Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Journal of Virological Methods 270 (2019) 106–112

Contents lists available at ScienceDirect

Journal of Virological Methods

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

PCR diagnostics: *In silico* validation by an automated tool using freely available software programs

Erik van Weezep[a], Engbert A. Kooi[a,1], Piet A. van Rijn[a,b,*]

[a] Department of Virology, Wageningen Bioveterinary Research (WBVR), Lelystad, the Netherlands
[b] Department of Biochemistry, North West University, Potchefstroom, South Africa

ARTICLE INFO

Keywords:
PCR test
Validation
In silico
Automated
Molecular diagnostics
Software

ABSTRACT

PCR diagnostics are often the first line of laboratory diagnostics and are regularly designed to either differentiate between or detect all pathogen variants of a family, genus or species. The ideal PCR test detects all variants of the target pathogen, including newly discovered and emerging variants, while closely related pathogens and their variants should not be detected. This is challenging as pathogens show a high degree of genetic variation due to genetic drift, adaptation and evolution. Therefore, frequent re-evaluation of PCR diagnostics is needed to monitor its usefulness. Validation of PCR diagnostics recognizes three stages, *in silico*, *in vitro* and *in vivo* validation. *In vitro* and *in vivo* testing are usually costly, labour intensive and imply a risk of handling dangerous pathogens. *In silico* validation reduces this burden. *In silico* validation checks primers and probes by comparing their sequences with available nucleotide sequences. In recent years the amount of available sequences has dramatically increased by high throughput and deep sequencing projects. This makes *in silico* validation more informative, but also more computing intensive. To facilitate validation of PCR tests, a software tool named PCRv was developed. PCRv consists of a user friendly graphical user interface and coordinates the use of the software programs ClustalW and SSEARCH in order to perform *in silico* validation of PCR tests of different formats. Use of internal control sequences makes the analysis compliant to laboratory quality control systems. Finally, PCRv generates a validation report that includes an overview as well as a list of detailed results. In-house developed, published and OIE-recommended PCR tests were easily (re-) evaluated by use of PCRv. To demonstrate the power of PCRv, *in silico* validation of several PCR tests are shown and discussed.

1. Introduction

Pathogens exhibit genetic variation as a result of genetic drift, adaptation and evolution, but also by random variation. Since the late nineties of the 20th century, due to the improved sequencing techniques and high throughput sequencing machines, the number of sequences submitted to databases like GenBank® has increased exponentially. This results in an enormous increase of identified variants and quasi-species as well as sequences of newly discovered pathogens from all over the world. A few examples are the discovery of coronaviruses causing Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS), Nipah and Hendra viruses, atypical pestiviruses, atypical and new serotypes of bluetongue virus, Schmallenberg virus and new variants of avian influenza viruses (Chua et al., 2000; Demmler and Ligon, 2003; Drosten et al., 2003; Hoffmann et al., 2012; Hofmann et al., 2008; Maan et al., 2011; Marcacci et al., 2018; Schirrmeier et al., 2004; van Boheemen et al., 2012; Wang, 2011; Zientara et al., 2014).

Currently, in many countries, the first line of pathogen detection is real-time PCR diagnostics. Favourably, PCR tests can be highly sensitive and specific, and are often designed to detect all variants of a defined family, genus or species, while not detecting closely related pathogens. In addition, PCRv can also be used to validate *in silico* PCR assays that differentiate between lineages, serotypes or variants. Therefore, PCR targets must be unique, and highly conserved. Nonetheless, false negative results can arise by genetic drifting or by emergence of new

Abbreviations: ASHV, African horse sickness virus; ASFV, African swine fever virus; BTV, bluetongue virus; CSFV, classical swine fever virus; EAV, equine arteritis virus; EBLV, European bat lyssa virus; EHDV, epizootic haemorrhagic disease virus; FICS, flagged internal control sequences; MSA, multiple sequence alignment; NCBI, national center for biotechnology information; OIE, world organisation for animal health; PCR, polymerase chain reaction; PPRV, peste des petits ruminants virus; PRV, pseudorabies virus; RVFV, rift valley fever virus; SGPV, sheep-and-goat pox virus; WNV, West-Nile virus

* Corresponding author at: Department of Virology, Wageningen Bioveterinary Research (WBVR), Lelystad, the Netherlands.
E-mail addresses: erik.vanweezep@wur.nl (E. van Weezep), ea.kooi@igj.nl (E.A. Kooi), piet.vanrijn@wur.nl (P.A. van Rijn).
[1] present address: Inspectie voor de Gezondheidszorg en Jeugd, Ministerie van Volksgezondheid, Welzijn en Sport.

https://doi.org/10.1016/j.jviromet.2019.05.002
Received 19 March 2019; Received in revised form 18 April 2019; Accepted 11 May 2019
Available online 13 May 2019

variants, while false positive results can be caused by new variants of closely related pathogens. It is therefore important to frequently re-evaluate and, if necessary, redesign PCR tests taking sequences of newly discovered pathogen variants into account.

Validation of PCR diagnostics should be organized in three stages, *in silico*, *in vitro* and *in vivo* validation. *In silico* validation covers the study on inventory of matching and non-matching sequences of the PCR target sequence in a nucleotide database. Matching sequences enable *in silico* sensitivity (detection of all variants), while non-matching sequences support *in silico* specificity (selective detection of variants of the respective group of pathogens). *In vitro* and *in vivo* validation include testing of cultured pathogens, and field samples of defined positive and negative status. *In vitro* and *in vivo* validation for all virus variants is practically impossible and extremely costly. Even more, not every pathogen variant has been cultured or isolated, and transport and handling of pathogens could imply safety issues. In contrast, sequences are rapidly becoming available by high throughput and deep sequencing, even without culturing of pathogens. Therefore, *in silico* re-evaluation of validated PCR diagnostics is and will be an attractive alternative to obtain detailed insight in detection of circulating and (re-) emerging virus variants, and should be frequently executed. It will however become an increasing task due to the rapid increase of available sequences and full genome sequences of numerous species.

