Title: A “pathotyping” multiplex PCR for *Haemophilus parasuis*: a tool for the prediction of virulence

Kate J. Howell¹#* (kjh52@cam.ac.uk), Lucy A. Weinert¹ (lucy.weinert@gmail.com), Sarah E. Peters¹ (sep34@cam.ac.uk), Jinhong Wang¹ (jw401@cam.ac.uk), Juan Hernandez-Garcia¹ (jh937@cam.ac.uk), Roy R. Chaudhuri²
(r.chaudhuri@sheffield.ac.uk), Shi-Lu Luan¹ (shilu.luan@gmail.com), Øystein Angen³ (Oystein.Angen@vetinst.no), Virginia Aragon⁴ (Virginia.Aragon@cresa.uab.cat), Susanna M. Williamson⁵ (Susanna.Williamson@apha.gsi.gov.uk), Paul R. Langford⁶ (p.langford@imperial.ac.uk), Andrew N. Rycroft⁷ (ARycroft@rvc.ac.uk), Brendan W. Wren⁸ (brendan.wren@lshtm.ac.uk), Duncan J. Maskell¹ (djm47@cam.ac.uk) and Alexander W. Tucker¹ (awt1000@cam.ac.uk) on behalf of the BRADP1T Consortium

¹Department of Veterinary Medicine, University of Cambridge, Madingley Road, Cambridge, CB3 0ES, UK.
²Department of Molecular Biology and Biotechnology, University of Sheffield, Firth Court, Western Bank, Sheffield, S10 2TN, UK.
³Norwegian Veterinary Institute, N-0106 Oslo, Norway
⁴Centre de Recerca en Sanitat Animal (CReSA), UAB-IRTA, Campus de la Universitat Autònoma de Barcelona, 08193, Bellaterra, and Institut de Recerca i Tecnologia Agroalimentàries (IRTA), Barcelona, Spain
⁵Animal and Plant Health Agency (APHA), Rougham Hill, Bury St Edmunds, Suffolk IP33 2RX
Abstract – word count 225 (Max 250)

*Haemophilus parasuis* is a diverse bacterial species found in the upper respiratory tract of pigs that can also cause Glässer’s disease and pneumonia. A previous pan-genome study of *H. parasuis* identified 48 genes that were associated with clinical disease. Here we describe the development of a generalised linear model (termed a pathotyping model) to predict the potential virulence of isolates of *H. parasuis*, based on a subset of 10 genes from the pan-genome. A multiplex PCR (mPCR) was constructed based on these genes, the results of which were entered into the pathotyping model to yield a prediction of virulence. This new diagnostic mPCR was tested on 143 field isolates of *H. parasuis* that had previously been whole-genome sequenced and a further 84 isolates from the UK from cases of *H. parasuis*-related disease in pigs collected between 2013 and 2014. The combination of the mPCR and
the pathotyping model predicted the virulence of an isolate with 78% accuracy for the original isolate collection, and 90% for the additional isolate collection, providing an overall accuracy of 83% (81% sensitivity and 93% specificity) when compared with the ‘current standard’ of detailed clinical meta-data. This new pathotyping assay has the potential to aid surveillance and disease control in addition to serotyping data.

1. Introduction

*Haemophilus parasuis* is a diverse Gram-negative bacterial species commonly found as a commensal in the upper respiratory tract (URT) of the pig. Some isolates of this bacterium can cause pneumonia as well as a systemic disease of pigs commonly known as Glässer’s disease, with more severe presentations including arthritis, meningitis, polyserositis and septicaemia (1–5). The UK Animal and Plant Health Agency (APHA) reported an increase in *H. parasuis* related outbreaks over the last few years (6), with fifteen outbreaks involving Glässer’s disease and seven from pneumonia due to *H. parasuis* in 2015. *H. parasuis* can cause disease throughout the lifecycle of the pig, affecting nursery herds (commonly alongside porcine reproductive and respiratory syndrome virus (7)) as well as contributing to multifactorial porcine respiratory disease complex (PRDC) in grower-finisher pigs (8).

Species level identification of *H. parasuis* can be useful diagnostically, but is not sufficient for all situations. For example, the presence of *H. parasuis* in a swab from the nose or from oral fluid is not very informative as this bacterium is a common member of the flora of the URT. Therefore, additional information would be very useful for this bacterium to allow for more thorough surveillance. Several subtyping methods or epidemiological tools are available with serotyping (9–11) being the most commonly used for this bacterium. However, serovar is regarded as a poor proxy for
virulence, with the exception of serovars 4 and 5 (12–14). The vtaA mPCR (15–17) is
the only method that has shown an association between the test result for a given
isolate and its virulence, but its wide scale use has not been reported.

