Preventive Veterinary Medicine 92 (2009) 9–19

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

journal homepage: www.elsevier.com/locate/prevetmed

Traditional descriptive analysis and novel visual representation of diagnostic repeatability and reproducibility: Application to an infectious salmon anaemia virus RT-PCR assay

Charles Caraguel [a,b,*], Henrik Stryhn [b], Nellie Gagné [c], Ian Dohoo [b], Larry Hammell [a,b]

[a] Centre for Aquatic Health Sciences, Department of Health Management, Atlantic Veterinary College, University of Prince Edward Island, Charlottetown, PEI C1A 4P3 Canada
[b] Centre for Veterinary Epidemiological Research, Department of Health Management, Atlantic Veterinary College, University of Prince Edward Island, Charlottetown, PEI C1A 4P3 Canada
[c] Fisheries and Oceans Canada, Molecular Biology Unit, Aquaculture and Environmental Sciences, Moncton, NB E1C 9B6 Canada

ARTICLE INFO

Article history:
Received 16 April 2009
Received in revised form 20 July 2009
Accepted 21 July 2009

Keywords:
Repeatability
Reproducibility
Agreement
Homogenization
Infectious salmon anaemia virus
RT-PCR
Proportion of agreement
Kappa values
Phylogram

ABSTRACT

As a component of diagnostic test evaluation, the estimation of repeatability and reproducibility of an assay is necessary to assess the robustness and the transferability of the method among laboratories. Respectively defined as the agreement within and between laboratories, repeatability and reproducibility of a qualitative diagnostic test are traditionally reported using observed proportion of agreement or Kappa values. Applied to a recently designed RT-PCR assay for the detection of infectious salmon anaemia virus, repeatability only within a national reference laboratory and reproducibility with two additional independent regional laboratories were investigated. Homogenization of fish kidney tissue was conducted to potentially provide more uniform submission material, and to assess the effect of homogenization on laboratory comparability. Comparison of agreement between non-homogenized and homogenized tissue samples revealed different patterns of test results and unexpected alterations of agreement due to homogenization. This observation may be explained by cross-contamination of some samples during the homogenization process. One of the laboratories was in clear disagreement with the two others and impacted the overall reproducibility of the assay. Agreement levels were visually described using a novel tree-shape representation inspired from phylogenetic studies. The resulting phylogram illustrated the proximity of test findings between repeated samples within a laboratory and between laboratories, and facilitated the interpretation of the agreement levels.

1. Introduction

1.1. Traditional evaluation of diagnostic precision

The Office International des Epizooties (OIE, World Organisation for Animal Health) aims to safeguard international trade by publishing standards and guidelines for health and self-declaration of disease-freedom for animals and animal products. To diagnose infectious diseases and associated pathogens, the OIE recommends use of certified or validated diagnostic assays (OIE, 2008a). Diagnostic validation is defined as the evaluation of a test method based on its fitness for a specific purpose (OIE, 2008a). Validation is a multiple-stage process that determines the operating characteristics of the test including the assessment of its characteristics and performance at the bench level (estimation of analytical sensitivity, specificity, and repeatability),

* Corresponding author at: 550, University Avenue, Charlottetown, Prince Edward Island, C1A 4P3 Canada. Tel.: +1 902 566 0995; fax: +1 902 566 0823.
E-mail address: ccaraguel@upei.ca (C. Caraguel).

0167-5877/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.prevetmed.2009.07.011

evaluation of its accuracy (diagnostic sensitivity and specificity) and estimation of its precision ("diagnostic" repeatability and reproducibility) at the population level.

The estimation of the test precision is an important step of the validation process, although sometimes overlooked and neglected. Diagnostic repeatability is defined as the variation in test results that are obtained with the same method on identical test items in the same laboratory by the same operator using the same equipment within short intervals of time (within-laboratory consistency); and diagnostic reproducibility is defined as the variation in test results that are obtained with the same method on identical test items in different laboratories with different operators using different equipment (between-laboratory consistency) (ISO 5725-1, 1994). The concept of variation in binary outcome diagnostics is associated with the concept of agreement between test runs. We defined as "test run" or "run" a set of results obtained using the same method under defined conditions relative to the testing laboratory and the nature of the sample (identical conditions for repeatability and similar conditions for reproducibility). Agreement is traditionally expressed using the proportion of agreement (Pa) (proportion of tests results that agree) or using Cohen's Kappa values (κ) (Dohoo et al., 2003). It has been suggested that precision for binary tests can also be assessed using predictive intervals of diagnostic sensitivity (DSe), specificity (DSp) or overall accuracy (Cleophas et al., 2008). This study was restricted to the evaluation of diagnostic repeatability and reproducibility.

1.2. Novel approach inspired by phylogenetics

Phylogenetics is a discipline that investigates the relationship among organisms according to their gene similarity. The pairwise comparison of aligned nucleotide or amino acid sequences determines the degree of similarity (agreement) between genes. The measure of similarity is calculated as the proportion of nucleotides, or amino acids, that are identical between two sequences (Vandamme, 2003). Proportions of similarity (or dissimilarity) are usually summarized in a pairwise genetic distance matrix. The distance matrix is then used to reconstruct a phylogenetic tree that illustrates the evolutionary relationship among compared organisms. Distance matrices are comparable to agreement matrices that are reported in diagnostic evaluation studies. Methods using distance matrices for phylogenetic tree inferences are referred to as distance-based methods in contrast with character-based methods that integrate additional character information. Genes with high sequence agreement will be positioned closer to each other, whereas genes with low sequence agreement will not group together. Similarly to genetic sequences, laboratory test results can be aligned and analyzed using distance-matrix based models to visually represent agreement among laboratories in a tree shape.

1.3. Infectious salmon anaemia virus

Infectious salmon anaemia virus (ISAV) is an Orthomyxovirus, genus *Isavirus*, causing a hemorrhagic syndrome in salmonids. Primarily pathogenic for Atlantic salmon, *Salmo salar* L., the viral agent causing high mortality is a serious threat to the economic sustainability of many salmon aquaculture industries around the world. Originally found in Norway in 1984 (Thorud and Djupvik, 1988), clinical ISA was then chronologically reported in Canada (Mullins et al., 1998), Scotland (Rodger et al., 1998), Faroe Islands (Anonymous, 2000), USA (Bouchard et al., 2001), and recently in Chile (Godoy et al., 2008).