We developed a software tool named PCRv to facilitate *in silico* validation of PCR tests entirely based on freely available software programs. PCRv links freely available software programs to automate the whole process, reduces labour, and generates a validation report that includes a brief summary as well as a list of detailed results.

2. Methods

The software tool PCRv is written in the Python programming language. PCRv consists of a user friendly graphical user interface and coordinates the use of software programs ClustalW2.1 (Larkin et al., 2007; Thompson et al., 2002) and SSEARCH (Brenner et al., 1998; Pearson, 1991; Pearson et al., 2017; Smith and Waterman, 1981; Smith et al., 1981) to perform *in silico* validation. PCRv is suitable to determine the *in silico* sensitivity (conservation of sequences) and *in silico* specificity (selectivity) of different PCR formats. To monitor the performance of PCRv, a set of flagged internal control sequences (FICS) are randomly added to the sequence database. PCRv processes data and analyses results, and generates a validation report that includes a summarizing table as well as a list of detailed results for an easy check of potential false positives and false negatives. An overview of all actions executed by PCRv is shown in Fig. 2.

2.1. In silico sensitivity

The sequences of a target organism are downloaded from the National Center for Biotechnology Information (NCBI) database (https://www.ncbi.nlm.nih.gov/nuccore/) by using the respective taxonomy ID number as search query. This guarantees that all available sequences of the defined taxon in the database are downloaded. To generate a multiple sequence alignment (MSA) of these sequences, a full genome sequence was selected as a reference sequence. Genome segments of pathogens with a segmented genome were concatenated to serve as an artificial full length genome. If a full genome sequence was not available, a representative large sequence of the taxon was selected as a reference sequence. A prerequisite is that this partial sequence contained the full target of the PCR test being validated. In order to drastically reduce computing time, pairwise alignments were calculated for each downloaded sequence to the reference sequence by using software program ClustalW 2.1 (Larkin et al., 2007). To correct for orientation errors in the database sequences, alignment in the reverse complement orientation was also attempted. A score was calculated using a scoring scheme as follows: match (+1), mismatch (-2), point deletion or gap (-3), every next adjacent point deletion (-2). The aligned orientation with the highest score was selected. To enable efficient alignment of large sequences, these large sequences were segmented in fragments of 10,000 nucleotides in length and individually aligned to the reference sequence and subsequently combined into one pairwise alignment. PCRv combined all individual pairwise alignments into one multiple sequence alignment (MSA), including the pairwise alignments of primers and probes. The calculation of the MSA was performed by a computer with an Intel® Xeon(R) CPU E5-1650 v2 @ 3.50 GHz processor and 16 Gb of internal computer memory. The regions corresponding to primers and probes were selected from the MSA to construct a conservation plot sorted in decreasing total number of mismatches. The *in silico* sensitivity was expressed as the percentage of hits with a cut-off value of a maximum of one mismatch per primer or probe.

2.2. In silico specificity

The entire nucleotide sequence database (compressed gzip file: nt.gz) was downloaded from the NCBI FTP-website (ftp://ftp.ncbi.nlm.nih.gov/blast/db/) using PCRv. The integrity of the download was confirmed by calculation of the MD5 checksum and subsequent comparison with the checksum published on the FTP-website (file nt.gz.md5). PCRv processed the data stream during download by several optimizations to improve the analysis. Nucleotide code 'N' was replaced by the meaningless code 'Z', which prevents infinite number of hits by the alignment search. The data stream was unpacked and subdivided into multiple fasta formatted text files. Fasta files with a maximum size of 500 MB were sequentially numbered and stored because the NCBI nucleotide database is too large to be analysed all at once. To increase the accuracy of the alignment search (see Discussion), large sequences were fragmented in sequences of maximal 3000 nucleotides with an overlap of 50 nucleotides to prevent the loss of hits of primer or probe sequences spanning the split site. Fragmented sequences were tagged with a unique code allowing reconstruction of the original sequence. Any nucleotide database in fasta format is compatible and could be added. Flagged internal control sequences (FICS) were added to enable validation of the alignment search. FICS consisted of randomly generated sequences of 3000 nucleotides in length containing primer and probe sequences of the PCR test being validated. Primer and probe sequences were inserted in all possible combinations and orientations potentially initiating amplification (Fig. 1). Multiple copies of each combination were inserted with an increasing number of randomly introduced mismatches from 0–10 in each primer and probe sequence (Fig. 1). In total, ten copies of each control sequence per number of mismatches were linearly spread in each 500 MB fasta file. An alignment search was performed with the default expectancy threshold value on all fasta files using primers and probes of the PCR test as search queries and the program SSEARCH available in the FASTA sequence analysis package (Brenner et al., 1998; Pearson, 1991; Pearson et al., 2017; Smith and Waterman, 1981; Smith et al., 1981). PCRv produced a list of hits of the alignment search of all possible primer/probe combinations potentially leading to detectable amplicons. Hits of FICS were stored separately. The percentage of returned hits of control sequences with an increasing number of mismatches was indicative for the sensitivity and accuracy of the alignment search per 500 MB fasta file. The maximum number of returned mismatches in the control sequences was determined by use of the Spearman-Kärber method and demonstrated the validity of the computing process (Wulff et al., 2012). An aborted search caused by an unknown error was visible by the incompleteness of returned FICS. If the accuracy of the alignment search was not acceptable, the alignment search was repeated with a higher expectancy threshold value, which usually resulted in a longer analysis time. The specificity check was limited to a maximum of 5000 nucleotides in amplicon length and up to four mismatches per primer or probe. This limitation was however not applied to the FICS in order to