The link between virulence and serovar for *H. parasuis* is predominantly based on
early experimental reproduction of disease using specific-pathogen free (SPF) pigs
using the reference strains (with somewhat inconsistent results) and the global
prevalence of disease causing isolates (1, 9, 13, 18). Serovars 4 and 5 are currently
the most prevalent disease-causing serovars of *H. parasuis* globally (12–14). However
generalisation of the virulence of a serovar is often made for all isolates based on SPF
challenge studies performed with small numbers of isolates and animals (1, 9, 18–21).

Even the serovar 3 reference strain, widely considered to be avirulent, resulted in
some clinical signs in a recent challenge study using colostrum-deprived piglets (22).
Therefore, better markers for virulence or understanding of virulence potential is
required for this bacterium.

Our previous analysis of more than 200 isolates of *H. parasuis* (23) revealed a diverse
pan-genome of over 7,000 genes, with 1,049 genes classified as part of the core
genome (i.e. present in every isolate). However, no link between previously suggested
virulence factors and clinical meta-data was found i.e. no previously suggested
virulence factors were present in a high proportion of isolates from one category
while absent from the other and, consequently, were not useful as indicators of the
virulence of an isolate from the pan-genome. Therefore we used a genome-wide
association study (GWAS) approach (23) using discriminant analysis of principal
components (24), which identified 48 genes associated with disease, that could be
found in either disease-associated or carriage isolates. While the relative importance
in virulence and the functional roles of these newly designated virulence-associated
genes has yet to be investigated, they have the potential to aid in the prediction of the virulence of an isolate based purely on their correlation with the virulent phenotype.

The definition of virulence for our studies has been carefully considered. The isolates used in our previous pan-genome analysis (23) were predominantly field isolates with detailed clinical meta-data based on post-mortem results (e.g. tissue of origin, serovar, age of pig, ante-mortem signs, post-mortem diagnosis, known welfare issues and known co-infections). This information was used to classify isolates of *H. parasuis* obtained from systemic or respiratory sites as disease-associated isolates and all isolates from the URT of healthy pigs as non-disease-associated or ‘carriage’ isolates. In reality this clinical meta-data is not a guaranteed means for determining disease-causing potential as many other factors may be involved in the outcome of an infection (additional host factors and variation in the microbiota) and so we call the clinical meta-data a silver standard dataset for comparison.

This paper describes the design and validation of an mPCR, termed a “pathotyping” tool, which can be used to predict the virulence of an isolate based on a subset of 10 genes from the pan-genome of *H. parasuis*.

2. Methods

2.1. Isolate collection and culture

The pan-genome of *H. parasuis* was previously defined (23) based on a collection of 212 diverse field isolates and reference strains. For ease of description we refer to the previously sequenced *H. parasuis* population as the original isolate collection, containing both the field isolates and the reference strains. In this study we use a subset of this sequenced isolate collection, with the most detailed clinical meta-data
(including tissue of origin, clinical signs from the post-mortem investigation, cause of death) for the development and testing of this mPCR (n=143). This was used as the training set for the mPCR. A summary table of the training set based on geographic location, serovar and virulence is included in Table 1.

An additional 84 disease-associated isolates of *H. parasuis* were subsequently collected by the APHA and this was termed the additional isolate collection. These isolates were used as an outside test set with known virulence but no expected amplicon pattern. Culture and gDNA extraction was performed as described previously (23) for evaluation using the mPCR.

Isolates of closely related commensal *Pasteurellaceae* including *Actinobacillus indolicus* (n=1, European Nucleotide Archive: ERS132160), *Actinobacillus minor* (n=3, ERS132116, ERS132158, ERS132165) and *Actinobacillus porcinus* (n=2, ERS132148, ERS132163) were identified from routine APHA diagnostic investigations and were also genome sequenced. These genome sequences were used to test the specificity of primers designed against *H. parasuis* (using BLASTn) and as part of the negative control panel. In addition, field isolates of the pathogens *Actinobacillus pleuropneumoniae* (n=3), *Bordetella bronchiseptica* (n=1) and *Streptococcus suis* (n=3), were used in the negative control panel for the mPCR.

### 2.2. Pathotyping Multiplex Design

#### 2.2.1. Gene choice

A list of candidate “virulence-associated” genes was constructed from three sources. First, 48 virulence-associated genes identified from a GWAS of the accessory genome (23, 24), including genes that were either positively or negatively associated with virulence. Second, 10 previously suggested virulence factors that were identified in more than 60% of disease-associated isolates and less than 40% of carriage isolates
from the pan-genome analysis (23). Third, Bayesian analysis of population structure (BAPS) (23, 25) of *H. parasuis* had identified five populations, two of which were predominantly of disease-associated isolates (BAPS 4 – 95% and BAPS5 – 98%).

