and Bootscanning and Siscan as secondary (summarized in [57]) as implemented in Rdp3 [39]. Windows and step sizes were varied to refine primary evidence of recombination as suggested in the user manual. In addition, recombination detection was performed on both genes by using a genetic algorithm approach looking at topological incongruence (GARD) [58] as implemented in the http://datamonkey.org[59] server.

Phylogenetic inference, molecular clock, substitution rates and demographic history

The best-fit model of evolution was selected under Akaike information Criterion [27] (AIC) and parameter values using model averaging as implemented in jModelTest 1.0.1 [60]. The best-fit model was used for subsequent analyses; nevertheless, base frequencies and rates were co-estimated with the phylogeny. A phylogenetic tree was inferred using Bayesian Inference (BI) as optimality criterion as implemented in the Beast 1.6.1 package [61]. Posterior probabilities (PP) were approximated by running four independent analyses of 1×10^8 steps sampled every 1×10^5 each. A burn-in period was defined using Tracer [57] by testing for convergence on likelihood scores. Burn-in was defined as the period before convergence and those runs were discarded accordingly. Also, we assessed convergence and mixing by checking the effective sample size statistic (ESS > 500)

and the trace itself. Strict molecular clock was rejected on the basis of Bayes Factor hypothesis testing between uncorrelated lognormal clock (H1) and strict clock (H0) as implemented in Tracer [62]. A lognormal prior with a mean centered on the reported Influenza A HA gene substitution rate [26] and tip dates were used to calibrate the clock. The same analysis was run 'on empty' to assess the influence of the selected priors. Estimating Association Index (AI), Parsimony Score (PS), and Maximum Clade number (MC) statistics tested the extent of geographic structure in BaTS [63]. Past population dynamics of both genes were inferred using the Bayesian Skyline Plot (BSP) model as a tree prior [45]. Ancestral State Reconstruction (ASR) was estimated used the method of Lemey et al. [64] implemented in BEAST as described online http://beast.bio.ed.ac.uk/Tutorials#Phylogeography_tutorials

Selection analyses and structure models

Both genes, *he* and *f*, were investigated for evidence of positive selection. First, a global analysis was performed using a counting method (SLAC). Then, a site-by-site analysis was performed using Fixed-Effects likelihood method as implemented in HyPhy 2.0 [65,66]. Differences in selective pressures between the Chilean clade and the rest of the tree were estimated with SelectionLRT.bf script also within the HyPhy 2.0 package.

To assess whether selected sites lay on structural and functionally relevant regions of proteins, *f* and *he* 3D structure models were build using the following protocol. The sequences of the *he* and the *f* genes were obtained from GenBank (ADF36510.1; ABG65799.1; AAX46277.1; AAX46259.1; ADF36509.1 and ADF36499.1). Alignments for HE protein were obtained by modeler 9v9 [67], and for F the server CLUSTALW was used [68]. To verify that the alignments were suitable, we used the secondary structure prediction server PSIPRED http://bioinf.cs.ucl.ac.uk/psipred/[69]. In order to generate the F protein model, we used two templates: hemagglutinin-esterase-fusion glycoprotein structure of influenza C virus (PDB ID: 1FLC) (3.20Å) and hemagglutinin-fusion glycoprotein structure of influenza A virus (PDB ID: 3EYJ) (2.60 Å), with identities of 21% and 20%, respectively. For the HE protein model, we used hemagglutinin-esterase fusion glycoprotein structure of influenza C virus (PDB ID: 1FLC) with 19% identity as template. Finally, both models were minimized (50,000 steps) and a trajectory of 1 ns was obtained by molecular dynamics using the CHARMM force field [70] in NAMD program [71]. The quality of the models was tested by the Anolea server [72].

Additional material

Additional file 1: Nucleotide alignments. Datasets used in this study. File is formatted as a nexus file, which can be read by most sequence

editor software. Taxa names are designated according to isolate name and geographic location. The *f* gene sequence goes from nucleotide 1 to 1368 and *he* gene sequence from 1369 to 2640. A MrBayes block with a character set is provided at the end of the file that can be read by BEAST.