Absent in some areas of Atlantic salmon production (e.g. Tasmania, Australia; British Columbia, Canada), ISAV is listed as a notifiable aquatic disease by the OIE (OIE, 2008b). Consequently, for international trade purposes, diagnostic methods used for screening, certification, confirmation and control require validation. The implementation of the National Aquatic Animal Health Program (NAAHP) in Canada, including national reference laboratories and surveillance programs, aims at controlling and preventing the emergence and spread of aquatic disease. Since ISAV surveillance is a goal of the program, it was required that a recently designed Reverse-Transcriptase Polymerase Chain Reaction (RT-PCR) assay for ISAV (Gagné et al., data unpublished) be validated.

1.4. Repeatability and reproducibility of the ISAV RT-PCR

A single study previously investigated the repeatability and reproducibility of an ISAV RT-PCR assay in three different laboratories (Nérette et al., 2005). The study revealed substantial differences in repeatability (Pa ranging from 76 to 98% and κ from 0.50 to 0.96). In addition, there was a serious disagreement explained by one laboratory with a higher proportion of positive tests although a substantial reproducibility was found between the two others (Pa = 91% and $\kappa = 0.79$). The authors proposed that factors associated with sample and testing conditions may have affected the assessment of reproducibility and repeatability: (i) heterogeneous distribution of virus in the organ may have resulted in virus quantity inconsistency among replicated samples; (ii) differences in testing protocols (i.e. different set of primers and methods) may have compromised the comparability of laboratories; (iii) differences in agarose gel interpretation and confirmation protocols (i.e. whether a sample with a weak gel band was retested) may have also affected interpretation of results in the three laboratories.

1.5. Objectives

The objective of this study was threefold. The first objective was to describe qualitative diagnostic precision of a newly designed ISAV RT-PCR (Gagné et al., data unpublished) in three different laboratories using identical standard operating procedures for testing and interpretation. Specifically, we estimated the repeatability only within the designated national reference laboratory for ISAV in Canada (the molecular biology laboratory of the Department of Fisheries and Oceans, Moncton, Canada), and the reproducibility by including two independent laboratories in the study. The second objective was to investigate the impact of potential heterogeneous distribution of viral

particles among replicate samples by assessing agreement of homogenized tissue samples. The third objective was to develop a novel visual approach to describe test agreement using distance-matrix based tree reconstruction inspired from phylogenetic studies. This new approach was not intended to replace former methods but to facilitate the illustration and complement interpretation of agreement with a new perspective.

2. Materials and methods

2.1. Study material

2.1.1. Sample selection

Kidney samples from 100 Atlantic salmon were selected from archives by combining different origins to target a prevalence of approximately 50% according to McClure et al. (2004); briefly, 45 apparently healthy fish were from three exposed cages (15 fish from each infected site) (expected prevalence of 28.1%), 35 apparently healthy fish were from an infected cage (expected prevalence of 41.5%), and 20 dead or moribund fish were from ISA clinically affected cages (10 fish from two different sites) (expected prevalence of 100%). From each fish, kidney samples were collected aseptically in replicates of six and stored in RNA*later* (Ambion Inc., Austin, TX, USA) at −80 °C after a 24 h period at 4 °C.

2.1.2. Sample allocation

Each sample was coded with a random identification number to blind laboratory operators and to avoid test-review bias (Ransohoff and Feinstein, 1978). Sample distribution and testing objectives are summarized in Fig. 1. From each salmon, duplicate samples were sent on dry ice to the reference laboratory (lab A) to estimate the repeatability, and single samples were transported on dry ice to two other laboratories (labs B and C) to estimate the reproducibility. Due to the restricted number of samples per fish, repeatability was only assessed in lab A. The remaining two samples were combined, homogenized and aliquoted with equal volume (250 μl) in four coded microtubes.

Homogenization was performed in lab A by transferring the two samples in a 2 ml microtube filled to the upper limit with RNA*later* and homogenized using a FastPrep® FP120A homogenizer (MP Biomedicals) at 5.5 m/s for 20 s twice. Two aliquots of 250 μl of homogenate from each fish stayed in lab A stored at −80 °C to estimate the repeatability and single aliquots of 250 μl of homogenates were sent frozen on dry ice to the other two laboratories to estimate the reproducibility (Fig. 1). Each of the participating laboratories agreed to test for the presence and absence of ISAV using the same RT-PCR protocol provided by the reference laboratory (lab A).

2.2. Testing protocol (Reverse-Transcriptase Polymerase Chain Reaction and interpretation)

For RNA extraction, a piece of tissue (approximately 30 ± 5 mg) was removed and homogenized in 1 ml of TRI reagent (Molecular Research Inc.) with a FastPrep FP120 (Savant Instruments). For homogenates, microtubes were centrifuged to remove the RNA*later* before adding 1 ml of TRI reagent and homogenizing. Manufacturer's instructions were followed, except for two additional washes of the RNA pellet with 75% ethanol. RNA pellets were resuspended in 50 μl of sodium citrate buffer 1 mM pH 6.4 containing an RNase inhibitor (Qiagen). RNA was further diluted if necessary and quantified on a spectrophotometer and normalized. A maximum of 1000 ng/μl was used for reverse transcription.

A one-step RT-PCR was used to detect ISAV, using the Qiagen One-step RT-PCR kit (Qiagen). The mixture comprised 5 μl of Q solution, 0.32 μM of each primer (404F: 5′ tgg gca atg gtg tat ggt atg a-3′ and RA3(583R): 5′ gaa gtc gat gaa ctg cag cga-3′), 1 μl of enzyme, 5 μl of buffer, 1 μl of dNTP, 11.2 μl of H_2O and ≤1 μg of RNA, for a total volume of 25 μl. PCR conditions consisted of an initial hold at 50 °C, 30 min, and 95 °C, 15 min, followed by 10 cycles of touchdown PCR starting with 94 °C, 40 s; 72 °C,