Genes that were found in a large proportion of isolates from these BAPS populations (>75%) were used as the third list (n=251). Only three genes were identified in both the BAPS and GWAS lists.

This list of 306 genes was refined for suitability as markers for a pathotyping mPCR. Genes were ruled out as candidates if the gene was less than 100 bp in length, if the predicted function of the gene was a transposase, phage gene or integrase, if there were no conserved regions in the gene to which primers could be designed, or if the gene was commonly found to have a pseudo-gene in the previously published pan-genome (cut-off <20% of representatives of a given gene were pseudogenes). Finally, a species-specific marker for *H. parasuis* was also included as previously described (26).

### 2.2.2. Model optimisation and statistical analysis

A series of step-wise logistic generalised linear models (in both directions using the MASS R library (26, 27)) was used to build a model using disease association as the phenotype and small subsets of the candidate genes (<20) as independent variables based on the presence and absence of each gene from the pan-genome. The candidates that were significant in the models (p-value<0.1) were taken forward to the second round. As combinations of genes could contribute to virulence, situations whereby up to five genes at a time were interacting were investigated in the series of models. For example the addition of genes A-E with interactions into the model estimated the importance of each gene individually, the importance of all combinations of two
genes, three genes, four genes and finally the combination of all five genes together and whether this improved the model. Finally, these two lists were combined and the model was assessed using step-wise selection again to remove spurious interactions.

The output of the model was a fitted value between 0 and 1. We assessed the performance of the model using the ROCR R package (28) to predict virulence by varying the cut-off on the scale of the fitted values; below the cut-off was considered “carriage” and above was considered “virulent”. The accuracy, sensitivity and specificity at each threshold were then calculated by comparing the predictions to the known clinical category for the original isolate collection. Accuracy was defined as the sum of true positives (TP) and true negatives (TN) divided by all isolates tested. Sensitivity was defined as the TP rate: TP divided by TP and false negatives (FN). Specificity was defined as the TN rate: TN divided by TN and false positives (FP). A receiver operating characteristic (ROC) curve was built using the ROCR R library (28), which is a plot of the sensitivity and specificity for each cut-off, and was used to identify a cut-off for the model that achieved greater than 80% sensitivity and 80% specificity. We also proposed that a buffer zone of +/-0.1 be added to the model cut-off, which we considered to be a category for “potentially virulent” isolates for the interpretation of the user.

2.2.3. Primer design

Primer3 (http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi/) was used to design primers between 21 and 30 bases in length with 40-60% G+C content based on the NEB master-mix guidelines for mPCR design. The primers for each gene were compared to the H. parasuis genomes using BLASTn (word size 7) to identify those that matched the presence/absence pattern of the genes in the original isolate.
Alignments of genes and primers were performed using MEGA(29), which were visually inspected to ensure the best choice of primers given the desired product sizes. An error of up to 10% in the matches of the primers to the *H. parasuis* genomes was allowed from the BLASTn results, allowing for the draft nature of the genome sequences. The primers were then compared (using BLASTn with a word size of 7) to the NCBI non-redundant (nr) nucleotide database and the closely related *Pasteurellaceae* bacterial genomes to check for non-specific primer matches. In addition to the primers for the virulence-associated genes, a previously published species-specific primer pair was added to the mPCR (26). Primers were ordered in dehydrated, de-salted form from Sigma-Aldrich (Haverhill, Cambridge).

### 2.3. One Step mPCR

The PCR master-mix was prepared using OneTaq® Quick-Load® 2X master mix with standard buffer (New England BioLabs) following product specifications and protocols. After optimisation, the final protocol for amplification of the targets was initiation at 94°C for 30 seconds, followed by 30 cycles of denaturation at 94°C for 30 seconds, annealing at 54°C for 30 seconds, and extension at 68°C for 60 seconds followed by a final extension at 68°C for 5 minutes. Each PCR contained 12.5 µl of OneTaq® Quick-Load® 2X master mix, 0.75 µl of DMSO (Sigma-Aldrich) 2.2 µl of the primer mix, 2 µl of genomic DNA (gDNA) for each isolate (at >10 ng/µl) and 7.55 µl UltraPure H₂O (Life Technologies) to a final volume of 25 µl. Ratios of the primers in the primer mix can be found in Table 2. Positive control isolates were chosen based on the presence of the target gene in their genome sequence. Negative control isolates were chosen based on the absence of the target gene in their genome sequence. A negative control of UltraPure H₂O was also used for each PCR performed in this study. Gel electrophoresis was performed using 2.0% agarose gel...
with 5% SybrSafe dye (Invitrogen) and run at 110V for 90 minutes using the Quick-
Load® 100 bp DNA Ladder (New England BioLabs) for all mPCRs. All mPCR
results were visualised using the GelDoc™ Imager (BioRad). The accuracy of the
results for these individual PCRs was compared to the expected results from the
genome sequences and BLAST matches for the primers.