Additional file 2: Hemagglutinin-Esterase Maximum clade credibility (MCC) phylogeny. Phylogenetic tree representing evolutionary relationships between the *he* genes of selected isolates. Colors represent country of origin and inferred ancestral states as in Figure 2.

Additional file 3: Sequence information and accession numbers. Excel readable spreadsheet listing isolate names, their country of origin, year of isolation, and their GenBank accession numbers.

Author details
[1]Department of Biology, 401 Widtsoe Building, Brigham Young University, Provo, UT 84602-5181, USA. [2]Centro de Biotecnología Acuícola, Laboratorio de Virología, Facultad de Química y Biología, Universidad de Santiago de Chile, Avenida Libertador Bernardo O'Higgins 3363 Santiago, Chile. [3]Facultad de Química y Biología, Universidad de Santiago de Chile, Avenida Libertador Bernardo O'Higgins 3363 Santiago, Chile.

Authors' contributions
ECN conceived the study, obtained and analyzed the molecular data and wrote the manuscript. MCS, C. Ma and C. Mo developed protein models, analyzed data and helped writing the manuscript. KAC conceived the study, advised on analyses, outlined, revised and edited the manuscript. All authors agreed with the final version of the manuscript.

References
1. Asche F, Hansen H, Tveteras R, Tveteras S: **The Salmon Disease Crisis in Chile.** *Mar Resour Econ* 2009, **24(4)**:405-411.
2. Bouchard D, Keleher W, Opitz HM, Blake S, Edwards KC, Nicholson BL: **Isolation of infectious salmon anemia virus (ISAV) from Atlantic salmon in New Brunswick, Canada.** *Dis Aquat Organ* 1999, **35(2)**:131-137.
3. Geoghegan F: **First isolation and identification of ISAV in Ireland.** *6th Annual Meeting of EU National Reference Laboratories for Fish Diseases. Brussels, Belgium* 2002.
4. Kibenge FSB, Munir K, Kibenge MJT, Joseph T, Moneke E: **Infectious salmon anemia virus: causative agent, pathogenesis and immunity.** *Anim Health Res Rev* 2004, **5(1)**:65-78.
5. Lovely J, Dannevig B, Falk K, Hutchin L, MacKinnon A, Melville K, Rimstad E, Griffiths S: **First identification of infectious salmon anaemia virus in North America with haemorrhagic kidney syndrome.** *Dis Aquat Organ* 1999, **35(2)**:145-148.
6. Godoy M, Aedo A, Kibenge M, Groman D, Yason C, Grothusen H, Lisperguer A, Calbucura M, Avendano F, Imilan M, *et al*: **First detection, isolation and molecular characterization of infectious salmon anaemia virus associated with clinical disease in farmed Atlantic salmon (Salmo salar) in Chile.** *BMC Vet Res* 2008, **4(4)**:28.
7. Kawaoka Y, Cox N, Haller O, Hongo S, Kaverin N, Klenk H, Lamb R, McCauley J, Palese P, Rimstad E, *et al*: **Infectious Salmon Anaemia Virus - Eight Report of the International Committee on Taxonomy Viruses.** New York: Elsevier Academic Press; 2005.
8. Cottet L, Rivas-Aravena A, Cortez-San Martin M, Sandino AM, Spencer E: **Infectious salmon anemia virus–genetics and pathogenesis.** *Virus Res* 2010, **155(1)**:10-19.
9. García-Rosado E, Markussen T, Kileng O, Baekkevold ES, Robertsen B, Mjaaland S, Rimstad E: **Molecular and functional characterization of two infectious salmon anaemia virus (ISAV) proteins with type I interferon antagonizing activity.** *Virus Res* 2008, **133(2)**:228-238.
10. Rimstad E, Mjaaland S: **Infectious salmon anaemia virus. An orthomyxovirus causing en emerging infection in Atlantic salmon.** *APMIS* 2002, **110(4)**:273-282.
11. Kibenge FS, Xu H, Kibenge M, Qian B, Joseph T: **Characterization of gene expression on genomic segment 7 of infectious salmon anaemia virus.** *Virology J* 2007, **4**:34-48.