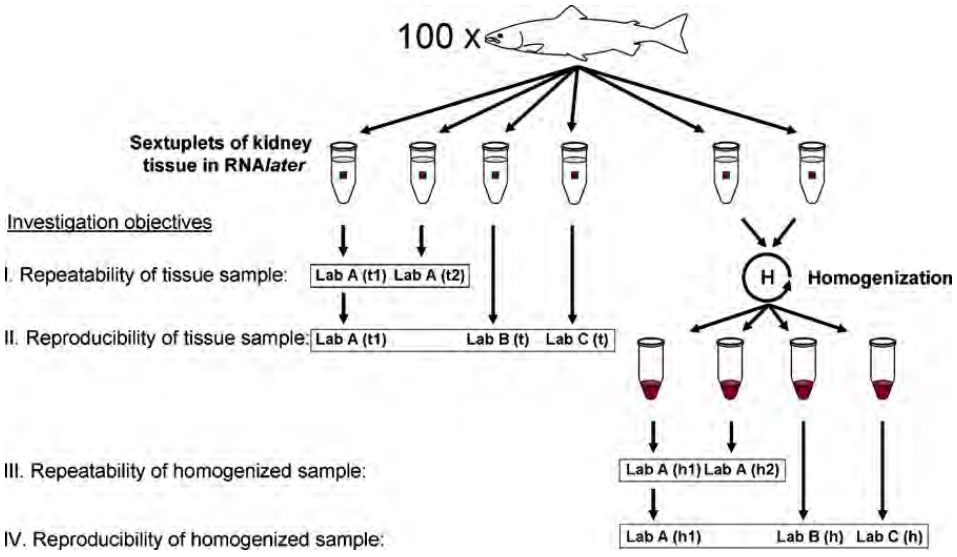

Fig. 1. Sample allocation and investigation objectives to study RT-PCR repeatability and reproducibility. (t1): Non-homogenized sample, duplicate 1; (t2): non-homogenized sample, duplicate 2; (t): non-homogenized sample; (h1): homogenized sample, duplicate 1; (h2): homogenized sample, duplicate 2; (h): homogenized sample.

40 s; 72 °C, 60 s, and lowering by 1 °C the annealing temperature after each cycle. Then 40 cycles at 94 °C, 40 s; 62 °C, 40 s; 72 °C, 60 s were added, and a final extension of 72 °C, 10 min and holding at 20 °C completed the program. PCR products (10 μl) were electrophoresed in 6% acrylamide, visualized with ethidium bromide, and compared to positive controls and a DNA ladder. A band at the same expected size (179 bp) as the control was considered positive. For quality control, extraction blanks (no sample) were included every 15th tube during extractions, and blanks (water) were added at the RT-PCR step. Electrophoresis gels were examined carefully, and PCR was repeated on samples where a very weak intensity band at the expected size was observed initially. If the second PCR result was positive again, the final result was positive; if not, it became negative.

2.3. Statistical analysis

2.3.1. Descriptive statistics

Test results from each laboratory were collated and first analyzed using Stata SE 10.0 (Stata Corp., College Station, TX, USA, 2007). The first agreement statistic computed was the observed proportion of agreement (Pa), giving the proportion of paired test results that agreed either on a positive or on a negative test result between two test runs. Exact confidence intervals (CIs) for observed agreement were computed. Average Pa were also computed as a mean of Pa estimates from all possible pairs of test runs within lab A (overall repeatability), and among the three laboratories (average reproducibility of homogenized and non-homogenized samples, and overall reproducibility). CIs were computed using 2.5 and 97.5% percentile values from bootstrapped estimates resampled 1000 times.

The second agreement statistic computed was Cohen's Kappa (κ), commonly used for subjective rating. Ranging from −1 to +1, this value represents the level of agreement beyond chance (Dohoo et al., 2003). The κ was computed for agreement among three laboratories data together (Fleiss et al., 2003). CIs for the κ statistic were computed using an analytical method for comparison of two test runs (Fleiss et al., 2003) and a bootstrap method for three runs (Lee and Fung, 1993). Prior to each paired κ estimation, a McNemar's test (exact binomial test for correlated proportions) was performed to assess if proportions of positive results differed between test runs. Evidence of proportion disagreement between runs would constitute disagreement between runs and reduce the interest in κ estimation (Dohoo et al., 2003).

Due to violation of the assumption of independent observations (test results obtained from the same fish), two Pa from different conditions (i.e. repeatability of non-homogenized vs. homogenized samples) could be compared using a McNemar's test by defining agreement/non-agreement (e.g. comparing two runs) as a binary outcome. However, this method does not consider proportion of agreement on a positive and on negative result. The effect of homogenization on agreement was then assessed by testing symmetry and marginal homogeneity in contingency tables. The symmetry test compares symmetrical cells around the agreement diagonal of the contingency table, whereas the marginal homogeneity test compares the marginal distributions of test results (Table 1). Agreement among all test results from non-homogenized samples were compared to agreement among all test results from homogenized samples using exact test for symmetry and Stuart-Maxwell test for marginal homogeneity (symmetry command; Stata Base Reference Manual, 2007). In addition, following the approach outlined in Agresti (2002), a quasi-symmetry model for the contingency table was fit, and marginal homogeneity was tested using a likelihood ratio test of symmetry in this model. The same approach was used to compare agreement in homogenized and non-homogenized samples within- and between-laboratory data.

Table 1
Contingency table comparing non-homogenized and homogenized sample results from the four tests (two tests in reference lab A and one test each in participating labs B and C)[a,b].

# of positive	Homogenized					
	0	1	2	3	4	Marginal
Non-homogenized						
0	**23**	6	6	0	0	*35*
1	4	**6**	4	1	3	*18*
2	1	0	**2**	1	2	*6*
3	0	0	0	**3**	4	*7*
4	0	0	0	2	**31**	*33*
Marginal	*28*	*12*	*12*	*7*	*40*	99

[a] Symmetry test compared symmetrical cells around the agreement diagonal (in bold).
[b] Marginal homogeneity test compared the marginal distributions of non-homogenized and homogenized samples (italicized).

2.3.2. Distance matrix

A summary matrix of observed agreement (Pa) and disagreement (i.e. proportion of results that disagree between the two test runs: 1 − Pa) was generated for all possible pairwise comparisons. In phylogenetic methods, the observed disagreement is also called *observed distance* or *p-distance* (Van de Peer and Salemi, 2003). Thus the distance matrix summarized the relative distance of the runs to each other based on their test results. Smaller distance values indicate closer result findings between two laboratories.

2.4. Test run phylogram

2.4.1. Pseudogold standard

A pseudogold standard (PGS) was created to provide a consensus reference baseline for the test results alignment. According to the PGS definition of Nérette et al. (2008), ISA positive and ISA negative classification criteria were arbitrarily based on the combination of six test results for each fish, excluding duplicate results in lab A. "Infected" fish were any fish with more than three positive tests out of the six (>3/6). "Non-infected" fish were any fish with three or less positive tests out of the six (≤3/6).

2.4.2. Alignment formatting

Initially formatted with individuals in rows and runs in columns, tests results were transposed so individuals were in columns and test runs in rows. Negative results, "*0*",

Table 2
Summary of ISAV diagnostic test descriptive agreement statistics, proportions and Kappa values, according to sample type and laboratories comparison. A1: reference laboratory A, duplicate 1; A2: reference laboratory A, duplicate 2; B: laboratory B; C: laboratory C.