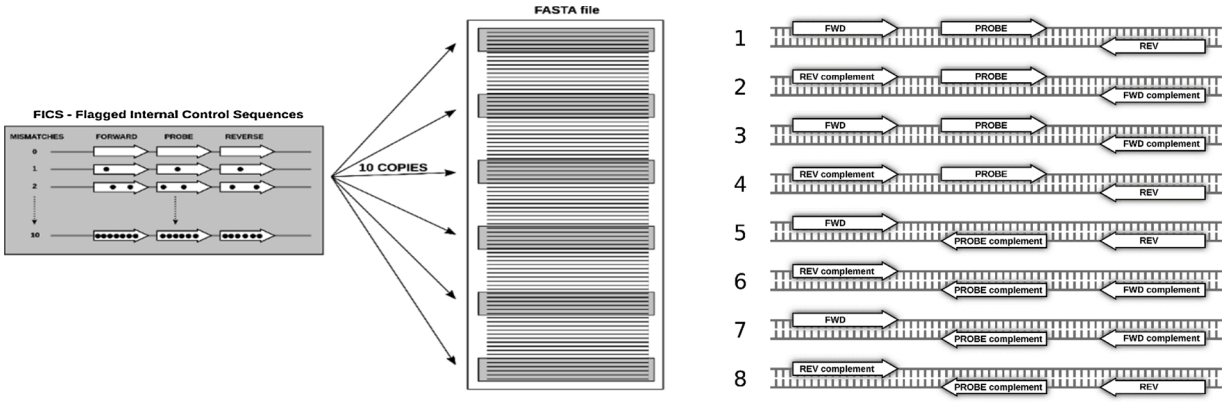

Fig. 1. Flagged internal control sequences (FICS). A) FICS consist of randomly generated sequences of 3000 nucleotides in length containing the primer and probe sequence of the PCR test being validated. Multiple copies were inserted with an increasing number of randomly introduced mismatches from 0–10 in each primer and probe sequence. Ten copies of each FICS per number of mismatches were linearly spread in each 500 MB fasta file. **B)** Overview of all eight possible combinations of positional orientations of forward primer (FWD), reverse (REV) primer and probe used as FICS which are all capable of initiating an (nonspecific) amplification reaction in combination with a detectable probe signal. Combinations of primers and probes according to other PCR formats (*e.g.* nested PCR, PCR using hybridisation probes or hydrolysis probe) are also supported by PCRv but are not shown.

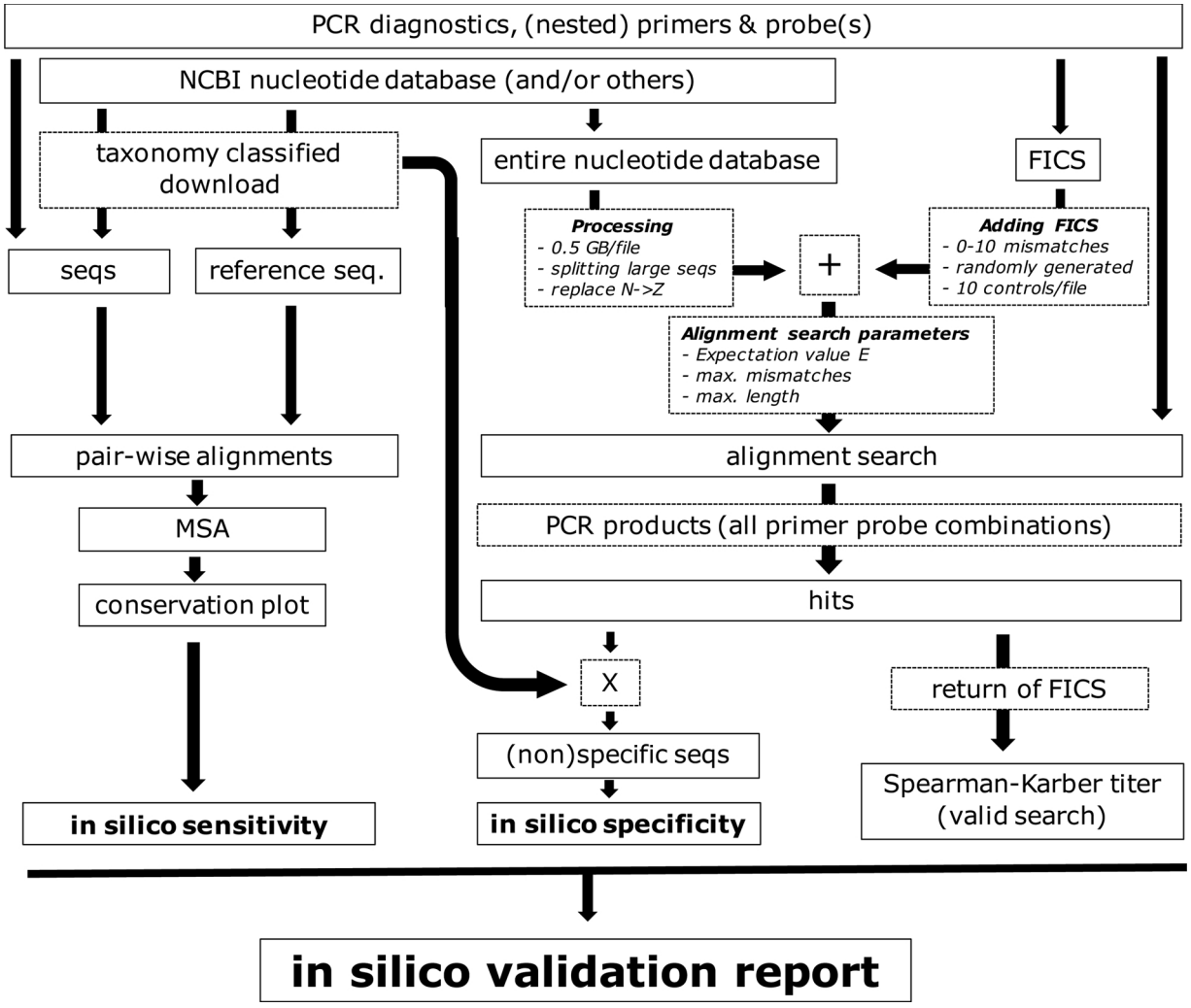

Fig. 2. Schematic overview of the *in silico* validation procedure automated by PCRv. PCRv requires sequences of primers and probes, a recent download of nucleotide sequences from the NCBI database, a selected reference sequence and the taxonomy code. A general overview combined with detailed results are presented in the validation report. FICS: Flagged Internal Control Sequences. MSA: Multiple Sequence Alignment. +: Nucleotide sequences downloaded from the NCBI website/database are combined with FICS for the PCR test being validated. X: The result of the alignment search is divided into specific and nonspecific hits for up to four mismatches per primer or probe. A hit is removed if its accession number is present in the list of downloaded taxonomy classified sequences.