2.4. Validation of the pathotyping mPCR

The mPCR was evaluated using gDNA of 143 isolates from the original collection and
gDNA of 84 isolates of the additional collection (gDNA concentration > 10 ng/µl).
The expected amplicon pattern for each isolate in the original collection was based on
the presence/absence of the 10 genes from their genome sequence (23). After the
mPCR was performed the amplicon pattern was entered into the final model and the
fitted model value was calculated. The accuracy, sensitivity and specificity of the
mPCR results were calculated using the cut-off of 0.72 and the known clinical
category. For the additional isolate collection, the mPCR results were entered into the
model and the output was compared to clinical meta-data. All mPCRs were repeated
on three occasions, each time using a separate master mix, to demonstrate the
repeatability and accuracy of the mPCR.

In order to determine whether the mPCR could detect *H. parasuis* directly from a
single colony, thus not requiring prior gDNA extraction, colony PCR was undertaken
on a subset of the additional isolate collection (n=20). For the colony PCR, a loopful
of bacteria was resuspended in 50 µl of UltraPure H2O, which was heated at 100°C for
30 minutes, and centrifuged at 4,000 x g for 1 minute before the supernatant was used
in the mPCR. The supernatant (2 µl) was used in the mPCR reaction.

2.5. Limit of Detection of the mPCR
The concentration of gDNA was measured using a Qubit fluorometer (Life Technologies) with broad-range standards for five isolates (strain name-serovar: HS145-S1, SW140-S2, Nagasaki-S5, C5-S8, D74-Aus–S9, and IA84/17975-S13). The gDNA was diluted six times in a serial dilution using UltraPure H₂O; this was then used as the template in the mPCR to estimate its limit of detection. The number of copies of the genome per microliter was calculated using an average genome size of 2.26 Mb.

3. Results

3.1. Design of the pathotyping mPCR

The final list of genes included in the model is shown in Table 1 alongside their primers and information regarding their importance in the model. This list included genes that were either positively or negatively associated with virulence and a species-specific marker that was designed based on one of the most conserved genes in the pan-genome of *H. parasuis* (26). Two genes (HPS_23879 and HPS_22976c) that were included in the model were found in high proportions (85% and 78% respectively) in the disease-associated BAPS populations (BAPS 4 & 5)(23). The genes included in the model were mostly of unknown function but included genes encoding an inosine-5-monophosphate dehydrogenase, an aspartate kinase monofunctional class protein, a helix-turn-helix family protein and a glycosyltransferase. This pathotyping model had an Akaike information criterion (estimate of the quality of the model) of 156.5 and the area under the curve (AUC) was 0.88 (Figure 1). The sensitivity and specificity of the model were both estimated
at greater than 80% using the ROC-curve (Figure 1) and led to the choice of a cut-off
of 0.72 for the model.

3.2. Validation of the pathotyping multiplex

The mPCR was validated using 143 of the original *H. parasuis* isolates, with 115
being disease-associated (80%) and 28 being carriage isolates (20%). An example gel
of the mPCR on 16 different isolates from the additional isolate collection can be
found in Figure 2. A summary of the results of the mPCR for these isolates can be
seen in Table 3, and detailed information in Supplemental Table 1. Overall 88% of
these isolates had the amplicon pattern ± one band compared to the expected pattern
from the presence or absence of each gene from the draft genome sequences of the
isolates. For those isolates where more than one amplicon difference was identified by
the mPCR (n=34), the model classified 85% as the same virulence category as the
clinical meta-data. Therefore only 5 isolates varied by more than one band and had a
different virulence prediction from the clinical meta-data. Taking the clinical meta-
data as the ‘current or silver standard’, the sensitivity and specificity of the model and
mPCR were calculated at 75% and 93% respectively, with an overall accuracy of
78%. For the additional clinical isolate collection (Table 3), the amplicon patterns
were entered into the model and 90% were predicted as being disease-associated
isolates.

The amplicon patterns were identical whether gDNA or colony PCR was used. Of the
eleven genes in the mPCR, two were amplified from the negative control panel
(Figure 3), including HPS_21058 from one of the two isolates of *A. porcinus* (HS206)
and HPS_23879 from the single isolate of *A. indolicus* (HS213), but no species-
specific band was produced for any these isolates. Six reference strains of *H. parasuis*
were used to test the limit of detection of the mPCR. The average minimum concentration of DNA detectable by the mPCR was determined to be 1.02 ng/µl for an individual pure gDNA preparation or \(3.4 \times 10^5\) genomes/µl across the isolates (Figure 4).

