Pathotyping the Zoonotic Pathogen *Streptococcus suis*: Novel Genetic Markers to Differentiate Invasive Disease-Associated Isolates from Non-Disease Associated Isolates from England and Wales.

Thomas M. Wileman, [a]# Lucy A. Weinert, [a] Kate J. Howell, [a]* Jinhong Wang, [a] Sarah E. Peters, [a] Susanna M. Williamson, [b] Jerry M. Wells, [c] Andrew N. Rycroft,[d] Brendan W. Wren,[e] Duncan J. Maskell, [a] and Alexander W. Tucker, [a#] on behalf of the BRaDP1T Consortium

[a]Department of Veterinary Medicine, University of Cambridge, Cambridge, UK
[b]Animal and Plant Health Agency (APHA), Bury St Edmunds, UK
[c]Host-Microbe Interactomics, Department of Animal Sciences, Wageningen Univeristy, Wageningen, the Netherlands
[d]The Royal Veterinary College, Hawkshead Campus, Hatfield, United Kingdom
[e]Faculty of Infectious & Tropical Diseases, London School of Hygiene & Tropical Medicine, London, UK

Running Head: Pathotyping *S. suis* Isolates from Pigs on UK Farms

#Address correspondence to Thomas M. Wileman, tmw37@cam.ac.uk;
Alexander W. Tucker, awt1000@cam.ac.uk.
*Present address: Kate J. Howell, Department of Paediatrics, Addenbrooke's Hospital, Cambridge, United Kingdom.

Abstract [limit: 250 words; word count: 248]

Streptococcus suis is one of the most important zoonotic bacterial pathogens of pigs causing significant economic losses to the global swine industry. *S. suis* is also a very successful coloniser of mucosal surfaces and commensal strains can be found in almost all pig populations worldwide, making detection of the *S. suis* species in asymptomatic carrier herds of little practical value in predicting the likelihood of future clinical relevance. The value of future molecular tools for surveillance and preventative health management lies in the detection of strains that genetically have increased potential to cause disease in presently healthy animals. Here we describe the use of genome-wide association studies to identify genetic markers associated with the observed clinical phenotypes i) invasive disease or ii) asymptomatic carriage on the palatine tonsils of pigs on UK farms. Subsequently we designed a multiplex-PCR to target three genetic markers that differentiated 115 *S. suis* isolates into disease-associated and non-disease associated groups; performing with a sensitivity of 0.91, specificity of 0.79, negative predictive value of 0.91, and positive predictive value of 0.79 in comparison to observed clinical phenotypes. We describe evaluation of our pathotyping tool, using an out-of-sample collection of 50 previously uncharacterised *S. suis* isolates, in comparison to existing methods used to characterise and subtype *S. suis* isolates. In doing so, we show our pathotyping approach to be a competitive method to characterise *S. suis* isolates recovered from pigs on UK farms, and one that can easily be updated to incorporate global strain collections.

Introduction

Streptococcus suis (*S. suis*) is one of the most important bacterial pathogens of pigs causing significant economic losses to the swine industry worldwide (1). The infectious agent is responsible for a wide range of clinical manifestations, including septicaemia with sudden death, meningitis, endocarditis, arthritis, and pneumonia amongst other diseases (2). *S. suis* is also a zoonotic pathogen associated with exposure to pigs or pork-derived products (3). Although cases in Europe are infrequently reported, in recent years the surveillance and number of reported human infections has increased substantially in Southeast Asia (4-9).

Importantly, *S. suis* is not only an invasive pathogen but also a very successful coloniser of mucosal surfaces (10). In fact, the upper respiratory tract of pigs, in particular the palatine tonsils, is considered to be both the natural habitat of *S. suis* and a principal route of invasion; although the bacterium can also be recovered from the gastrointestinal and genital tracts (2). Colonisation of adult pigs is common in almost all pig populations sampled, meaning that transfer of *S. suis* from sow to piglet during parturition and suckling is an important route of transmission (10).