Table 1
Details of OIE recommended PCR tests for West Nile virus. The conventional and nested PCR test (Johnson et al., 2001), and the real time PCR test have been described (Eiden et al., 2010).

PCR format	Primers and probe
conventional	1401F: ACC-AAC-TAC-TGT-GGA-GTC 1845R: TTC-CAT-CTT-CAC-TCT-ACA-CT
real-time	Forward primer: GGG-CCT-TCT-GGT-CGT-GTT-C Reverse primer: GAT-CTT-GGC-YGT-CCA-CCT-C Probe: FAM-CCA-CCC-AGG-AGG-TCC-TTC-GCA-A-BHQ
nested	Outer primers: 1401F: ACC-AAC-TAC-TGT-GGA-GTC 1845R: TTC-CAT-CTT-CAC-TCT-ACA-CT Nested primers: 1485F: GCC-TTC-ATA-CAC-ACT-AAA-G 1732R: CCA-ATG-CTA-TCA-CAG-ACT

fully ascertain the validity of the executed alignment search. Hits were interpreted as specific or nonspecific according to the taxonomy classified sequences as used to generate the MSA. The *in silico* specificity is expressed as the percentage of specific hits of taxonomy classified sequences with a maximum of one mismatch per primer or probe as these are considered to be detected with the respective PCR test.

3. Results

To demonstrate the suitability of our in-house developed software tool PCRv, we determined the *in silico* sensitivity and specificity of three PCR tests for West Nile virus (WNV) recommended by the World Organisation for Animal Health (OIE) (Eiden et al., 2010; Johnson et al., 2001). These WNV PCR tests represented three different formats; a real-time PCR test, a conventional PCR test and a nested PCR test (Table 1).

3.1. In silico sensitivity

Available West Nile virus nucleotide sequences were downloaded from the NCBI website using taxonomy ID 11,082 (search query NCBI:txid11082 on January 15th, 2019). In total, the download contained 20,964 WNV sequences. A MSA was calculated using the full genome sequence with accession number NC_009942 as a reference sequence (Borisevich et al., 2006). Primer and probe sequences were included in the alignment. The calculation of the MSA with PCRv was completed in about 4.5 h. A limited number of 10–15% of the aligned sequences encompassed the locations of primers or probes of the selected OIE-recommended WNV PCR tests. The regions corresponding to primers and probes were taken from the alignment in order to construct a conservation plot. Detailed results were sorted according to the number of mismatches to easily select individual sequences with > 1 mismatch in order to check their origin (Supplemented data A). Note, sequences incorrectly classified as WNV as well as synthetically derived sequences should be discarded as these are irrelevant. Results of the conservation plot were summarized according to the number of mismatches to a maximum of four mismatches per primer or probe (Table 2).The overall *in silico* sensitivity of each PCR test was calculated and expressed as the percentage of sequences with a maximum of one mismatch per primer or probe. The real time PCR test for WNV showed the highest *in silico* sensitivity of 98.8% (83.3% + 15.47%). The conventional and nested PCR tests showed an *in silico* sensitivity of 87.1% and 86.5%, respectively.

3.2. In silico specificity

The entire nucleotide sequence database from the NCBI FTP-website was downloaded as a compressed gzip file (nt.gz) of 502 GB on January 7th, 2019. The download was valid according to the calculated MD5

Table 2
Summary of the *in silico* sensitivity and specificity check of OIE recommended PCR tests for WNV. The *in silico* sensitivity is expressed as the percentage of hits with a maximum of one mismatch per primer or probe. The real time PCR test shows the highest *in silico* sensitivity of 98.77%. The conventional and nested PCR tests show an *in silico* sensitivity of 87.09% and 86.45%, respectively. Found sequences with up to 4 mismatches per primer or probe were classified as specific or nonspecific according to taxonomy number 11,082 for WNV. The *in silico* specificity is expressed as the percentage of specific hits with a maximum of one mismatch per primer or probe. The *in silico* specificity of the real time PCR test is 99.8% ((1783 + 336)/(1783 + 336 + 3 + 2) x 100%). The conventional and nested PCR tests show an *in silico* specificity of 100%, since nonspecific hits with 0 or 1 mismatch were not found. The number of specific hits in the specificity check differs from that of the sensitivity check, see discussion. 1: total number of sequences with ID taxonomy number 11,082 (WNV). 2: number of PCR target sequences found by PCRv. 3: total number of sequences in the downloaded database. 4: mean maximum number of mismatches found in the recovered Flagged Internal Control Sequences (FICS) according to the Spearman-Kärber method. 5: number of hits with indicated mismatches per primer or probe. 6: for the *in silico* specificity, the maximum is 4 mismatches per primer or probe. Note: hits of non-natural sequences were not discarded in the *in silico* specificity check.