Agreement level	Repeatability		Reproducibility					
Sample type	Non-homogenized	Homogenate	Non-homogenized			Homogenate		
Lab comparison	A1/A2	A1/A2	A1/B	A1/C	B/C	A1/B	A1/C	B/C
0–0	49	35	51	39	40	40	31	35
1–1	35	45	36	39	41	41	46	45
1–0	7	8	6	3	2	12	6	2
0–1	9	12	7	19	17	7	16	17
Total (count)	100	100	100	100	100	100	99	99
Pa (CI)	0.84 (0.75–0.90)	0.80 (0.71–0.87)	0.87 (0.79–0.93)	0.78 (0.69–0.86)	0.81 (0.72–0.88)	0.81 (0.72–0.88)	0.78 (0.68–0.85)	0.81 (0.72–0.88)
Pa average[a] (CI)	0.81 (0.75–0.86)			0.82 (0.76–0.88)			0.80 (0.74–0.86)	
					0.82 (0.77–0.86)			
McNemar's test (*P*-value)	0.610	0.370	0.780	0.000*	0.000*	0.250	0.033*	0.000*
Kappa (Cohen's)	0.674	0.597	0.734	0.571	0.621	0.621	0.550	0.628
CI	0.53–0.82	0.44–0.75	0.60–0.87	0.42–0.72	0.47–0.77	0.47–0.77	0.40–0.71	0.48–0.77
3-Rater Kappa	na	na		0.639			0.595	
CI (bootstrap = 1000)	na	na		0.52–0.76			0.47–0.71	

na: non-applicable; Pa: observed proportion of agreement; CI: confidence interval.
[a] Computed as the mean of all possible Pa estimates between runs within lab A or among the three laboratories.
* Significant McNemar's test ($P < 0.05$): significant difference of proportion of positive results between the two test runs; thus corresponding Kappa value is less relevant.

were recoded with an "*a*" (corresponding to adenine) and positive results, "*1*", were recoded with a "*g*" (corresponding to guanine) in a FASTA format to suit the requirements of the DNA sequence alignment editor software BioEdit version 7.07 (Hall, 1999) and to allow for further phylogenetic analyses. Later, test results were edited and displayed as a sequence alignment of "n" and "p" ("negative" and "positive", respectively) where only results in disagreement with the PGS are highlighted, and test results in agreement were symbolized by a "." as a placeholder.

2.4.3. Distance-matrix based tree reconstruction model

To conduct tree reconstruction, the FASTA alignment was transferred into the MEGA format using the package MEGA version 4 (Tamura et al., 2007). The alignment was considered as a non-protein-coding nucleotide sequence and phylograms were obtained using the distance based Neighbor-Joining (NJ) method. The model used distances based on the number of differences, and missing data were handled by pairwise deletion. Statistical support for tree topologies was bootstrap-resampled 1000 times (Felsenstein, 1985). Bootstrap support values (proportion of resampled trees that include the node of interest) were reported in percentage on the nodes of the original tree. Phylograms were edited using the TreeExplorer software appended to the MEGA package.

3. Results

3.1. Descriptive statistics

Results were obtained for all eight test conditions for all 100 samples, except that lab C had insufficient material for one homogenized sample (only 99 results for homogenates). Among the 100 non-homogenized samples, duplicates 1 and 2 of lab A detected respectively 42 and 44 positives, lab B detected 43 and lab C detected 58. Among the 100 homogenized samples, duplicates 1 and 2 of lab A detected respectively 53 and 57 positives, lab B detected 48 and lab C detected 62 (out of 99 results). Agreement statistics of interest (Pa and κ) and CIs are summarized in Table 2.

Overall repeatability revealed slightly lower Pa than overall reproducibility (0.81 and 0.82, respectively), although the overlapping of CIs provided little evidence of significant difference (Table 2). Tests from pairwise comparisons involving lab C showed serious disagreement with the two other laboratories regardless of the sample type (significant McNemar's test). Estimates of κ ranged from 0.57 to 0.73 and supported Pa results (Table 2).

The average proportion of positive results for non-homogenized samples was 46.8%; and the average proportion of positive results for homogenized samples was 56.4%. Table 1 shows the contingency table of overall test results comparing non-homogenized and homogenized samples. Both symmetry and marginal homogeneity tests showed a significant difference ($P < 0.05$) in overall agreements, repeatabilities and reproducibilities between non-homogenized and homogenized sample results. As an example, for overall agreement, we observed more complete agreements (all four tests agree for a given sample type) on positive results than on negative results in homogenized samples whereas non-homogenized samples showed the opposite pattern (Table 1). However, intermediate agreements (two to three tests that agree) were quite comparable (Table 1). Quasi-symmetry modelling procedure showed a good fit to the data and the hypothesis of marginal homogeneity was significantly rejected against that model for overall agreement, repeatability and reproducibility data (all $P < 0.05$).

Table 3 represents a summary matrix of observed agreement (Pa) and disagreement ($1 - \text{Pa}$) of all possible pairwise comparisons. The minimum disagreement or distance (0.09) was observed between lab A, duplicate 2, and lab B with non-homogenized sample; and the maximum was observed between non-homogenized sample in lab A and homogenized sample in lab C (0.25). Additionally, significant McNemar's test revealed serious disagreement despite high Pa for several pairwise comparisons (Table 3).

3.2. Test runs phylogram

According to the PGS, out of the 100 salmon sampled, 48 were positive and 52 negative for ISAV. Assuming that the PGS is correct, the targeted prevalence in the submitted samples (~50%) was reached which was fortuitous since the estimates within each salmon group did not agree with the ones observed in McClure et al. (2004). Using the PGS as a reference sequence, the alignment of test results from the eight runs (three laboratories, two sample types, and

Table 3
Agreement matrix with proportion of agreement (lower left corner) and proportion of disagreement or distance (top right corner in bold) between runs; (t1): non-homogenized sample, duplicate 1; (t2): non-homogenized sample, duplicate 2; (t): non-homogenized sample; (h1): homogenized sample, duplicate 1; (h2): homogenized sample, duplicate 2; (h): homogenized sample.

Runs	Lab A(t1)	Lab A(t2)	Lab B(t)	Lab C(t)	Lab A(h1)	Lab A(h2)	Lab B(h)	Lab C(h)
Lab A(t1)	–	**0.16**	**0.13**	**0.22***	**0.19***	**0.19***	**0.14**	**0.25*$_{Max}$**
Lab A(t2)	0.84	–	**0.09$_{Min}$**	**0.20***	**0.23***	**0.19***	**0.14**	**0.25*$_{Max}$**
Lab B(t)	0.87	0.91$_{Max}$	–	**0.19***	**0.22***	**0.18***	**0.11**	**0.24***
Lab C(t)	0.78*	0.80*	0.81*	–	**0.19**	**0.13**	**0.12***	**0.15**
Lab A(h1)	0.81*	0.77*	0.78*	0.81	–	**0.20**	**0.19**	**0.22***
Lab A(h2)	0.81*	0.81*	0.82*	0.87	0.80	–	**0.11***	**0.14**
Lab B(h)	0.86	0.86	0.89	0.88*	0.81	0.89*	–	**0.19***
Lab C(h)	0.75*$_{Min}$	0.75*$_{Min}$	0.76*	0.85	0.78*	0.86	0.81*	–

Min: minimum; Max: maximum.
* Significant McNemar's test ($P < 0.05$): significant difference of proportion of positive results between the two runs; thus serious disagreement.