Several methods exist to investigate strain diversity and identify phylogenetic groups of *S. suis*. Simple biochemical tests cannot always differentiate *S. suis* from *S. suis*-like strains when performed on cultured isolates recovered from diseased animals, and to date remain of little practical use in differentiating invasive disease-associated strains from asymptomatic commensal-like strains both of which may contribute sub-clinically to the respiratory microflora of colonised pigs (1). Other existing methods used to characterise and subtype *S. suis* as part of epidemiological studies have recently

been the subject of a comprehensive review by Xia *et al*. (11). Each approach has its limitations often requiring either large amounts of sample DNA, which is labour intensive and cumbersome, or high levels of technical competence making the comparison of results between laboratories difficult.

To date, serotyping remains the most widely used method to subtype *S. suis* isolates and is an important part of the routine diagnostic procedure (2, 12). A total of 35 serotypes have been described for *S. suis* based on differences in the capsular polysaccharide antigens, but since their original descriptions evidence now exists for the reclassification of a number of serotypes as other *Streptococcus* species meaning current opinion considers there to be just 29 "true" *S. suis* serotypes (namely 1-19, 21, 23-25, 27-31, and 1/2) (13). Serotype 2 predominates among clinical cases of disease in most countries, although serovars 1-9, 14 & 1/2 have all been documented as being of clinical importance in certain geographical locales (14-18). As a result, serotyping has been used as a proxy for predicting the virulence potential of *S. suis* isolates. However, the use of serotyping alone as a predictor of virulence has the limitation that strains of the same serotype can vary substantially in virulence (19, 20).

Given the limitations of serotyping to reliably predict virulence potential of *S. suis* strains other markers have been investigated. A wide range of homologs of bacterial virulence factors and virulence-associated factors found in other Gram-positive organisms has been shown to affect the virulence of *S. suis* strains through targeted mutagenesis studies (21-23). However, clear association with specific roles in the onset and development of disease has not been found for many proposed factors (24, 25). Despite this, the 'virulence-associated markers' (rather than virulence factors *per se*) extracellular protein factor (EF, encoded by

the *epf* gene) (26) and muramidase-released protein (MRP, encoded by the *mrp* gene) (27), as well as, the thiol-activated toxin hemolysin, suilysin (SLY, encoded by the *sly* gene) (28, 29) have been extensively used to predict the virulence potential of *S. suis* strains in certain mainly European countries, particularly for strains of serotype 2 (17, 24, 30). Unfortunately, genotyping of *epf, mrp*, and/or *sly* also fails to provide clear classification of a *S. suis* isolate as virulent (or not) because isogenic mutants devoid of such factors have been found to be as virulent as their respective parental strains, emphasising the importance of their consideration as virulence associated markers rather than true virulence factors *per se* (31).

Advances in sequencing technologies now allow whole-genome sequencing (WGS) of multiple strains of the same species, including *S. suis* (32-36). This explosion in the amount of detailed genetic information has allowed Bayesian analysis of population structure and the investigation of *S. suis* recombination rates, revealing enormous species diversity and significant genomic differences between *S. suis* isolates responsible for systemic disease in pigs when compared to non-clinical isolates recovered from the upper respiratory tract (35). Indeed, in 2015 Weinert *et al.* proposed loss of protein-encoding sequences had led to a smaller systemic disease-associated genome with increased virulence potential and an overrepresentation of genes encoding previously reported virulence-factors associated with *S. suis* (35).

Minimum core genome (MCG) sequence typing is a recently described typing scheme that also takes advantage of the increase in available *S. suis* WGS data, using population genetics-based sub-divisions for strain identification and typing (33, 37). MCG sequence typing exploits advances in next-generation

sequencing to identify novel regions of the core-genome that can be used to identify and type *S. suis* isolates into "MCG groups" that can later be associated with clinical phenotypes. In fact, during its design, MCG group 1 was reported as being assigned to all highly virulent isolates tested and associated with the greatest occurrence of previously reported virulence genes (33). However, MCG sequence typing like multilocus sequence typing (MLST), also described for *S. suis* (38), is difficult to apply to routine diagnostic testing and can sometimes lack the discriminatory power to differentiate bacterial strains into virulent and avirulent sub-populations, limiting its usefulness in epidemiological studies.