real time				
ID taxonomy [1]	20,964			
PCR target [2]	2,204			
database [3]			49,967,663	
FICS [4]			4.1	
nr. of	*in silico* sensitivity		*in silico* specificity	
Mismatches [5]	nr	%	specific	nonspecific
0	1,836	83.30	1783	3
1	341	15.47	336	2
2	6	0.27	3	0
3	7	0.32	1	7
>/= 4 [6]	14	0.64	0	522
Conventional				
ID taxonomy [1]	20,964			
PCR target [2]	3,688			
database [3]			49,967,663	
FICS [4]			4.3	
nr. of	*in silico* sensitivity		*in silico* specificity	
Mismatches [5]	nr	%	specific	nonspecific
0	2,835	76.87	2,460	0
1	377	10.22	317	0
2	76	2.06	18	1
3	69	1.87	19	217
>/= 4 [6]	331	8.98	24	8,033
Nested				
ID taxonomy [1]	20,964			
PCR target [2]	3,704			
database [3]			49,967,663	
FICS [4]			3.7	
nr. of	*in silico* sensitivity		*in silico* specificity	
Mismatches [5]	nr	%	specific	nonspecific

(*continued on next page*)

Table 2 (*continued*)

Nested				
ID taxonomy [1]	20,964			
PCR target [2]	3,704			
database [3]			49,967,663	
FICS [4]			3.7	
nr. of	*in silico* sensitivity		*in silico* specificity	
Mismatches [5]	nr	%	specific	nonspecific
0	2,645	71.41	2,285	0
1	557	15.04	476	0
2	58	1.57	19	0
3	23	0.62	0	0
> / = 4 [6]	421	11.36	0	4

checksum compared to the one published on the FTP-website. The data stream was processed using PCRv as described (Methods). Briefly, downloaded large sequences were fragmented, all sequences were stored in separate fasta formatted text files of 500 MB, and FICS were added. A total of 371 files (215 GB) were stored containing 49,967,663 sequences. An alignment search with primer and probe sequences was performed with a cut-off expectation value E of 5000. The search per PCR test was completed in less than two hours. About 3.7–6.9 million individual primer and probe alignment hits were found and processed by PCRv as described (Fig. 2). FICS were found homogeneously in all 371 database files indicating that the alignment search was completed properly. FICS for each PCR test were returned with a mean of 3.7–4.3 mismatches per primer or probe demonstrating completeness and acceptable accuracy of the alignment search (Table 2). Potential amplicons were interpreted as specific or non-specific according to the presence of its NCBI accession number in the list of sequences as used for the *in silico* sensitivity check (Table 2). We noticed that the number of specific hits differed from the numbers as scored by the *in silico* sensitivity check (Table 2). However, several reasons for this apparent inconsistency can be considered, see Discussion. In summary, using WNV PCR tests as an example, PCRv easily determined the *in silico* sensitivity and specificity of these PCR tests of different formats in a highly automated manner. All results are included in the validation report generated by PCRv, such as a summarizing table of results, conservation plot and a list of nonspecific hits. The summarizing table clearly demonstrates the differences of the *in silico* sensitivity and specificity between these PCR tests (Table 2). In addition, the detailed conservation plot (Supplemented data A) and detailed list of nonspecific hits up to 4 mismatches per primer or probe (Supplemented data B) support manual check of individual sequences on correctness, background, submission details, and other information.

4. Conclusion and discussion

Validation of diagnostics by testing all variants of a target pathogen in cultured or field samples, named *in vitro* and *in vivo* validation, respectively, is hardly feasible. Because of the availability of sequences of pathogens in databases, checking conservation and uniqueness of primer and probe sequences, so-called *in silico* validation, has become an attractive and reliable alternative to (re-) evaluate specificity and sensitivity of molecular diagnostics. Exponential expansion of available sequences, genetic drift of pathogens, and discovery of new pathogens drive the need to frequently validate established PCR tests. This, however, will also become an increasing significant effort. We automated the *in silico* validation process by integrating freely available software programs into a single tool named PCRv. Public databases, such as NCBI as well as other available databases and sequences formatted in single sequence fasta files are compatible with PCRv.

PCRv generates a multiple sequence alignment (MSA) using a selected reference sequence, which is preferably a full length genome but at least a partially large sequence encompassing the PCR target. Software program ClustalW2.1 (Larkin et al., 2007) is used to calculate pair-wise alignments of each sequence to the reference sequence, and subsequently a MSA is generated using these pair-wise alignments. This strategy exponentially reduces calculation time, in particular for large numbers of sequences. Additionally, more than one reference sequence could be used to improve the generation of a MSA in case of extreme variability among a group of pathogens. The MSA is used to determine the *in silico* sensitivity, since this is less prone to mismatches in primers or probes (not shown). For example, sequences with numerous mismatches in one of the primers or probes will not be found by an alignment search using these primer or probe sequences as search queries. However, such sequences will be present in the MSA, see conservation plots of WNV PCR tests. Supplemented data A shows the summarised - without accession numbers - conservation plots of the three WNV PCR tests. PCRv generates a conservation plot listing all hits according to decreasing number of mismatches. Hits with the most mismatches needs attention as these could lead to false negative PCR results. We calculated and defined the *in silico* sensitivity as the percentage of hits with a maximum of one mismatch per primer or probe as these are assumed to be detected with the respective PCR test.

The software program SSEARCH that is available in the FASTA sequence analysis package from the University of Virginia (Pearson, 1991) uses a calculated expectation value E in combination with a supplied threshold value to determine whether a hit is returned. The expectation value E depends on the number and length of sequences in the database. Consequently, the E value of a search hit depends on the location of the found sequence in the database. Large sequences are therefore segmented into fragments of maximal 3000 nucleotides in length. This reduces the variability in sequence length leading to a more homogenous sensitivity of SSEARCH across the database and improves the overall sensitivity of SSEARCH.

