Molecular evolution and phylogeography of infectious hematopoietic necrosis virus with a focus on its presence in France over the last 30 years

Laure Bellec,1,2 Lénàïg Louboutin,3 Joëlle Cabon,3 Jeanne Castric,3 Joëlle Cozien,4 Richard Thiéry5 and Thierry Morin3,*

Abstract
Infectious hematopoietic necrosis virus (IHNV) is among the most important pathogens affecting the salmonid industry. Here, we investigated the molecular evolution and circulation of isolates from 11 countries or regions all over the world, with a special focus on the epidemiological situation in France. The phylogeography, time to the most recent common ancestor (TMRCA) and nucleotide substitution rate were studied using 118 full-length glycoprotein gene sequences isolated from 9 countries (5 genogroups) over a period of 47 years. The TMRCA dates back to 1943, with the L genogroup identified as the likely root (67%), which is consistent with the first report of this pathogen in the USA. A Bayesian inference approach was applied to the partial glycoprotein gene sequences of 88 representative strains isolated in France over the period 1987–2015. The genetic diversity of these 88 sequences showed mean nucleotide and amino-acid identities of 97.1 and 97.8%, respectively, and a dN/dS ratio (non-synonymous to synonymous mutations) of 0.25, indicating purifying selection. The French viral populations are divided into eight sub-clades and four individual isolates, with a clear spatial differentiation, suggesting the predominant role of local reservoirs in contamination. The atypical ‘signatures’ of some isolates underlined the usefulness of molecular phylogeny for epidemiological investigations that track the spread of IHNV.

INTRODUCTION
Belonging to the Salmonid novirhabdovirus species of the genus Novirhabdovirus, infectious hematopoietic necrosis virus (IHNV) is the causative agent of a severe aquatic disease affecting wild and farmed salmonid species [1–5]. Structurally, the typically bullet-shaped virion encapsidates a non-segmented, negative-sense, single-stranded RNA of about 11 000 nucleotides [6]. The linear genome encodes six proteins in the following order (3’–5’): a nucleoprotein (N), a phosphoprotein (P), a matrix protein (M), a glycoprotein (G), a non-virion protein (NV) and a polymerase (L) [7–10].

The first outbreak of IHNV was detected in hatcheries in western North America in the 1950s [11]. Since then, the aetiological agent has spread through North America, Asia and Europe – mostly due to the international trade of juvenile fish or eggs – and has become a major threat to the aquaculture industry. The detection of IHNV in Japan dates from 1971 and the virus was probably imported with a shipment of contaminated fish eggs from Alaska, USA [12]. In Europe, the virus was reported for the first time in 1987 from two independent cases, one in France and one in Italy, and later in Germany (1992) [13–15]. Many outbreaks have since been reported around the world [16, 17] and phylogenetic analyses based on the complete or partial sequence of the G gene, which shows relatively high genetic diversity compared with the other genes, have led to the definition of five major genogroups: U (upper), M (middle), L (lower), E (Europe) and J (Japanese rainbow trout) [10, 18]. U, M and L correspond to the observed geographical range in North America [19–21]. Several diverse genotypes have been identified in Europe and seem closely related to the
M genogroup [22]. The Japanese isolates show high divergence, but share a common origin with genogroup U [23, 24].

Due to its high infectivity and wide distribution, IHNV represents a serious economic impact in aquaculture species such as rainbow trout (*Oncorhynchus mykiss*) and Atlantic salmon (*Salmo salar*). The associated disease is therefore listed as one of the regulated non-exotic fish diseases in the European Union [25] and is notifiable to the World Organization for Animal Health (OIE) [26]. Historically based on virus isolation in susceptible cell cultures followed by confirmation steps, the direct official methods for surveillance and diagnosis now also include RT-qPCR detection [27, 28]. The extensive use of these methods for health monitoring and veterinary controls of international fish trade have significantly reduced the number of incidents associated with IHNV in Europe in the past few years. Recent surveillance data indicate that about 0.6 % of the listed European fish farms are considered infected [29], but this number is probably underestimated, particularly because it is difficult to detect the virus in fish that are not in the most susceptible stage (juvenile) and/or in permissive conditions [30].