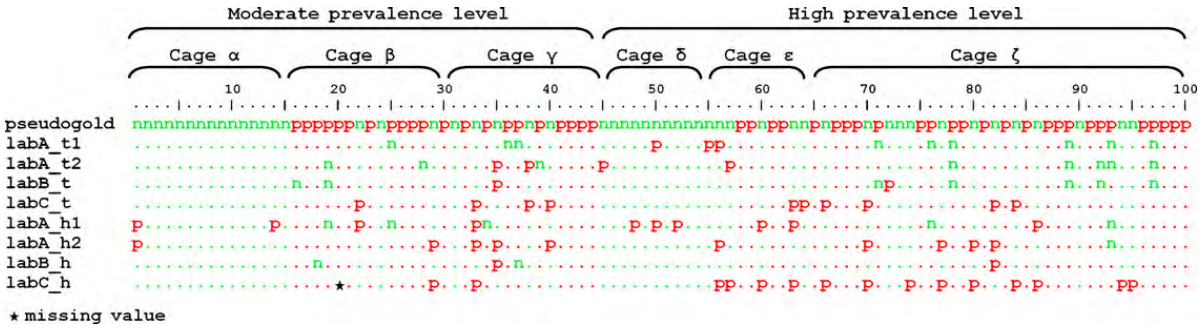

Fig. 2. Test result alignment: sampled salmon (in column) were clustered by cage origin and expected prevalence level population grouping (moderate level: apparently healthy fish from exposed cage; high level: mix of apparently healthy, mortality and moribund fish from infected cage). Negative was recoded as "n"; positive as "p"; and by column a dot (.) indicates same result as the first row. Greek letters: arbitrary cage number. (t1): Non-homogenized sample, duplicate 1; (t2): non-homogenized sample, duplicate 2; (t): non-homogenized sample; (h1): homogenized sample, duplicate 1; (h2): homogenized sample, duplicate 2; (h): homogenized sample.

duplicates in lab A) highlighted the differences among test results (Fig. 2).

The computed unrooted tree represents the relative position among the eight test runs (Fig. 3). Except for lab C, all non-homogenized samples were grouped together and formed a cluster supported by a low bootstrap value (61%). Within the cluster, lab A (duplicate 2) and lab B were the closest and associated with a high bootstrap value (88%) as previously shown in the distance matrix (Table 3). Except for lab B, all homogenized samples were grouped together including the lab C non-homogenized sample and formed a cluster not supported by bootstrap (50%). Within the homogenates cluster, both lab C samples and lab A homogenate (duplicate 2) were grouped based on a low bootstrap value (60%). Lab C homogenate and lab A homogenate (duplicate 2) group was not supported by bootstrap (43%) which does not separate them from lab C non-homogenized. The homogenized sample from lab B was consistently separated from the two main clusters.

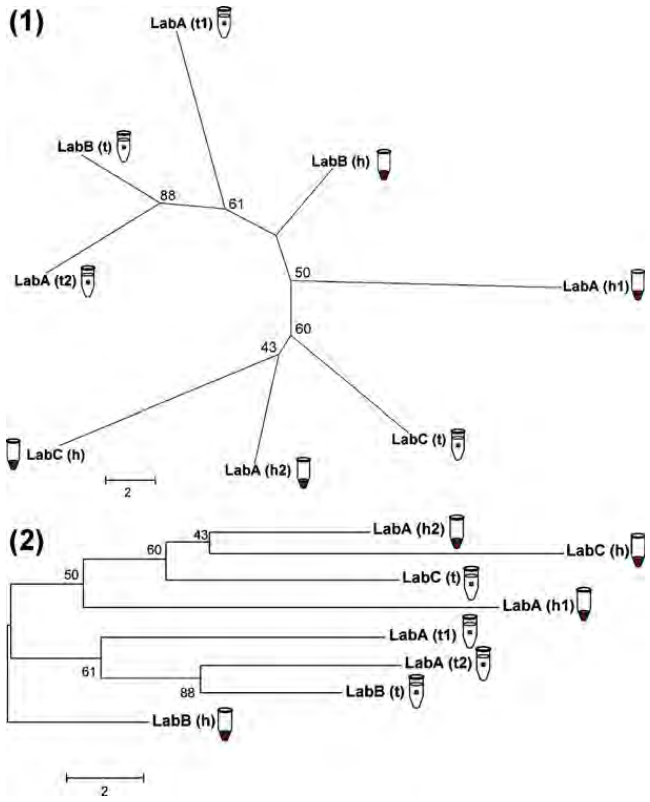

Fig. 3. Unrooted phylogram representing agreement among test runs. Star topology (1); tree topology (2). The distance between two runs is visually assessed by the relative length of branches that connect them and are scaled based on the number of differing results out of the 100 samples tested. (t1): Non-homogenized sample, duplicate 1; (t2): non-homogenized sample, duplicate 2; (t): non-homogenized sample; (h1): homogenized sample, duplicate 1; (h2): homogenized sample, duplicate 2; (h): homogenized sample.

4. Discussion

4.1. Formal descriptive analysis of agreement

Kappa values (κ) are usually used to compare test results agreement beyond expected agreement (Dohoo et al., 2003). However, κ was estimated for all test runs at an identical prevalence level of approximately 50%, and therefore, expected agreement due to chance would be consistent for all agreement estimates. Accordingly, we decided to base most of the discussion on Pa with little reference to κ values.

4.1.1. Repeatability

For a full evaluation of repeatability, estimation of agreement within labs B and C should have been conducted. However, due to restricted sampling material, these estimations are missing in this study and it would require further repeatability evaluation for the test to be validated in these two participating laboratories. Based on the set of samples tested, the overall RT-PCR repeatability in the reference laboratory was approximately 81% (proportion of agreement). One could interpret this value as one in every set of five samples tested does not provide the same result when tested a second time (or 19% of the samples do not repeat the same result). When two results from the same individual disagree, based on a dichotomous outcome, one of the results has to be incorrect. Consequently, qualitative diagnostic tests that lack repeatability also lack accuracy. In this particular case, 9.5% ($1/2 \times 19\%$) of the combined results are either false positive or false negative.