The aim of this study was to design and then evaluate a pathotyping tool to predict the virulence potential of *S. suis* isolates using genome-wide association studies, a so-far unexploited method for the identification of *S. suis* virulence-associated markers. The statistical power to allow the identification of robust associations between genotype and phenotypes including virulence in many different bacterial species is now possible due to the rapid increases in the availability of detailed WGS data (39, 40). Here we have combined WGS data with high-quality clinical metadata in order to identify genetic markers in the *S. suis* accessory genome (i.e. genes absent from one or more isolates or unique to a given isolate) associated with i) invasive disease or ii) asymptomatic carriage on the palatine tonsils of pigs on UK farms. Subsequently, we designed a multiplex-PCR (mPCR) to target three genetic markers that differentiated 115 *S. suis* isolates into i) invasive disease-associated and ii) non-disease associated groups. We also describe evaluation of our pathotyping tool (generalised linear model and mPCR), using an out-of-sample collection of 50 previously uncharacterised *S. suis* isolates, in comparison to existing methods used to

characterise and subtype *S. suis* isolates. In doing so, we show our approach to be a competitive method to subtype *S. suis* isolates recovered from pigs on UK farms, and one that can easily be updated to incorporate global strain collections.

Materials and Methods

Bacterial isolates. Two groups of *S. suis* isolates were used in this study a) a training collection of 115 isolates and b) an out-of-sample test collection of 50 previously uncharacterised isolates. **The original training collection** was used to identify genetic markers which could differentiate *S. suis* isolates into i) invasive disease-associated and ii) non-disease associated phenotypic groups. The 'training' collection consisted of laboratory reference strain P1/7 (NC_012925) originally recovered from an ante-*mortem* blood culture from a pig that died with meningitis in the United Kingdom (32, 41). The other 114 isolates of the training collection were a subset recovered from pigs on farms in England and Wales during routine diagnostic investigations at the Animal Health and Veterinary Laboratories Agency (AHVLA; now the Animal and Plant Health Agency, APHA) in 2010, and contribute to a larger collection previously described in 2015 by Weinert *et al.* (35). Well-defined phenotypic metadata were available based on which each isolate was categorised as being associated with invasive *S. suis* disease (n=53; recovered from systemic sites in the presence of clinical signs (arthritis, meningitis, septicaemia) and/or gross pathology consistent with *S. suis* infection) or as being non-disease associated (n=62; recovered from the tonsil or trachea-bronchus of pigs without any typical signs of streptococcal disease but diagnosed with disease unrelated to *S. suis*, such as enteric disease). **The out-of-sample test collection** was used to evaluate our pathotyping tool. Out-of-sample forecasting is a common approach used to evaluate the performance of binary diagnostic tests. To avoid reducing statistical power, rather than split the training collection, an additional out-of-sample 'test' collection was put together consisting of 23 invasive disease-associated

(recovered from systemic, non-respiratory locations of pigs diagnosed with *S. suis* disease at the APHA during 2013) and 27 non-disease associated isolates (recovered from material scraped from the palatine tonsils of pigs exhibiting no signs of *S. suis* disease on farms in England and Wales between June 2013 and May 2014). Site of recovery, ante-, and post-mortem findings of all isolates described in this study are summarised in Table S1.