Here, we applied a Bayesian coalescent method to the five genogroups of IHNV to better understand the evolutionary history and phylogeography of this pathogen. In addition, the genetic diversity and geographical distribution of 88 strains isolated in France over a period of 30 years were analysed for the first time and used to develop hypotheses for the different possible routes of viral spread among trout farms.

**RESULTS**

**Worldwide phylogeography of IHNV**

A discrete phylogeographical analysis of IHNV was conducted on 118 complete G gene sequences from 11 locations (China, Croatia, France, Germany, Italy, Japan, South Korea, Switzerland, USA lower, USA middle and USA upper) and covered all recognized genogroups: J, E, U, M and L (Fig. 1). The distribution clearly suggests that the USA lower specimens constitute the root of the tree with over 67 % probability. The topology defined three clear genogroups (L, E and M) and two others (U and J) that were more ambiguous. For example, U showed a basal position compared with J, but they clustered together. Genogroup J was composed of isolates from Japan, South Korea and China. Isolates from South Korea formed two groups with one Japanese isolate at a basal position each time. Recent (2012–2013) Chinese isolates constituted one group and appeared to emerge from Japan. The European group derived from the M genogroup, with a probable German origin that gave rise to separate infections in Italy, Switzerland, Croatia and France. Several introduction events involving French-related viruses in Germany or Switzerland were also observed.

The nucleotide substitution rate (per site and per year) for the G gene was estimated to be $7.14 \times 10^{-4}$ substitutions (subs) site$^{-1}$ year$^{-1}$ and the emergence of IHNV was estimated to date from 1943 (Table 1). For a more accurate analysis, the most recent common ancestor (TMRCA) and mean genetic diversity were also calculated for the five genogroups (Table 2). The oldest genogroup was the L genogroup and the youngest genogroup appeared to be the E genogroup, whose divergence dates back to 1981.

**Phylogenetic analysis of French isolates**

We sequenced a region of the G gene from 88 representative French IHNV isolates collected by our National Reference Laboratory (NRL) and covering a 30-year period (1987–2015) (Fig. 2). A dataset of a partial G-gene sequence of 570 bp was created, which allowed us to group the 88 French isolates plus 5 outgroups (from the M genogroup), and was used for Bayesian phylogenetic reconstruction (Fig. 3). The structures of the French isolates in the phylogeny show eight sub-clades (six with posterior support $>0.75$) and four individual isolates. Two main monophyletic groups, both including 23 French viruses (sub-clades A and E) and defined by posterior support of 0.75 and 1, respectively, were observed. The other genetic sub-clades (B, C, D, F, G and H) were smaller and composed of 3 to 15 sequences. Interestingly, isolates corresponding to the first French outbreak (71_87_1987) and the most recent one (SA15.1442_2015) did not cluster with the other sequences. The temporal link within the sub-clades is not obvious, for example the two most abundant sub-clades, A and E, have a large range of isolation dates, with 1999–2014 and 1990–2005, respectively. However, we observed that viruses were isolated from three distinct areas in France: the north (orange), east (purple) and centre/south (blue) (Fig. 3). All of the viruses from the east area belong to sub-clade A, with one exception (isolate 016650_2007), suggesting that all the viruses from the east are derived and spread from a single common infection. The IHNV isolates from the centre/south form three different sub-clades (B, C and H). The isolates from the north seem to be the more divergent, with four sub-clades (D, E, F and G) plus four individual isolates representing multiple independent introductions or the parallel evolution of separate viral lineages. However, some strains had genetic profiles that were not related to the geographical distribution of the strains isolated in the respective area/time period. For example, isolate 503_2000 from the north only showed 97.8 % nucleotide identity with the two other strains from the same area, but had 100 % identity with three isolates (Y6_2000, J13325_1999 and K2165_2000) from the east group. In another example, isolate 016650_2007, identified in the east area, near the French–Swiss border, clustered with the north group, sharing 100 % nucleotide identity with an isolate (LDO1n_2007) from this group, despite the groups being separated by a distance of more than 600 km. Similarly, isolate M13316_2002, isolated from fish in the south/centre area, was 100 % identical to a strain (L5889_2001) isolated 1 year before in the east area.