Greater frequencies of false negative results imply decreased diagnostic sensitivity (DSe) that can be explained by several factors. The most likely reason for false negative results is a limited analytical sensitivity. Defined as the minimum threshold of detection, the analytical sensitivity of a RT-PCR depends on several method-specific factors including primer stringency (i.e. design of primer, nature and freshness of reagents), reaction preparation (i.e. ratio of target and primers), and thermocycling protocol (i.e. annealing temperature). It is possible that samples with a low concentration of target and/or a complex molecular matrix might not be detected if the primers do not bind to the target during the first cycles of the reaction. Thus, for infected samples with a concentration close to the limit of detection, the target is sometimes detected or not, and the repeatability decreased. Among the 19% of non-repeatable results, some may be due to low concentration of target; hence the estimate of repeatability depends strongly on the nature of the sample tested. During routine surveillance, the pathogen load of screened apparently healthy individuals is likely to be low and the frequency of false negative results is expected to be high, whereas for diagnostic confirmation, the pathogen load of clinically suspected individuals is likely to be high and the frequency of false negative results is expected to be low.

Greater frequencies of false positive results imply decreased diagnostic specificity (DSp) that can be explained by several factors. The most likely reason for false positive results with RT-PCR is cross-contamination (Wilson, 1997). Among the 19% of non-repeatable results, some may be due to contamination. Agreement among false positive results was not complete indicating that contamination was most likely random and not systematic. In theory, the probability of contamination should be associated with the prevalence of infection in the sample pool tested. During routine surveillance, the prevalence of infected samples is usually low and the frequency of false positive results is expected to be low, whereas for diagnostic confirmation, the prevalence of infected samples is likely high and the frequency of false positive results is expected to be high. As discussed previously (Begg, 1987; Greiner and Gardner, 2000), test operating characteristics depend strongly on the targeted population. Thus, the specific purpose and use of the diagnostic method must be clearly defined to reflect the assay performances (OIE, 2008a).

Although no minimum threshold has been set as suitable for validation of qualitative diagnostic tests, it is assumed that greater repeatability and reproducibility is preferred, provided that the McNemar's test is non-significant. In general, authorities in charge to design ISAV control programs need to make decisions about which tests to include in the program. Tests with low repeatability should not be considered. This study made available repeatability estimates for the developed RT-PCR for comparison with other considered diagnostic tests for ISAV control and surveillance programs.

4.1.2. Reproducibility

The overall reproducibility of 82% proportion of agreement was estimated slightly higher than the overall repeatability of 81% (Table 1). In theory, it is expected to observe a larger variation in results between than within laboratories. Factors influencing the reproducibility include the ones influencing the repeatability plus factors differing among laboratory practices such as technician habit and training, equipment, facilities structure and organisation. For example, false positive or false negative results could arise from the subjective reading and interpretation of bands in electrophoresis gels. Laboratory technicians must decide the dichotomous result (i.e. positive or negative) according to the test protocol, its own training and experience. Although mostly expected to influence the assay reproducibility, gel interpretation may also affect the repeatability due to human error and multiple laboratory staff.

In general, a major source impacting reproducibility of amplification methods is electrophoresis gel reading and the decision to retest or not a sample showing a band of weak intensity. In routine, not all laboratories do proceed to the retesting of sample with a weak gel band since they consider the band present and the sample positive. However, in this study all laboratories retested the weak gel bands (cf. Section 2). A sample will be retested or not based on the subjective call from the operator. The subsequent interpretation in series of the repeated test result (if the second test result is negative the sample is declared negative) aims to remove potential false positive results and increase DSp. However, this procedure may also interpret as negative some truly low infected samples that are poorly repeatable and decrease DSe. The ability of the operators to classify a gel band as weak or strong according to the intensity is a source of variation that could have explained some test results discrepancy in this study. Overall, the gel reading subjectivity had little influence in this study since between laboratories agreement did not show substantial differences with within laboratory agreement. However, note that between laboratories values report an average and do not separate individual pairwise agreements.

Regardless of sample type, lab C had significantly higher proportions of positive results, suggesting serious disagreement with other laboratories (Table 1). This can be explained either by a higher DSe or by lower DSp in lab C. Assuming that the PGS is correct, lab C tended to have more false positive results (Fig. 2). Mostly with homogenates, lab B had overall a lower proportion of positive results. This can be explained by a lower DSe or a higher DSp in lab B. Assuming again that the PGS is correct, lab B seemed to have more true results than the two other laboratories (Fig. 2). However, pairwise comparisons between labs A and B revealed better agreement than comparisons involving lab C (Table 3). Although, when two proportions of agreement are similar, the proportion of tests that agree on a positive result and the proportion of tests that agree on a negative result can still differ. As an example, McNemar's test detected significant proportions of positive results on homogenized tissue samples between labs B and C and not between labs A and B for identical Pa (Table 1). More sophisticated modelling, using multilevel logistic regression models, could alleviate the assumption of independent observations to simultaneously explore the

effects of laboratories and homogenization on agreement levels. Nonetheless, there are evidences of variation of performance in lab C associated with low reproducibility.

No international or standard guidelines are available to define acceptable levels of reproducibility. However, reproducibility estimates of the test provides an indication of how easily a test can be distributed to other laboratories if this is necessary as part of the control program. Clearly, the evaluated RT-PCR for ISAV was not properly transferred to lab C and it would require further protocol harmonization for this laboratory to be included in conjoint ISAV control program.

4.2. Homogenization effect

The second objective of this study was to assess the effect of homogenization by comparing agreement between non-homogenized and homogenized samples. On average, homogenized samples had a higher proportion of positive results than non-homogenized samples (56.4% vs. 46.8%), which implies that homogenization impacted the test performances with either increased analytical sensitivity and DSe, decreased DSp, or both. Homogenized samples revealed slightly lower repeatability and reproducibility compared to non-homogenized samples (Table 1). We expected a strong improvement of agreement with homogenized sample as the supposedly heterogeneous distribution of ISAV particles in the salmon kidney was one suggested explanation for the low RT-PCR reproducibility in Nérette et al. (2005). Furthermore, significant symmetry and marginal homogeneity tests suggested a shift in testing pattern with homogenized samples. The marginal distribution of overall agreement revealed that the proportion of complete agreement (all test results agree for a given sample type) was higher with positive than with negative test results for homogenized samples while it was the opposite with non-homogenized samples (Table 2). Although it was expected that higher complete agreement was observed for positive results with homogenized samples, a decrease of complete agreement for negative results was totally unexpected in particular with the assumed dilution effect of homogenization (see below).