4. Discussion

We have described the development of a pathotyping mPCR assay exploiting the recent discovery of putative virulence-associated genes based on whole genome analysis of a large population of \(H. parasuis\) with well-characterized clinical origins. Analysis of the pan-genome of the original isolate collection showed that the majority of previously published putative “virulence factors” for \(H. parasuis\) did not show a strong relationship with virulence as they were core genes (i.e. present in all isolates). Furthermore, no single marker of virulence could be identified, emphasizing the necessity for an mPCR approach to any future pathotyping tool. These findings meant that the design of a diagnostic test to predict virulence of isolates would not be straightforward. The decision was therefore made to develop a pathotyping model for \(H. parasuis\) based on a combination of genes found in higher proportions of disease-associated isolates, and genes found in predominantly disease-associated BAPS populations, together with a set of virulence-associated genes from the GWAS analysis of the original isolate collection (23). The resulting mPCR generated a pattern of 11 amplicons including a species-specific marker (26) which, when entered into the model, correctly predicted the allocation of 78% of the original isolate collection to their observed virulence category from our silver-standard clinical meta-data. The majority of the genes targeted by this mPCR have not previously been linked to virulence in \(H. parasuis\) or other bacterial species, an exception being inosine 5-monophosphate dehydrogenase, which has been implicated in the adhesion.
of S. suis serotype 2 (30). Testing of the additional collection of disease-associated isolates resulted in 90% accuracy. Combining the results from both collections gave an overall accuracy of 83%.

There was no obvious pattern based on this mPCR for the prediction of virulence as each gene has its own importance in the model and may have a positive or negative effect. In addition, the amplicon-based outcome was not cumulative, i.e. the number of bands did not relate to the virulence potential and so it is necessary to input the amplicon pattern into the model to interpret the results of the mPCR and predict the virulence potential. Therefore we have developed a user-friendly online tool (https://hps-pathotyping.shinyapps.io/Patho-app) to implement the model, built using R and the Shiny package (31), which requires only the input of a table of the mPCR amplicon pattern. It then provides a fitted value for the amplicon pattern using the model and an associated interpretation of this value into a simple virulence category: virulent, potentially virulent and carriage. The additional category of potentially virulent (based on a buffer zone of ±0.1 on the fitted model value) allows the user to interpret the risk of a particular isolate on a case-by-case basis. The threshold for this buffer zone was chosen to minimise the false positive and false negative rate for our model but in clinical practice there are many additional factors that contribute to the outcome of disease beyond bacterial genotype such as the interplay between host immunity and co-infections with additional pathogens present at the time of challenge. For any approach that only attempts to measure the genotype of the bacterial pathogen it is unrealistic to be able to predict the outcome of disease with 100% accuracy.

We identified a number of discrepancies between expected versus actual amplicon patterns. These could be explained through errors in sequence assemblies as we used...
predominantly draft genome sequences. In addition, the assessment of species-specificity of the mPCR revealed that two of the genes could be identified in single representatives of commensal *Actinobacillus* species. This result was not predicted by the BLAST comparison of the available draft genomes and may have occurred via horizontal gene transfer in these closely related species.

The mPCR was developed from, and designed for use on, DNA extracted from individual bacterial isolates. The presence of multiple isolates of *H. parasuis*, particularly in URT samples (32–34) is one of the major challenges in designing a prospective molecular tool for determining virulence. Current surveillance methods for other pathogens such as *Actinobacillus pleuropneumoniae* operate on an individual isolate level, where the detection of the bacterium initiates a more thorough investigation at the single colony level to look for the presence of toxins (35, 36). We see this pathotyping mPCR as a useful tool for surveillance as a preventative measure. For example, if two herds were about to be mixed then prospective surveillance sampling by nasal swab of these populations might enable the isolation of *H. parasuis* colonies (37), which, in turn, could be tested by mPCR to determine whether different isolates are being carried and if any are likely to cause disease when introduced to a naïve population. This could also be used in conjunction with a recently published mPCR for molecular serotyping (26). Therefore this mPCR approach to the pathotyping of *H. parasuis* presents a major step forward for preventive health programs through its rapidity as well as its specificity.