The nucleotide sequence identity ranged from 95.4 to 100 %, with a mean value of 98.1 %, for all isolates considered (Table 3).
Genetic diversity

The worldwide sequences of the infectious hematopoietic necrosis virus displayed low genetic diversity for the G gene, with a mean nucleotide diversity of 3.4%; within the five genogroups, the diversity reached a maximum of 3.5% and a minimum of 1.3% (Table 2). Nevertheless, the nucleotide diversity between the five genogroups showed that the Japanese group was the most distant, with its diversity ranging from 4.2 to 5.6% with respect to each of the four other groups. The genetic diversity, \(d_\text{ys}/d_\text{S}\) ratio (non-synonymous to synonymous mutations) and Tajima’s \(D\) (test based on polymorphism frequencies) were calculated for our 88 French isolates (Table 3). The partial G region displayed more synonymous than non-synonymous changes (ratio <1), suggesting purifying selection on the viral population. Tajima’s test was negative, indicating negative selection on the viral population.

DISCUSSION

Infectious hematopoietic necrosis (IHN) is one of the most economically important viral diseases in farmed rainbow trout in Europe, and in most parts of the northern hemisphere. Here, we performed a discrete phylogeographical analysis on 118 sequences from the five genogroups of IHNV using BEAST2. Further, we conducted, for the first time, an extensive study on the genetic diversity and geographical distribution of 88 strains collected by our laboratory, the French NRL. These strains are representative of all outbreaks reported since the first description of the disease in rainbow trout farms in 1987 up until 2015. Our results (i) confirm the origin and the history of spread of this worldwide virus; (ii) demonstrate a correlation between molecular phylogeny and geographical distribution in France; and (iii) illustrate how phylogeny can contribute to epidemiological investigations.
During the last six decades, extensive phylogenetic studies on IHNV have led to the identification of five genogroups that clearly correlate with specific geographical origins [10]. Our analysis of the evolutionary history (phylogeography and date of divergence) of these genogroups confirms that the L genogroup is located at the putative root (origin) of the tree, while the U and M diverged from this group early on. The finding that the L genogroup is at the origin of the phylogenetic evolution is consistent with the first reported outbreaks in hatcheries in the state of Washington, USA in the early 1950s [11] and with previous studies on IHNV [18, 20, 22]. A second significant divergence event involves the derivation of the J genogroup from the U genogroup, presumably in Japan via an import from Alaska, USA in the mid-1970s. Since then, Japanese strains have probably been introduced into Korea [31, 32] and China [33]. Infections with isolates from the U genogroup in Japan were also observed. The youngest genogroup, the E genogroup (1981), derived from genogroup M and spread in different European countries [21, 22, 34]. It was suggested that all European isolates were derived from the first introductions of the virus in France and Italy in 1987 [22]. Our analysis suggests that Germany could also be a potential initial source of introduction, despite later detection (Fig. 1). The first representative outbreaks in France is clearly associated with an individual sequence (71_87_1987 in Fig. 3), as was previously shown in a phylogenetic tree of *Salmonid novirhabdovirus* in Europe [21]. The first sequence in Italy has a basal position belonging to a large clade (IO-87_FJ711518, Fig. 1). This finding was observed previously by Enzmann and co-authors [21], and confirmed by a recent study on Italian IHNV isolates [34]. Although the origin of genogroup E remains to be identified through further investigations, IHNV circulation within the European Union appears very clearly and demonstrates the difficulty of controlling the spread of this virus.