12. Mjaaland S, Hungnes O, Teig A, Dannevig BH, Thorud K, Rimstad E: **Polymorphism in the infectious salmon anemia virus hemagglutinin gene: importance and possible implications for evolution and ecology of infectious salmon anemia disease.** *Virology* 2002, **304(2)**:379-391.
13. Markussen T, Jonassen CM, Numanovic S, Braaen S, Hjortaas M, Nilsen H, Mjaaland S: **Evolutionary mechanisms involved in the virulence of infectious salmon anaemia virus (ISAV), a piscine orthomyxovirus.** *Virology* 2008, **374(2)**:515-527.
14. Kibenge FSB, Kibenge MJT, Wang Y, Qian B, Hariharan S, Sandi : **Mapping of putative virulence motifs on infectious salmon anemia virus surface glycoprotein genes.** *J Gen Virol* 2007, **88(11)**:3100-3111.
15. Cottet L, Cortez-San Martín M, Tello M, Olivares E, Rivas-Aravena A, Vallejos E, Sandino AM, Spencer E: **Bioinformatic Analysis of the Genome of Infectious Salmon Anemia Viruses associated with outbreaks of high mortality in Chile.** *J Virol* 2010, **84(22)**:11916-11928.
16. Nylund A, Devold M, Karlsen M: **Sequence analysis of the fusion protein gene from infectious salmon anemia virus isolates: evidence of recombination and reassortment.** *J Gen Virol* 2006, **87(Pt 7)**:2031-2040.
17. Aspehaug V, Mikalsen AB, Snow M, Biering E, Villoing S: **Characterization of the infectious salmon anemia virus fusion protein.** *J Virol* 2005, **79(19)**:12544-12553.
18. Kibenge FSB, Grate ON, Johnson G, Arriagada R, Kibenge MJT, Wadowska D: **Isolation and identification of infectious salmon anaemia virus (ISAV) from Coho salmon in Chile.** *Dis Aquat Organ* 2001, **45(1)**:9-18.
19. Vike S, Nylund S, Nylund A: **ISA virus in Chile: evidence of vertical transmission.** *Arch Virol* 2009, **154(1)**:1-8.
20. Kibenge F, Godoy M, Wang Y, Kibenge M, Gherardelli V, Mansilla S, Lisperger A, Jarpa M, Larroquete G, Avendano F, *et al*: **Infectious salmon anaemia virus (ISAV) isolated from the ISA disease outbreaks in Chile diverged from ISAV isolates from Norway around 1996 and was disseminated around 2005, based on surface glycoprotein gene sequences.** *Virology J* 2009, **6**:88.
21. Elena SF, Bedhomme S, Carrasco P, Cuevas JM, de la Iglesia F, Lafforgue G, Lalić J, Pròsper A, Tromas N, Zwart MP: **The evolutionary genetics of emerging plant RNA viruses.** *Mol Plant Microbe Interact* 2011, **24(3)**:287-293.
22. Castao A, Ruiz L, Elena SF, Hernández C: **Population differentiation and selective constraints in Pelargonium line pattern virus.** *Virus Research* 2011, **155(1)**:274-282.
23. Wu B, Alexandra , Liu Y, Zhou G, Wang X, Elena SF: **Dynamics of molecular evolution and phylogeography of Barley yellow dwarf virus-PAV.** *PloS One* 2011, **6(2)**:e16896-e16896.
24. Lalić J, Agudelo-Romero P, Carrasco P, Elena SF: **Adaptation of tobacco etch potyvirus to a susceptible ecotype of Arabidopsis thaliana capacitates it for systemic infection of resistant ecotypes.** *Philos Trans R Soc Lond B Biol Sci* 2010, **365(1548)**:1997-2007.
25. Bhatt S, Holmes EC, Pybus OG: **The genomic rate of molecular adaptation of the human influenza A virus.** *Mol Biol Evol* 2011, **28(9)**:2443-2451.
26. Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC: **The genomic and epidemiological dynamics of human influenza A virus.** *Nature* 2008, **453(7195)**:615-619.
27. Akaike H: **A new look at the statistical model identification.** *IEEE Trans Automat Contr* 1974, **19(6)**:716-723.
28. Felsenstein J: **Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates.** *Genet Res* 1992, **59(2)**:139-147.
29. Hughes GJ, Fearnhill E, Dunn D, Lycett SJ, Rambaut A, Leigh Brown AJ, UK HIV Drug Resistance Collaboration: **Molecular Phylodynamics of the**

Heterosexual HIV Epidemic in the United Kingdom. *Plos Pathog* 2009, **5**(9):e1000590.