Reviewing the homogenization protocol and the fact that 12 fish with some positive results for homogenized samples were negatives for the four tests in non-homogenized samples (Table 2), it is plausible that cross-contamination occurred during homogenization. The use of pipette tips that lacked a filter might explain the potential false positives among the homogenates. Homogenization protocols, in particular of solid tissue, must be optimized and standardized in order to reach the maximal homogeneity in the sub-aliquots. However, even in a scenario of contamination, we would have expected all four homogenized aliquots to be contaminated. Random contamination with few viral RNA might thus explain a decreased repeatability or reproducibility with homogenized samples.

Repeatability and reproducibility estimates of tissue samples depend on the assumption that sub-samples from a same fish are identical and that the detection threshold of the assay is constant. Both can be either associated or independent. For example, in the initial phase of the infection, only clusters of low numbers of viral particles may be present in the salmon kidney to be tested. At this stage, homogenization would dilute already low levels of virus and produce more false negative results and lower agreement. Further, the progression of the infection would produce clusters of high numbers of viral particles as a result of viral replication. Homogenization would harmonize viral concentration among sub-samples at a detectable level despite dilution. Finally, later stages of infection are expected to result in high numbers of viral particles throughout the organ. Homogenization would then provide little advantage since all tissue samples will contain high virus load. Although unrealistic according to the ISAV histopathology (Byrne et al., 1998), another scenario would be a spread of low numbers of viral particles throughout the organ. Homogenization would then provide limited benefit since each tissue sample would already have similar levels of particles. Agreement level would diminish mainly due to inconsistent detection of low virus load.

Overall, homogenization was of limited value in this precision evaluation; we suspect that occasional non-systematic cross-contamination in non-infected samples affected the specimen's comparability and the agreement estimation. Also, repeatability was lower in homogenized samples which would go against the hypothesis of heterogeneous distribution of viral particles in the infected salmon kidney. However, homogenization increased the proportion of complete agreement on a positive test result which, in infected fish, would support the variable virus distribution. Tissue homogenization has diverse application for ISAV control program (sample pooling, certified reference and control material, laboratory proficiency testing) and is greatly needed but a more detailed evaluation of its influence on test comparisons requires a close monitoring and protocol optimization.

4.3. Novel descriptive analysis of agreement

4.3.1. Test result alignment

The approach offered in this study of using column (individual fish) and row (test run) to represent the test results similar to a genetic sequence alignment has not been previously published. This is a convenient and intuitive way for the reader and the investigator to screen and visually compare test results (Fig. 2), whereby each result is compared within a fish (column) to the first aligned test, in this case the PGS. No alignment algorithm is needed as each test result corresponds to a defined fish (or column). From the alignment, it is possible to generate a matrix of pairwise comparisons among sequences, also called a distance or similarity matrix.

4.3.2. Test runs phylogram

The phylogram graphically represents the matrix of agreement and facilitates the visualization of the relative position among test runs. Distance-based phylograms are generated from the matrix of pairwise genetic distances. A matrix of pairwise genetic distances is very comparable to

a matrix of tests disagreement (1 − Pa) (Table 2). However, due to variable pressure of evolutionary changes, it is common in phylogeny to correct the estimates of genetic distance for multiple events per site (Van de Peer and Salemi, 2003). Since the probability that test results will be first positive then negative then positive again is extremely low, an evolutionary correction in the distance computation was judged not necessary. However, future development of this approach may benefit from incorporating different weights for results changing from negative to positive and from positive to negative. Indeed, depending on the diagnostic test method being assessed, the probability of a false positive result (e.g. contamination) might be higher than the probability of false negative result (e.g. target decay during transport). However, more knowledge on the assay performances is required to implement this refinement.

The distance matrix obtained from the alignment was identical to the initially computed disagreement matrix. Distance-matrix based tree reconstruction differentiates methods that are character-based and non-character-based (Van de Peer and Salemi, 2003). The reconstruction generated by this study used only two arbitrary characters (adenine and guanine for negative and positive result respectively) with equal weights of substitution, giving no value to the character chosen.

Also referred to as pairwise distance methods, non-character-based methods include cluster or minimum evolution analyses (Van de Peer and Salemi, 2003). The latter was preferred to the former because cluster analysis only assumes constant evolution (existence of a molecular clock) and would position all test runs in the tree equidistantly from the baseline or root. The commonly used method to estimate the minimal evolution tree is the Neighbor-Joining (NJ) method (Van de Peer and Salemi, 2003). We selected only pairwise deletion in cases of missing data to avoid losing all the information from a fish when only one test result was missing. The obtained tree is an unrooted phylogram scaled for distances (set as the number of differing results) among test runs (Fig. 3).

Bootstrap analysis is commonly used to evaluate the robustness of nodes that support tree branches. The magnitude of the bootstrap values is intimately correlated to the numbers of variable sites (or fish in this instance) that are informative in the alignment. A variable site is informative if there are at least two different characters that are represented at least twice at the given site. All bootstrap values, except for one node, were lower than 70% (Fig. 3). The low resolution of the tree suggests some caution in its interpretation. With only 100 fish, the number of fish that discriminate the test runs might be limited and a higher number of salmon might provide a better tree resolution. However, poor tree resolution will also be expected when test runs greatly agree (high consistency and precision).

The obtained phylogram illustrates the relative agreement among test runs (Fig. 3). The distance between two runs is visually assessed by the relative length of branches that connect them. Non-homogenized samples were clearly clustered and showed some testing consistency, although lab C was separated, confirming poorer reproducibility. Within this cluster, non-homogenized samples of lab A (duplicate 2) and lab B were grouped separately from lab A (duplicate 1) which supported previous observations of undifferentiated repeatability and reproducibility on non-homogenized samples.

The cluster of homogenates, excluding lab B but including non-homogenized lab C, was poorly supported by bootstrapping (50%). This weak separation presumed a tendency of homogenized samples to test differently. However, the wide distribution of homogenates in the tree supported a serious inconsistency in the testing pattern compared to non-homogenized samples. The homogenization protocol appeared to be inadequately refined or standardized to harmonize the testing pattern. Lab C revealed a distinct testing pattern with a reasonable repeatability regardless of the sample type and more closely resembling homogenized samples. However, lab C clearly decreased the overall assay reproducibility and must standardize its testing procedure to be comparable to the other laboratories.

The distance-matrix based tree reconstruction approach helps the investigator and the reader to visualize the relative proximity among test runs and to understand the distinctive testing patterns reflected by each of them.