The sensitivity of the well-known and commonly used BlastN alignment search program was compared to that of SSEARCH (Fig. 3). Clearly, SSEARCH returns 100% of the primers up to six mismatches. In contrast, the percentage of returns with BlastN is slightly less than 100% for three mismatches and rapidly declines by an increasing number of mismatches. We conclude that SSEARCH is much more accurate, and thus more suitable than BlastN to determine the *in silico*

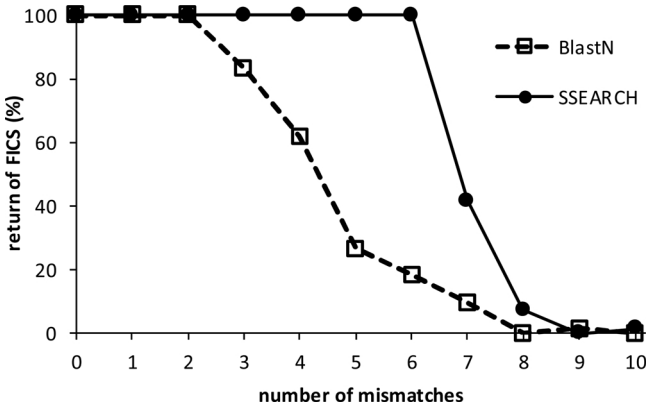

Fig. 3. Comparison of the accuracy of an alignment search performed by the BlastN and the SSEARCH software programs. A test database of randomly generated nucleotide sequences was generated containing 10,000 sequences of 3000 nucleotides in length. 875 sequences contained a primer sequence of 24 nucleotides in length. Each primer contained randomly 0–10 mismatches. The cut-off expectation value E used in both programs was 1000. The inserted primer with up to 2 mismatches completely returned with BlastN, whereas SSEARCH completely returned the primer with up to 6 mismatches.

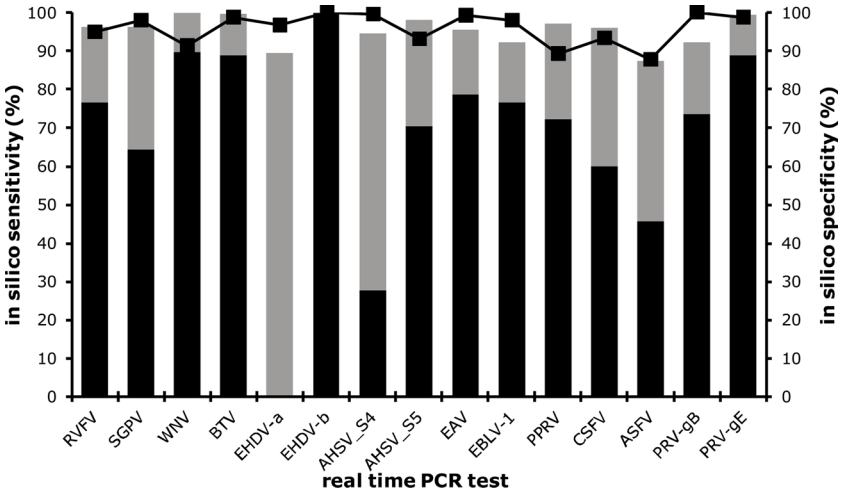

Fig. 4. Overview of the *in silico* sensitivity and specificity of several real time PCR tests at WBVR as determined by PCRv. The *in silico* sensitivity of PCR tests is expressed as the percentage of hits with a maximum of one mismatch per primer or probe (squares, line). The *in silico* specificity is expressed as the percentage of specific hits with 0 mismatches (black) and 1 mismatch per primer of probe (grey). Real time PCR tests are indicated: WNV; West-Nile virus (Eiden et al., 2010; Johnson et al., 2001), BTV; bluetongue virus (van Rijn et al., 2012; 2013), PPRV; peste des petits ruminants virus (van Rijn et al., 2018a), ASHV_S4; African horse sickness virus segment 4 (van Rijn et al., 2018b), ASHV_S5; African horse sickness virus segment 5 (van Rijn et al., 2018b); in-house developed assays: RVFV; Rift Valley fever virus, SGPV; sheep-and-goat pox virus, EHDV-a; epizootic haemorrhagic disease virus test a, EHDV-b; epizootic haemorrhagic disease virus test b, EAV; equine arteritis virus, EBLV-1; European bat lyssa virus type 1, CSFV; classical swine fever virus, ASFV; African swine fever virus, PRV-gB; pseudorabies virus glycoprotein gene gB, PRV-gE; pseudorabies virus glycoprotein gene gE. Results of PCRv could demonstrate the need to optimize or redesign a PCR test, like for EHDV-a and AHSV_S4. Note: hits of non-natural sequences were not discarded. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

specificity. We also noticed that BlastN tends to find partial/fractional nucleotide alignment hits which is not desirable for primers and probes. In addition, PCRv using SSEARCH is suitable for use in a laboratory quality control system, since the search process is monitored per 500 MB fasta file for completeness and accuracy/sensitivity by returned hits of flagged internal control sequences (FICS). An overview of this monitoring is added to the validation report. Examples of incomplete, inaccurate or alignment searches with a low sensitivity are presented (Supplemented data C). In case alignment search results are not sufficient, the threshold value can be changed to increase the sensitivity but the calculation time will also increase.

Here, we showed *in silico* validation results of WNV PCR tests of different formats as an example. PCRv was also used to validate real time PCR tests at WBVR (Fig. 4). SSEARCH quantifies hits for any combination of primers and probes potentially leading to detectable amplicons, see Fig. 2. This can result in more hits for the *in silico* specificity check by SSEARCH than for the *in silico* sensitivity check by ClustalW 2.1. For example, sequences partially overlapping with the PCR target sequence will not be found by the *in silico* specificity check, since this check only finds complete amplicons. Further, NCBI only stores unique nucleotide sequences in its downloadable database export file “nt.gz”. Identical sequences are combined as one sequence with the sequence name as a concatenation of all individual sequence names separated by the ASCII code 1. PCRv does not recognize merged names as multiple sequences, resulting in less hits by SSEARCH.