Significant efforts for health monitoring and veterinary control of the spread of IHNV and viral haemorrhagic septicemia (VHS) have been undertaken in recent years in Europe [25, 28, 35]. Overall, they have reduced the number of outbreaks and have helped improve the availability of molecular characterization data for these viruses. The high capability of IHNV to circulate in continental Europe can be assumed from the phylogeny and topology of the E genogroup, in which there is some evidence for a spread starting from German and Italian isolates. The virus has been demonstrated in previously IHN-free countries, such as Croatia or the Netherlands, but to date it has never been detected in the UK, Ireland or Scandinavia [21]. It still represents a serious threat to the large salmon-producing countries of northern Europe [17]. This latent dissemination may be due to the difficulty of observing clinical signs during veterinary controls, particularly if the more susceptible juvenile stages are protected in farms by careful containment from the subadult and adult stages, in which infection is generally latent and silent [30, 36, 37]. Subclinical infections are also associated with lower virus levels in affected fish compared to fish undergoing clinical infection and can lead to false negative diagnosis results [38]. This ability of the disease to go unnoticed (compared with VHS for example) probably leads to underestimation of the number of infected farms in Europe. Violations of European regulations can also not be noticed (compared with VHS for example) probably leads to underestimation of the number of infected farms in Europe.

In France, phylogenetic analysis of the partial G gene sequences of representative isolates collected over the past 30 years is in line with the presumed date of introduction [13]. The observed nucleotide and amino-acid diversities

### Table 1. Details for the dataset and estimates related to the IHNV glycoprotein (G) gene

<table>
<thead>
<tr>
<th>Parameter</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence length (nt)</td>
<td>1518</td>
</tr>
<tr>
<td>No. of sequences</td>
<td>118</td>
</tr>
<tr>
<td>Time span</td>
<td>1966–2013</td>
</tr>
<tr>
<td>Mean substitution rate, $10^{-4}$ site $^{-1}$ year $^{-1}$ (95 % HPD)</td>
<td>7.14 (5.2–9.13)</td>
</tr>
<tr>
<td>TMRCA, years (95 % HPD)</td>
<td>1943 (1918–1961)</td>
</tr>
<tr>
<td>Coefficient of variation (95 % HPD)</td>
<td>0.94 (0.84–1.04)</td>
</tr>
<tr>
<td>Mean genetic diversity (%)</td>
<td>3.4</td>
</tr>
</tbody>
</table>

HPD, highest probability density; TMRCA, time to the most recent common ancestor.

### Table 2. Estimates for the diversity and divergence for the glycoprotein (G) according to the five genogroups of IHNV

<table>
<thead>
<tr>
<th>Genogroup</th>
<th>No. of sequences</th>
<th>TMRCA, years (95 % HPD)</th>
<th>Mean genetic diversity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>5</td>
<td>1954 (1941–1966)</td>
<td>1.4</td>
</tr>
<tr>
<td>U</td>
<td>11</td>
<td>1965 (1957–1970)</td>
<td>1.3</td>
</tr>
<tr>
<td>M</td>
<td>7</td>
<td>1970 (1962–1977)</td>
<td>2.5</td>
</tr>
<tr>
<td>E</td>
<td>74</td>
<td>1981 (1976–1985)</td>
<td>1.4</td>
</tr>
</tbody>
</table>

HPD, highest probability density; TMRCA, time to the most recent common ancestor.
are low and correspond to a purifying selection process stabilizing the viral population [34]. The most striking finding was the clear spatial distribution of genetically related isolates in three areas. This pattern of outbreaks, corresponding to the first diffusive patterns for IHNV spread in Italy described by Abbadi and colleagues, suggests the existence of local reservoirs around which recontaminations regularly occur through passive diffusion via water, fomites (landing nets, boots and vehicles) or piscivorous birds [34, 39, 40]. In this context of relative genetic homogeneity, the detection of isolates with a divergent genetic 'signature' strongly suggests a transfer of isolates from other areas, with diffusion mainly being associated with fish movement and trade practices [34]. In all cases, genetic data analysis provides information of interest to guide epidemiological investigations and take the most appropriate health measures.