30. Bennett SN, Drummond AJ, Kapan DD, Suchard MA, Munoz-Jordan JL, Pybus OG, Holmes EC, Gubler DJ: **Epidemic Dynamics Revealed in Dengue Evolution.** *Mol Biol Evol* 2010, **27**(4):811-818.

31. Tazi L, Pérez-Losada M, Gu W, Yang Y, Xue L, Crandall K, Viscidi R: **Population dynamics of Neisseria gonorrhoeae in Shanghai, China: a comparative study.** *BMC Infectious Diseases* 2010, **10**(1):13.

32. Zehender G, De Maddalena C, Giambelli C, Milazzo L, Schiavini M, Bruno R, Tanzi E, Galli M: **Different evolutionary rates and epidemic growth of hepatitis B virus genotypes A and D.** *Virology* 2008, **380**(1):84-90.

33. Holmes EC: **The Evolutionary Genetics of Emerging Viruses.** *Annu Rev Ecol Evol S* 2009, **40**:353-372.

34. Posada D, Crandall KA, Holmes EC: **Recombination in evolutionary genomics.** *Annu Rev Genet* 2002, **36**:75-97.

35. Boni MF, de Jong MD, van Doorn HR, Holmes EC: **Guidelines for Identifying Homologous Recombination Events in Influenza A Virus.** *PLoS One* 2010, **5**(5):e10434.

36. Boni MF, Zhou Y, Taubenberger JK, Holmes EC: **Homologous Recombination Is Very Rare or Absent in Human Influenza A Virus.** *J Virol* 2008, **82**(10):4807-4811.

37. He CQ, Han GZ, Wang D, Liu W, Li GR, Liu XP, Ding NZ: **Homologous recombination evidence in human and swine influenza A viruses.** *Virology* 2008, **380**(1):12--20.

38. Kosakovsky Pond SL, Posada D, Stawiski E, Chappey C, Poon AFY, Hughes G, Fearnhill E, Gravenor MB, Leigh Brown AJ, Frost SDW: **An Evolutionary Model-Based Algorithm for Accurate Phylogenetic Breakpoint Mapping and Subtype Prediction in HIV-1.** *PLoS Comput Biol* 2009, **5**(11):e1000581.

39. Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P: **RDP3: a flexible and fast computer program for analyzing recombination.** *Bioinformatics* 2010, **26**(19):2462-2463.

40. Devold M, Falk K, Dale B, Krossøy B, Biering E, Aspehaug V, Nilsen F, Nylund A: **Strain variation, based on the hemagglutinin gene, in Norwegian ISA virus isolates collected from 1987 to 2001: indications of recombination.** *Dis Aquat Organ* 2001, **47**(2):119-128.

41. Cunningham CO, Gregory A, Black J, Simpson I, Raynard RS: **A novel variant of the infectious salmon anaemia virus (ISAV) haemagglutinin gene suggests mechanisms for virus diversity.** *Bull Eur Assoc Fish Pathol* 2002, **22**(6):366-374.

42. Posada D: **Unveiling the molecular clock in the presence of recombination.** *Mol Biol Evol* 2001, **18**(10):1976-1978.

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

44. Domingo-Calap P, Sanjuán R: **Experimental Evolution of RNA versus DNA Viruses.** *Evolution* 2011, **65**(10):2987-2994.

45. Drummond AJ, Rambaut A, Shapiro B, Pybus O: **Bayesian coalescent inference of past population dynamics from molecular sequences.** *Mol Biol Evol* 2005, **22**(5):1185-1192.

46. Ho S, Shapiro B: **Skyline-plot methods for estimating demographic history from nucleotide sequences.** *Mol Ecol Resour* 2011, **11**(3):423-434.

47. Pannell JR: **Coalescence in a Metapopulation with Recurrent Local Extinction and Recolonization.** *Evolution* 2003, **57**(5):949-961.