5. Conclusion

Utilisation of basic phylogenetic reconstruction techniques provides a convenient and intuitive approach to visually compare and assess agreement among test runs. The interpretation and validation of repeatability and reproducibility estimates, particularly using natural field samples, are complicated by the fact that no international standards and guidelines are established. Until guidelines are provided, we recommend considering as evidence of acceptable agreements results that show (i) fairly large κ estimates with (ii) a fairly narrow confidence interval obtained from (iii) a medium range prevalence, and (iv) conditional on a non-significant McNemar's test. Repeatability and reproducibility levels and the associated test accuracy appear to vary strongly with the intended use of assay. Appropriate assessment of consistency of test performance is critical to the interpretation of surveillance and control results and requires further development to model agreement across a range of population covariates (e.g. infection prevalences, infection stages).

Acknowledgements

We wish to thank the technical staff and casual employees of the AVC-Centre for Aquatic Health Sciences for the impressive effort and efficacy furnished during the endless sample collection. The authors also wish to thank managers and staff of the participating laboratories for their participation in sample testing. We want to thank Dr. Greenwood for its critical review of the tree reconstruction approach. We greatly appreciate the direct and indirect support of the following agencies and companies: Atlantic Innovation Fund, New Brunswick Department of Agriculture and Aquaculture (Total Development Fund), Fisheries and Oceans Canada (Aquaculture Collaborative Research &

Development Program), and many salmon farms in New Brunswick.

References

Agresti, A., 2002. Categorical Data Analysis. Wiley, New York.

Anonymous, 2000. ISA hits the Faroes. Fish Farming Int. 27, 47.

Begg, C.B., 1987. Biases in the assessment of diagnostic tests. Stat. Med. 6, 411–423.

Bouchard, D.A., Brockway, K., Giray, C., Keleher, W., Merrill, P.L., 2001. First report of infectious salmon anaemia (ISA) in the United States. Bull. Eur. Assoc. Fish Pathol. 21, 86–88.

Byrne, P.J., MacPhee, D.D., Ostland, V.E., Johnson, G., Ferguson, H.W., 1998. Haemorrhagic kidney syndrome of Atlantic salmon, *Salmo salar* L. J. Fish Dis. 21, 81–91.

Cleophas, T.J., Droogendijk, J., van Ouwerkerk, B.M., 2008. Validating diagnostic tests, correct and incorrect methods, new developments. Curr. Clin. Pharmacol. 3, 70–76.

Dohoo, I., Martin, W., Stryhn, H. (Eds.), 2003. Veterinary Epidemiologic Research. AVC Inc., Charlottetown, Canada.

Felsenstein, J., 1985. Confidence-limits on phylogenies—an approach using the bootstrap. Evolution 39, 783–791.

Fleiss, J.L., Levin, B., Paik, M.C., 2003. Statistical Methods for Rates and Proportions, 3rd ed. Wiley, New York, USA.

Godoy, M.G., Aedo, A., Kibenge, M.J., Groman, D.B., Yason, C.V., Grothusen, H., Lisperguer, A., Calbucura, M., Avendaño, F., Imilán, M., Jarpa, M., Kibenge, F.S., 2008. 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. 4, 28.

Greiner, M., Gardner, I.A., 2000. Epidemiologic issues in the validation of veterinary diagnostic tests. Prev. Vet. Med. 42, 2–22.

Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. 41, 95–98.

International Standard 5725-1, 1994. Accuracy (Trueness and Precision) of Measurement Methods and Results. Part 1. General Principles and Definition. International Organisation for Standardisation (ISO), ISO Central Secretariat, 1 rue de Varembé, Case Postale 56, CH—1211, Geneva 20, Switzerland.

Lee, J., Fung, K.P., 1993. Confidence interval of the kappa coefficient by bootstrap resampling. Psychiatry Res. 49, 97–98.

McClure, C.A., Hammell, K.L., Dohoo, I.R., Nerette, P., Hawkins, L.J., 2004. Assessment of infectious salmon anaemia virus prevalence for different groups of farmed Atlantic salmon, *Salmo salar* L., in New Brunswick. J. Fish Dis. 27, 375–383.

Mullins, J.E., Groman, D., Wadowska, D., 1998. Infectious salmon anaemia in salt water Atlantic salmon (*Salmo salar* L.) in New Brunswick, Canada. Bull. Eur. Assoc. Fish Pathol. 18, 110–114.

Nérette, P., Dohoo, I., Hammell, L., Gagné, N., Barbash, P., MacLean, S., Yason, C., 2005. Estimation of the repeatability and reproducibility of three tests for infectious salmon anaemia virus. J. Fish Dis. 28, 101–110.

Nérette, P., Stryhn, H., Dohoo, I., Hammell, L., 2008. Using pseudogold standards and latent-class analysis in combination to evaluate the accuracy of three diagnostic tests. Prev. Vet. Med. 85, 207–225.

Office International des Epizooties, 2008a. OIE Quality Standard and Guidelines for Veterinary Laboratories: Infectious Diseases. Office International des Epizooties (OIE), 12 rue de Prony, 75017 Paris, France, 70 pp.

Office International des Epizooties, 2008b. OIE Aquatic Animal Health Code, 11th ed. Office International des Epizooties (OIE), 12 rue de Prony, 75017 Paris, France, pp. 99–104.

Ransohoff, D.F., Feinstein, A.R., 1978. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N. Engl. J. Med. 17, 926–930.

Rodger, H.D., Turnbull, T., Muir, F., Millar, S., Richards, R., 1998. Infectious salmon anaemia (ISA) in United Kingdom. Bull. Eur. Assoc. Fish Pathol. 18, 115–116.

Stata Base Reference Manual, vol. 2, Q-Z Release10. A Stata Press Publication, Stata Corporation LP, College Station, Texas, USA, 536 pp.

Tamura, K., Dudley, J., Nei, M., Kumar, S., 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24, 1596–1599.

Thorud, K.E., Djupvik, H.O., 1988. Infectious salmon anaemia in Atlantic salmon (*Salmo salar* L.). Bull. Eur. Assoc. Fish Pathol. 8, 109–111.

Van de Peer, Y., Salemi, M., 2003. Phylogeny inference based on distance methods. In: Salemi, M., Vandamme, A.M. (Eds.), Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny. Cambridge University Press, Cambridge, pp. 101–136.

Vandamme, A.M., 2003. Basics concept of molecular evolution. In: Salemi, M., Vandamme, A.M. (Eds.), Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny. Cambridge University Press, Cambridge, pp. 1–23.

Wilson, I.G., 1997. Inhibition and facilitation of nucleic acid amplification. Appl. Environ. Microbiol. 63, 3741–3751.