Our estimate of the evolutionary dynamics of IHNV is within the range of previously published data (7.14 × 10⁻⁴ subs/site/year in our study; 8.1 × 10⁻⁴ subs/site/year⁻¹ in [18]; 12 × 10⁻⁴ subs/site/year⁻¹ in [41]; and 11 × 10⁻⁴ subs/site/year⁻¹ in [34]) and in agreement with RNA viruses that evolve quickly with a rate between 10⁻² and 10⁻⁵ subs/site/year⁻¹ [42, 43]. Our results on the G gene are also consistent with findings for other fish viruses, such as the Anguillid rhabdovirus, with a rate of 4.23 × 10⁻⁴ subs/site/year⁻¹ [44], VHSV with 5.91 × 10⁻⁴ subs/site/year⁻¹ [45], or the spring viraemia of carp virus, with 5.47 × 10⁻⁴ subs/site/year⁻¹ [46].
Host switching is an important parameter for IHNV evolutionary dynamics. The J and M genogroups of IHNV showed higher genetic diversity compared with the three other genogroups (L, U and E). The lower viral genetic diversity observed for strains from genogroups U and L, mostly found in Pacific salmon, is indicative of evolutionary equilibrium. For the J genogroup, we confirm hypotheses that the host switch [from Pacific salmon (genogroup U) to rainbow trout] is associated with virus adaptation through an increase in genetic diversity. Evolution in the G gene of isolates from farmed rainbow trout has been reported to be up to sixfold higher than that from salmonid fish having an ocean migration phase [20, 47, 48].

This study brings new elements to the understanding of evolutionary history of IHNV. In France, the epidemiological situation is relatively stable due to the implementation of European legislation since the early 2000s. A national monitoring and eradication plan, also covering VHSV, is currently being set up. It will probably lead to the discovery of new isolates that were previously difficult to detect on the basis of clinical observations. This molecular evolutionary work on IHNV in France has provided new data on how it is spread and will help enhance the control strategy.

METHODS

Virus isolates and dataset compilation

A dataset of 118 complete G genes from the five IHNV genogroups was extracted from GenBank (all sequences available to date). The dates and country of isolation were established based on the available literature [19, 21, 23, 31, 49, 50].

A dataset of 88 partial G genes was used in this study to represent IHNV isolates collected from France during the period 1987–2015. The isolates, geographic areas, isolation dates and GenBank accessions (nos MF464270 to MF464340 and no MF476202) are reported in Table 4.

RNA extraction, PCR amplification and sequencing

RNA was extracted from various internal organs, such as the anterior kidney, spleen, heart and brain, or from cell culture supernatants using the Nucleospin RNA virus kit (Macherey-Nagel, France). The partial glycoprotein gene, a sequence of 693 bp, was amplified using the primers 5'-AGAGATCCCTACACCAGAGAC-3' and the antisense primers 5'-GGTGAGTTTCTTCGCAGCAA-3'[51]. Reverse-transcription (RT) and amplification were performed with the SuperScript III One-Step RT-PCR system (Invitrogen, France) using the following mix: approximately 1 µg of RNA was added to 20 µM of each primer, 1 µl of high-fidelity Platinum Taq, 25 µl of reaction mix (2×) and water in a final volume of 50 µl. The G-gene RT-PCRs were conducted in a Mastercycler (Eppendorf, France) with an initial step at 52°C (30 min) followed by one step at 94°C (2 min) and then 40 cycles of 94°C (15 s), 52°C (30 s) and 68°C (60 s), with a final extension at 68°C (8 min). After agarose gel electrophoresis, the PCR products of interest were purified with a NucleoSpin gel and PCR clean-up kit (Macherey-Nagel, France) and then cloned using the TOPO TA cloning kit (Invitrogen, France). For each PCR product, three clones were selected and sequenced in both orientations using the Sanger method and a 3130 genetic analyser (Applied Biosystems). Sequence files were analysed visually using VectorNTI v 11.5 software.