48. Duffy S, Shackelton LA, Holmes EC: **Rates of evolutionary change in viruses: patterns and determinants.** *Nat Rev Genet* 2008, **9**(4):267-276.

49. Mikalsen AB, Sindre H, Mjaaland S, Rimstad E: **Expression, antigenicity and studies on cell receptor binding of the hemagglutinin of infectious salmon anemia virus.** *Arch Virol* 2005, **150**(8):1621-1637.

50. Kibenge FS, Kibenge MJ, Wang Y, Qian B, Hariharan S, McGeachy S: **Mapping of putative virulence motifs on infectious salmon anemia virus surface glycoprotein genes.** *J Gen Virol* 2007, **88**(Pt 11):3100-3111.

51. Gouy M, Guindon S, Gascuel O: **SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building.** *Mol Biol Evol* 2010, **27**(2):221-224.

52. Katoh K, Kuma K, Toh H, Miyata T: **MAFFT version 5: improvement in accuracy of multiple sequence alignment.** *Nucleic Acids Res* 2005, **33**(2):511-518.

53. Watterson GA: **On the number of segregating sites in genetical models without recombination.** *Theor Popul Biol* 1975, **7**(2):256-276.

54. Librado P, Rozas J: **DnaSP v5: a software for comprehensive analysis of DNA polymorphism data.** *Bioinformatics* 2009, **25**(11):1451-1452.

55. Kuhner MK, Yamato J, Felsenstein J: **Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling.** *Genetics* 1995, **140**(4):1421-1430.

56. Kuhner MK: **LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters.** *Bioinformatics* 2006, **22**(6):768-770.

57. Posada D, Crandall KA: **Evaluation of methods for detecting recombination from DNA sequences: Computer simulations.** *Proc Natl Acad Sci USA* 2001, **98**(24):13757-13762.

58. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD: **GARD: a genetic algorithm for recombination detection.** *Bioinformatics* 2006, **22**(24):3096-3098.

59. Delport W, Poon AFY, Frost SDW, Kosakovsky Pond SL: **Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology.** *Bioinformatics* 2010, **26**(19):2455-2457.

60. Posada D: **jModelTest: Phylogenetic Model Averaging.** *Mol Biol Evol* 2008, **25**(7):1253-1256.

61. Drummond AJ, Rambaut A: **BEAST: Bayesian evolutionary analysis by sampling trees.** *BMC Evol Biol* 2007, **7**(1):214.

62. Suchard MA, Weiss RE, Sinsheimer JS: **Bayesian Selection of Continuous-Time Markov Chain Evolutionary Models.** *J Mol Evol* 2001, **18**(6):1001-1013.

63. Parker J, Rambaut A, Pybus OG: **Correlating viral phenotypes with phylogeny: Accounting for phylogenetic uncertainty.** *Infect Genet Evol* 2008, **8**(3):239-246.

64. Lemey P, Rambaut A, Drummond AJ, Suchard MA: **Bayesian Phylogeography Finds Its Roots.** *PLoS Comput Biol* 2009, **5**(9).

65. Kosakovsky Pond SL, Frost SDW: **Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection.** *Mol Biol Evol* 2005, **22**(5):1208-1222.

66. Kosakovsky Pond SL, Frost SDW, Muse SV: **HyPhy: hypothesis testing using phylogenies.** *Bioinformatics* 2004, **21**(5):676-679.

67. Sali A, Blundell TL: **Comparative Protein Modelling by Satisfaction of Spatial Restraints.** *J Mol Biol* 1993, **234**(3):779-815.

68. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22**(22):4673-4680.

69. Jones DT: **Protein secondary structure prediction based on position-specific scoring matrices.** *J Mol Biol* 1999, **292**(2):195-202.

70. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M: **CHARMM: A program for macromolecular energy, minimization, and dynamics calculations.** *J Comput Chem* 1983, **4**(2):187-217.

71. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kalé L, Schulten K: **Scalable molecular dynamics with NAMD.** *J Comput Chem* 2005, **26**(16):1781-1802.

72. Guex N, Peitsch MC: **SWISS-MODEL and the Swiss-Pdb Viewer: An environment for comparative protein modeling.** *Electrophoresis* 1997, **18**(15):2714-2723.