Ideally, any new research tool or diagnostic test should be compared against an accepted ‘gold standard’ assay. This enables meaningful comparisons of sensitivity, specificity and overall accuracy for the new test. For *H. parasuis* this is challenging for a number of reasons. First, the ideal standard would entail a panel of isolates for
which a series of consistently controlled in vivo challenge experiments had been undertaken using pigs of identical immune status and genetics (1, 9, 13, 18). Both bacterial and host factors impact significantly on the disease outcome for a given isolate and, although in vivo challenge data is available for a number of isolates of H. parasuis, most of this data comes from separately published studies using small sets of isolates and under differing conditions (1, 9, 13, 18). Our approach, which has its caveats, was to take a large collection of field-derived isolates from pigs with or without clinical evidence of H. parasuis-related disease, and to use this for comparative purposes as a silver standard. This data was obtained from detailed post-mortem results including clinical signs, tissue of isolation and the presence of any additional pathogens, and so a certain number of false negatives and false positives were expected based on the unique balance between host immunity versus bacterial virulence potential in each individual clinical context. We were, however, able to perform an in silico comparison of our isolates (23) to the existing vtaA mPCR primers (16) (data not shown) as the leading method for the prediction of virulence to date. This was performed using BLAST, which resulted in 82% sensitivity but only 56% specificity when compared to our clinical meta-data. Our mPCR had similar accuracy (overall accuracy 83%) to the vtaA mPCR (accuracy 80%) when tested on our isolate collection but, although sensitivity was slightly lower, it had far higher specificity and exceeds the usefulness of previous methods such as fingerprinting methods that have limited repeatability between laboratories or usefulness in field situations (38–41).

While our mPCR performed very well a number of false negatives were identified when testing on the original isolate collection. Of these, 59% were linked to co-infections or respiratory disease and since pneumonia in pigs is typically a
multifactorial disease (42), it is likely that these isolates were co-infections. Similarly, for the additional isolate collection, 4 of the 9 (43%) false negatives were isolated from the lung or from animals that had respiratory disease. A far lower proportion of isolates was recorded as false positive from the original isolate collection with only 2 recorded from a total of 143 tested. These isolates were predicted, using our model, as being capable of causing disease, but were isolates from the URT of the pig with no clinical evidence of disease. This phenomenon could be explained by pre-existing passive or active immunity resulting in subclinical carriage, which has been described in the field and in controlled challenge studies using disease-associated isolates (1, 43, 44).

This method was designed using a UK-biased isolate collection (54% of the original isolate collection) from diverse serovars, and will require testing against isolates from a wider range of locations. However, the method offers flexibility and can be easily updated as additional training sets are studied, whether by changing the model cut-off or the addition of new virulence markers or population-based markers that would give further indications of mixed isolates in a sample. Overall, the proposed mPCR is an accurate, sensitive and specific method for predicting the virulence of *H. parasuis*, which could be applied to a variety of samples and with further optimisation could become part of prospective surveillance procedures.
Figure Legends:

Figure 1: ROC curve for the final model based on the ten genes chosen for the pathotyping PCR. The curve is labelled with the different cut-offs for the model to achieve the sensitivity and specificity at each point on the curve. The dashed line represents the desired boundary for 80% specificity and sensitivity to aid the choice of model cut-off.

Figure 2: Results of the pathotyping multiplex on a subset of *H. parasuis* isolates from the additional isolate collection. M - Quick-Load 100 bp DNA Ladder.

Figure 3: Negative control panel showing specificity of the primer sets across a range of commensal and pathogenic bacteria of the pig respiratory tract (*Haemophilus parasuis* positive controls serovar 2 (SW140) and serovar 5 (Nagasaki), *Actinobacillus minor*, *A. porcinus*, *A. indolicus*, *Streptococcus suis*, *A. pleuropneumoniae*, *Bordetella bronchiseptica*). M - Quick-Load 100 bp DNA Ladder.

Figure 4: Limit of detection for the pathotyping multiplex PCR using pure genomic DNA for seven isolates representing all of the bands present in the pathotyping multiplex PCR. (M) represents the Quick-Load 100 bp DNA Ladder.

Table Legends:

Table 1: Summary of isolates used in the validation of the mPCR based on serovar and country of origin

Table 2: Pathotyping genes, including the predicted function, and multiplex primers and estimated product sizes. The genes are numbered according to the pan-genome reference (gene num.) and their respective prevalence amongst disease-associated...
(DA) and carriage (C) isolates from the original collection and category of marker are also shown.