Phylogenetic reconstruction

An alignment of partial G DNA sequences was performed with MUSCLE using SeaView 4 [52]. Phylogenetic reconstruction was done using Bayesian inference (BI) from an alignment of 93 sequences of 570 bp. Partition schemes and evolutionary models were selected via the Bayesian information criterion calculated in PartitionFinder v 1.1.1[53]: for the first and second codon partitions the K81+I+G model were used; for the third codon partition a model K81+G. BI was carried out with MrBayes 3.2.6 [54] on the CIPRES Science Gateway [55], with four chains of 5×10⁶ generations, trees sampled every 500 generations, and burn-in the value set to 20% of the sampled trees. We checked that the standard deviation of the split frequencies fell below 0.01 and confirmed the convergence of the runs to ensure convergence in tree search using Tracer v 1.6 software (http://tree.bio.ed.ac.uk/software/tracer/). The tree was visualized using FigTree v 1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/).

Ancestral reconstruction and discrete phylogeography

The BEAST software package v 2.1.3 was used to estimate the date of divergence and the discrete phylogeography from 118 full-length IHNV G gene sequences [56]. We used discrete phylogeography to perform ancestral reconstruction on a single character (i.e. location) and to determine the geographical location for the root of the tree [57]. The dataset was analysed using a HKY+G model under a relaxed uncorrelated exponential molecular clock. This model was evaluated with the coefficient of variation (CoV) [58], where a CoV value >0 was considered to be evidence of non-clock-like evolutionary behaviour. The ‘exponential coalescent tree prior’ was used to infer the complex population dynamics of IHNV. Three independent Bayesian MCMC runs were carried out for 30 million generations (to obtain effective sample size values of at least

Table 3. Molecular data for partial G sequences from IHNV

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Partial G</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of sequences</td>
<td>88</td>
</tr>
<tr>
<td>No. of nucleotides (amino acids)</td>
<td>570 (190)</td>
</tr>
<tr>
<td>% NT ID, (min–max)</td>
<td>98.15 (95.4–100)</td>
</tr>
<tr>
<td>% AA ID, (min–max)</td>
<td>97.82 (93.1–100)</td>
</tr>
<tr>
<td>% variable nucleotides</td>
<td>20.17</td>
</tr>
<tr>
<td>dS/dN</td>
<td>0.247</td>
</tr>
<tr>
<td>Tajima’s D</td>
<td>–1.85</td>
</tr>
</tbody>
</table>

AA ID, mean amino-acid identity; dS, number of non-synonymous mutations; dN, number of synonymous mutations; N, number; NT ID, mean nucleotide identity.
The convergences of the runs were confirmed using Tracer v 1.5, and an MCC tree was generated using TreeAnnotator v 1.5 and visualized using FigTree v1.4.

Genetic diversity

The basic population statistics and single nucleotide polymorphisms were calculated using two programs: DnaSp v5 [59, 60] and MEGA v 6 [61].

Funding information

The authors received no specific grant from any funding agency.
Acknowledgements

We are grateful to the following laboratories of the French surveillance network for providing IHNV isolates for characterization: Laboratoire des Pyrénées et des Landes (LPL), Laboratoire Départemental d’Analyses du Jura, Laboceca, Laboratoire Départemental Vétérinaire de l’Hérault, Laboratoire Départemental de l’Orne, Laboratoire Départemental d’Analyses du Pas de Calais and Laboratoire Agro-Vétérinaire Départemental de Seine Maritime. We also thank the ANSES sequencing facility (Unit Viral genetic and biosecurity) located in Ploufragan, Dr. A.M. Hattenberger, and D. Potratz, translator in the “External Communication Unit” at Anses, for editorial assistance.

Conflicts of interest

The authors declare that there are no conflicts of interest.

References