Detailed analysis of *in silico* validation results enables a focus on specific test problems, as shown for the PCR test for peste-des-petits ruminants virus (PPRV) of WBVR that presumably does not detect PPRV strain Ghana2010 because of three mismatches in the probe sequence. Indeed, the PCR target of this PPRV strain was amplified but was not detected by the Taqman probe (van Rijn et al., 2018a). We used PCRv to analyse OIE-recommended and published PCR tests for other pathogens in order to select the best option for implementation in laboratory diagnostics. Upon preparedness on incursions, frequent *in silico* (re-)validation could also show the need for adaptation of operational PCR tests to emerging epidemics caused by new variants in other parts of the world.

PCRv depends on compatible and reliable nucleotide databases. For example, *in silico* validation by PCRv depends on submission of accurately determined sequences which are coded with the correct taxonomy ID number. For example, classical swine fever virus (CSFV) sequences that are taxonomy classified as bovine viral diarrhoea virus type 2 (BVDV II) were consequently interpreted as false positives in the CSFV PCR test and as false negatives in the BVDV PCR test. Further, in our example of WNV PCR tests, five nonspecific hits appeared to be sequences without taxonomy ID. Still, these sequences are definitely WNV sequences, although 2 out of 5 nonspecific hits have been synthetically derived (Supplemented data B). On the other hand, a more specific taxonomy classification or labelling of sequences in databases could be used for the development of PCR tests specific for subspecies, serotypes or lineages.

Considering the expected rapid expansion of available sequences, PCRv will be further improved by allowing incremental analyses in which only newly submitted sequences with respect to the previously analysed sequences are processed. This will keep the required analysis time manageable for *in silico* re-validation of PCR tests. The number of hits for the *in silico* sensitivity and specificity are not representative for the field situation but represents that of the sequences in the database. In other words, the percentages could be skewed by a small number of sequences in the database, or by a large number of very closely related sequences caused by a huge effort during one epidemic.

Submitted sequences are sometimes not trimmed for synthetic adaptors like PCR primers causing misleading positive analysis results. Synthetic or optimized genes of pathogens can lead to misleading negative PCRv results. Synthetic and genetically modified sequences should be labelled as ‘nonnatural’ in databases to prevent misleading results of *in silico* validation efforts. Finally, negative PCRv results can be created on purpose by development of DIVA (Differentiating Infected from Vaccinated) vaccine viruses with a deleted or mutated DIVA target, like gE deletion mutants of bovine herpes virus type 1 and pseudorabies virus (Kaashoek et al., 1994; van Oirschot et al., 1990), NS3 deletion mutants of bluetongue virus and African horse sickness virus (Feenstra et al., 2014; van Rijn et al., 2018b, 2013), and live-attenuated lumpy skin disease (LSD) vaccine (Agianniotaki et al., 2017).

Viral pathogens belonging to the same taxon showing an extreme variation in their sequence cannot be aggregated in one MSA using one reference sequence. Further, large scale genomic rearrangements, such as duplication, deletion, insertion, inversion, and translocation, are very common in genomes of bacterial pathogens, and will undoubtedly challenge the calculation of a MSA, if this is even possible. Currently, we are investigating alignment-free analysis methods to address these challenges. Even more, we foresee the development of a next generation *in silico* tool, partially based on PCRv, to find highly conserved targets for new or confirmatory PCR tests.

Conflict of interest

All authors declare no conflict of interest.

Acknowledgements

The authors are grateful to colleagues of WBVR, in particular to Jan Boonstra and René van Gennip, for fruitful discussions and suggestions. This research was financially supported by project WOT-01-003-015 of the Dutch ministry of Agriculture, Nature and Food Quality (LNV) (WBVR-project number 1600013-01).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.jviromet.2019.05.002.

References

Agianniotaki, E.I., Chaintoutis, S.C., Haegeman, A., Tasioudi, K.E., De Leeuw, I., Katsoulos, P.D., Sachpatzidis, A., De Clercq, K., Alexandropoulos, T., Polizopoulou, Z.S., Chondrokouki, E.D., Dovas, C.I., 2017. Development and validation of a TaqMan probe-based real-time PCR method for the differentiation of wild type lumpy skin disease virus from vaccine virus strains. J. Virol. Methods 249, 48–57.

Borisevich, V., Seregin, A., Nistler, R., Mutabazi, D., Yamshchikov, V., 2006. Biological properties of chimeric West Nile viruses. Virology 349, 371–381.

Brenner, S.E., Chothia, C., Hubbard, T.J., 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc. Natl. Acad. Sci. U. S. A. 95, 6073–6078.

Chua, K.B., Bellini, W.J., Rota, P.A., Harcourt, B.H., Tamin, A., Lam, S.K., Ksiazek, T.G., Rollin, P.E., Zaki, S.R., Shieh, W., Goldsmith, C.S., Gubler, D.J., Roehrig, J.T., Eaton, B., Gould, A.R., Olson, J., Field, H., Daniels, P., Ling, A.E., Peters, C.J., Anderson, L.J., Mahy, B.W., 2000. Nipah virus: a recently emergent deadly paramyxovirus. Science 288, 1432–1435.

Demmler, G.J., Ligon, B.L., 2003. Severe acute respiratory syndrome (SARS): a review of the history, epidemiology, prevention, and concerns for the future. Semin. Pediatr. Infect. Dis. 14, 240–244.

Drosten, C., Gunther, S., Preiser, W., van der Werf, S., Brodt, H.R., Becker, S., Rabenau, H., Panning, M., Kolesnikova, L., Fouchier, R.A., Berger, A., Burguiere, A.M., Cinatl, J., Eickmann, M., Escriou, N., Grywna, K., Kramme, S., Manuguerra, J.C., Muller, S., Rickerts, V., Sturmer, M., Vieth, S., Klenk, H.D., Osterhaus, A.D., Schmitz, H., Doerr, H.W., 2003. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. N. Engl. J. Med. 348, 1967–1976.