Table 3: Summary table of the mPCR results for the original and additional isolate collections using the model cut-off of 0.72. The table shows the categorisation of the isolates according to available clinical meta-data and according to mPCR. The corresponding rates for true positive (TP), true negatives (TN), false positives (FP) and false negatives (FN) are given

References:


31. RStudio and Inc. 2014. shiny: Web Application Framework for R. 0.10.2.1.


Table 1: Summary of isolates used in the validation of the mPCR based on serovar and country of origin

<table>
<thead>
<tr>
<th>Country</th>
<th>Disease-associated Isolates (Serovars)</th>
<th>Carriage Isolates (Serovars)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 4 or 5 6 7 9 13 14 15 Total</td>
<td>1 2 3 4 or 5 6 7 10 13 14 Total</td>
</tr>
<tr>
<td>Argentina</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Denmark</td>
<td>1 1 6 5 2 3 1 6 1 1 27</td>
<td>1</td>
</tr>
<tr>
<td>Germany</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Greece</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Italy</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Japan</td>
<td>1</td>
<td>1 2 2 1 5</td>
</tr>
<tr>
<td>Spain</td>
<td>1</td>
<td>1 9 1 1 1 2 2 7</td>
</tr>
<tr>
<td>Sweden</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Switzerland</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>UK</td>
<td>2 7 23 11 2 9 2 8 1 3 67 1 2 3 1 2 3 13</td>
<td>1</td>
</tr>
<tr>
<td>Unknown</td>
<td>2</td>
<td>2 1 1 1 3</td>
</tr>
<tr>
<td>USA</td>
<td>1</td>
<td>1 1 3</td>
</tr>
<tr>
<td>Total</td>
<td>6 8 29 27 6 13 3 14 4 5 115 3 4 3 1 3 3 2 1 3 2 28</td>
<td>5 2 8 4 3 4 3 1 3 3 2 1 2 3 2 3 2 2 3 2 28</td>
</tr>
</tbody>
</table>

Table 1: Summary of isolates used in the validation of the mPCR based on serovar and country of origin

<table>
<thead>
<tr>
<th>Country</th>
<th>Disease-associated Isolates (Serovars)</th>
<th>Carriage Isolates (Serovars)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 4 or 5 6 7 9 13 14 15 Total</td>
<td>1 2 3 4 or 5 6 7 10 13 14 Total</td>
</tr>
<tr>
<td>Argentina</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Denmark</td>
<td>1 1 6 5 2 3 1 6 1 1 27</td>
<td>1</td>
</tr>
<tr>
<td>Germany</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Greece</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Italy</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Japan</td>
<td>1</td>
<td>1 2 2 1 5</td>
</tr>
<tr>
<td>Spain</td>
<td>1</td>
<td>1 9 1 1 1 2 2 7</td>
</tr>
<tr>
<td>Sweden</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Switzerland</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>UK</td>
<td>2 7 23 11 2 9 2 8 1 3 67 1 2 3 1 2 3 13</td>
<td>1</td>
</tr>
<tr>
<td>Unknown</td>
<td>2</td>
<td>2 1 1 1 3</td>
</tr>
<tr>
<td>USA</td>
<td>1</td>
<td>1 1 3</td>
</tr>
<tr>
<td>Total</td>
<td>6 8 29 27 6 13 3 14 4 5 115 3 4 3 1 3 3 2 1 3 2 28</td>
<td>5 2 8 4 3 4 3 1 3 3 2 1 2 3 2 3 2 2 3 2 28</td>
</tr>
</tbody>
</table>
Table 2: Pathotyping genes, including the predicted function, and multiplex primers and estimated product sizes. The genes are numbered according to the pan-genome reference (gene num.) and their respective prevalence amongst disease-associated (DA) and carriage (C) isolates from the original collection and category of marker are also shown.

<table>
<thead>
<tr>
<th>Gene Num.</th>
<th>Existing reference</th>
<th>Predicted Function</th>
<th>DA</th>
<th>C</th>
<th>P-value</th>
<th>Marker Forward Primer 5'-3'</th>
<th>Reverse Primer 5'-3'</th>
<th>Primer Ratio</th>
<th>Product Size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPS_210</td>
<td>WP_02111436.0.1</td>
<td>inosine-5-monophosphate dehydrogenase</td>
<td>41%</td>
<td>16%</td>
<td>0.07</td>
<td>Virulent CCGAAGACATAGATCACAATGC</td>
<td>CCACCTGTGTACTGGCTGTGC</td>
<td>1</td>
<td>590</td>
</tr>
<tr>
<td>HPS_210</td>
<td>K056_02725 (AE015787.1)</td>
<td>Unknown function</td>
<td>45%</td>
<td>11%</td>
<td>0.06</td>
<td>Virulent CGTAGCATACGCACACCTAAAG</td>
<td>GAAAGGGCAATAGATACATTTCCG</td>
<td>2</td>
<td>720</td>
</tr>
<tr>
<td>HPS_210</td>
<td>gb</td>
<td>EQA03873.1</td>
<td>aspartate kinase mono-functional class</td>
<td>17%</td>
<td>53%</td>
<td>0.006</td>
<td>Carriage TGATAATGCACAGATGTTGATGCTGC</td>
<td>TATGACTACTCAGAAATGTTGCTGC</td>
<td>2</td>
</tr>
<tr>
<td>HPS_229</td>
<td>WP_02935 (gb</td>
<td>KDB47008.1)</td>
<td>Unknown function</td>
<td>67%</td>
<td>37%</td>
<td>0.79</td>
<td>Virulent CAAGGAAGTGTATTATTGGGAAGAGG</td>
<td>GCTCGATCCACCCCTGAATTTC</td>
<td>1.5</td>
</tr>
<tr>
<td>HPS_230</td>
<td>gb</td>
<td>EQA03144.1</td>
<td>helix-turn-helix family protein</td>
<td>43%</td>
<td>81%</td>
<td>0.000</td>
<td>Carriage CCTGTGATTGAATGGGGTCCTCG</td>
<td>GTGTATTAGATCATACATCCATTAC</td>
<td>2</td>
</tr>
<tr>
<td>HPS_233</td>
<td>HS327_01128 (KEZ222556.1)</td>
<td>Unknown function</td>
<td>26%</td>
<td>60%</td>
<td>0.06</td>
<td>Carriage GGATACAGATCTGATTAAGCCTCCAT</td>
<td>GCAGGTTTCTCTGATTAGCTTTTC</td>
<td>2</td>
<td>120</td>
</tr>
<tr>
<td>HPS_235</td>
<td>HS327_02008 (KEZ17271.1)</td>
<td>Unknown function</td>
<td>30%</td>
<td>76%</td>
<td>0.004</td>
<td>Carriage GAACAGCTAACCAGGAAGGAAATG</td>
<td>TAAGGATATATGCAACTCTCCCGG</td>
<td>1.5</td>
<td>170</td>
</tr>
<tr>
<td>HPS_238</td>
<td>gb</td>
<td>EQA03144.1</td>
<td>Putative glycosyltransferase</td>
<td>13%</td>
<td>9%</td>
<td>0.04</td>
<td>Carriage GATATTAGCTGAAGCTCAAGACA</td>
<td>GCAAAGCGACATAACCTGGCTGATCTC</td>
<td>1.5</td>
</tr>
<tr>
<td>HPS_238</td>
<td>HPS/M011434.1 (EQA00635.1)</td>
<td>Unknown function</td>
<td>29%</td>
<td>48%</td>
<td>0.006</td>
<td>Carriage GGATACAGATCTGATTAAGCCTCCAT</td>
<td>GCAAAGCGACATAACCTGGCTGATCTC</td>
<td>1.5</td>
<td>210</td>
</tr>
<tr>
<td>HPS_229</td>
<td>SVR5_00410 (CDH08961.1)</td>
<td>Unknown function</td>
<td>55%</td>
<td>41%</td>
<td>0.34</td>
<td>Carriage ATTAGACATTGACAGACCCAGGAAGA</td>
<td>TTACCCAAGCGCCTAATGTTGTTT</td>
<td>1</td>
<td>375</td>
</tr>
<tr>
<td>HPS_sp-sp</td>
<td>-</td>
<td>Unknown function</td>
<td>100%</td>
<td>5%</td>
<td>NA</td>
<td>Specie s marker ACAAACGTGAAGTCTATCAGGAG</td>
<td>TAGGCTACTGTTGTAGATATCCACG</td>
<td>1</td>
<td>275</td>
</tr>
</tbody>
</table>
Table 3: Summary table of the mPCR results for the original and additional isolate collections using the model cut-off of 0.72. The table shows the categorisation of the isolates according to available Disease-associated meta-data and according to mPCR. The corresponding rates for true positive (TP), true negatives (TN), false positives (FP) and false negatives (FN) are given.

<table>
<thead>
<tr>
<th>Contingency Table</th>
<th>Disease association from meta-data</th>
<th>Disease-associated</th>
<th>Carriage</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Results by PCR</td>
<td></td>
<td>85 (TP 75%)</td>
<td>2 (FP 7%)</td>
<td>87</td>
</tr>
<tr>
<td>Original isolate collection</td>
<td>Disease-associated</td>
<td>30 (FN 25%)</td>
<td>26 (TN 93%)</td>
<td>56</td>
</tr>
<tr>
<td>Carriage</td>
<td></td>
<td>76 (TP 90%)</td>
<td>0</td>
<td>76</td>
</tr>
<tr>
<td>Additional isolate collection</td>
<td>Disease-associated</td>
<td>8 (FN 10%)</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>199</td>
<td>28</td>
<td>227</td>
</tr>
</tbody>
</table>