Eiden, M., Vina-Rodriguez, A., Hoffmann, B., Ziegler, U., Groschup, M.H., 2010. Two new real-time quantitative reverse transcription polymerase chain reaction assays with unique target sites for the specific and sensitive detection of lineages 1 and 2 West Nile virus strains. J. Vet. Diagn. Invest. 22, 748–753.

Feenstra, F., Maris-Veldhuis, M., Daus, F.J., Tacken, M.G., Moormann, R.J., van Gennip, R.G., van Rijn, P.A., 2014. VP2-serotyped live-attenuated bluetongue virus without NS3/NS3a expression provides serotype-specific protection and enables DIVA. Vaccine 32, 7108–7114.

Hoffmann, B., Scheuch, M., Hoper, D., Jungblut, R., Holsteg, M., Schirrmeier, H., Eschbaumer, M., Goller, K.V., Wernike, K., Fischer, M., Breithaupt, A., Mettenleiter, T.C., Beer, M., 2012. Novel orthobunyavirus in Cattle, Europe, 2011. Emerg. Infect Dis. 18, 469–472.

Hofmann, M.A., Renzullo, S., Mader, M., Chaignat, V., Worwa, G., Thuer, B., 2008. Genetic characterization of toggenburg orbivirus, a new bluetongue virus, from goats, Switzerland. Emerg. Infect Dis. 14, 1855–1861.

Johnson, D.J., Ostlund, E.N., Pedersen, D.D., Schmitt, B.J., 2001. Detection of North American West Nile virus in animal tissue by a reverse transcription-nested polymerase chain reaction assay. Emerg. Infect Dis. 7, 739–741.

Kaashoek, M.J., Moerman, A., Madic, J., Rijsewijk, F.A., Quak, J., Gielkens, A.L., van Oirschot, J.T., 1994. A conventionally attenuated glycoprotein E-negative strain of bovine herpesvirus type 1 is an efficacious and safe vaccine. Vaccine 12, 439–444.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. Clustal W and clustal X version 2.0. Bioinformatics 23, 2947–2948.

Maan, S., Maan, N.S., Nomikou, K., Batten, C., Antony, F., Belaganahalli, M.N., Samy, A.M., Reda, A.A., Al-Rashid, S.A., El Batel, M., Oura, C.A., Mertens, P.P., 2011. Novel bluetongue virus serotype from Kuwait. Emerg. Infect Dis. 17, 886–889.

Marcacci, M., Sant, S., Mangone, I., Goria, M., Dondo, A., Zoppi, S., van Gennip, R.G.P., Radaelli, M.C., Camma, C., van Rijn, P.A., Savini, G., Lorusso, A., 2018. One after the other: a novel Bluetongue virus strain related to Toggenburg virus detected in the Piedmont region (North-western Italy), extends the panel of novel atypical BTV strains. Transbound. Emerg. Dis. 65, 370–374.

Pearson, W.R., 1991. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. Genomics 11, 635–650.

Pearson, W.R., Li, W., Lopez, R., 2017. Query-seeded iterative sequence similarity searching improves selectivity 5-20-fold. Nucleic Acids Res. 45, e46.

Schirrmeier, H., Strebelow, G., Depner, K., Hoffmann, B., Beer, M., 2004. Genetic and antigenic characterization of an atypical pestivirus isolate, a putative member of a novel pestivirus species. J. Gen. Virol. 85, 3647–3652.

Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. J. Mol. Biol. 147, 195–197.

Smith, T.F., Waterman, M.S., Fitch, W.M., 1981. Comparative biosequence metrics. J. Mol. Evol. 18, 38–46.

Thompson, J.D., Gibson, T.J., Higgins, D.G., 2002. Multiple Sequence Alignment Using ClustalW and ClustalX. Curr Protoc Bioinformatics Chapter 2, Unit 2 3.

van Boheemen, S., de Graaf, M., Lauber, C., Bestebroer, T.M., Raj, V.S., Zaki, A.M., Osterhaus, A.D., Haagmans, B.L., Gorbalenya, A.E., Snijder, E.J., Fouchier, R.A., 2012. Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. MBio 3, 1–9.

van Oirschot, J.T., Gielkens, A.L., Moormann, R.J., Berns, A.J., 1990. Marker vaccines, virus protein-specific antibody assays and the control of Aujeszky's disease. Vet. Microbiol. 23, 85–101.

van Rijn, P.A., Heutink, R.G., Boonstra, J., Kramps, H.A., van Gennip, R.G., 2012. Sustained high-throughput polymerase chain reaction diagnostics during the European epidemic of Bluetongue virus serotype 8. J. Vet. Diagn. Invest. 24, 469–478.

van Rijn, P.A., van de Water, S.G., van Gennip, H.G., 2013. Bluetongue virus with mutated genome segment 10 to differentiate infected from vaccinated animals: a genetic DIVA approach. Vaccine 31, 5005–5008.

van Rijn, P.A., Boonstra, J., van Gennip, H.G.P., 2018a. Recombinant Newcastle disease viruses with targets for PCR diagnostics for rinderpest and peste des petits ruminants. J. Virol. Methods 259, 50–53.

van Rijn, P.A., Maris-Veldhuis, M.A., Boonstra, J., van Gennip, R.G.P., 2018b. Diagnostic DIVA tests accompanying the Disabled Infectious Single Animal (DISA) vaccine platform for African horse sickness. Vaccine 36, 3584–3592.

Wang, L.F., 2011. Discovering novel zoonotic viruses. N. S. W. Public Health Bull. 22, 113–117.

Wulff, N.H., Tzatzaris, M., Young, P.J., 2012. Monte Carlo simulation of the Spearman-Kaerber TCID50. J. Clin. Bioinforma. 2, 5.

Zientara, S., Sailleau, C., Viarouge, C., Hoper, D., Beer, M., Jenckel, M., Hoffmann, B., Romey, A., Bakkali-Kassimi, L., Fablet, A., Vitour, D., Breard, E., 2014. Novel bluetongue virus in goats, Corsica, France, 2014. Emerg. Infect Dis. 20, 2123–2132.