Identification of genetic markers associated with observed clinical phenotype. Genetic markers to pathotype *S. suis* were identified using positive detection data of putative protein-encoding sequences making up the *S. suis* accessory-genome (i.e. genes absent from one or more isolates or unique to a given isolate). The accessory genome was taken from Weinert *et al.* (35). Briefly, *de novo* assemblies of Illumina fastq reads were produced, protein-encoding genes were then identified and used in MCL clustering to find orthologue groups, which were manually checked. Two complementary genome-wide association studies i) the univariate Chi-squared test for independence and ii) the multivariate Discriminant Analysis of Principal Components (DAPC) were combined to define a preliminary list of genetic markers associated with the observed clinical phenotypes i) invasive disease or ii) asymptomatic carriage on the palatine tonsils of pigs. The Chi-squared test for independence, implemented in the R package: *stats* (42), was used to compare the observed positive detection of protein-encoding sequences with expected frequencies, in doing so calculating a test statistic that if greater than the critical value was reason enough to reject the null hypothesis of independence (p-value <0.05). Bonferroni adjustment (α/n) was used to control for family-wise error associated with multiple sampling.

DAPC (43, 44), implemented in the R package: *adegenet* (45, 46), was used to identify genetic differences between pre-defined phenotypic groups. The total amount of original variation retained in the DAPC model affected which genetic markers contributed most to the separation of genetic structures. As a result, four independent DAPC analyses were performed retaining 60, 70, 80 or 90% of the original genetic variation, and the 1% of ranked genetic markers contributing most to the discrimination of pre-defined phenotypic groups was then analysed and genetic markers consistently output by two or more DAPC analyses taken forward as candidates for pathotyping *S. suis*.

Analysis of the distribution of previously reported virulence factors associated with *S. suis* disease. Protein-encoding sequences present in P1/7, taken from the list of previously published virulence and virulence-associated factors complied as part of a comprehensive review by Fittipaldi *et al.* (24), were extracted from GenBank (Table S2). P1/7 protein-encoding sequences were used as tBLASTn queries against a bespoke BLAST database consisting of the draft genome assemblies of all isolates described in this study. Amino acid level matches to >80% of >80% of the total length of each translated protein-encoding sequence were considered hits.

Selection of genetic markers to pathotype *S. suis*. Logistic regression analysis in the form of a generalised linear model (GLM) with backwards-stepwise selection using penalised likelihood ratio tests, implemented in the R package: *logistf* (47), was used to identify the fewest statistically significant (p-value <0.05) markers to differentiate *S. suis* isolates into pre-defined i) invasive disease-associated and ii) non-disease associated groups. A receiver operating characteristic (ROC) curve, implemented in the R package: *ROCR* (48), was used

to visualise the GLM performance metrics true positive rate (sensitivity) and false positive rate (1-specificity) in comparison to the observed clinical phenotype (considered to be the 'gold-standard' in this study), and a cutoff threshold selected to convert the real-valued output (fitted values) of the logistic regression (probability of causing invasive disease) into a binary class decision: invasive disease-associated (1)/non-disease associated (0). As no cutoff was optimal according to all possible performance criteria, cutoff choice involved a trade-off between different performance metrics where low false negative rate (1-sensitivity, analogous to Type II error) was chosen as the most valuable performance metric for pathotyping *S. suis*, with a view to establish and then maintain a pig population free of invasive disease-associated strains.

All statistical analyses were performed using the standard R environment for statistical computing and graphs (version 3.1.1) (49).

Identification of *S. suis*-species specific genetic markers. We designed a mPCR to target genetic markers associated with observed clinical phenotype, along with a *S. suis* species-specific marker as a positive control. The most conserved protein-encoding sequences of the *S. suis* core-genome (i.e. present in all isolates) were used to select a species-specific marker to complement the pathotyping markers. To do this, all annotated protein-encoding sequences of *S. suis* strain P1/7 were used as BLASTn queries against a bespoke BLAST database of all *de novo* assemblies and known *S. suis* complete genome sequences. Protein-encoding sequences with identities >95% across >80% of the total length of each query sequence were then used to query the NCBI non-redundant (nr) database to identify matches only to *S. suis*.

Multiplex-PCR and detection of PCR amplicons. The online software, primer3 version 4.0.0 (http://primer3.ut.ee) was used to design mPCR primers. All mPCR primers were designed to target conserved regions within the protein-encoding sequence of genetic markers (as opposed to flanking regions) and are summarised in Table 1. Primers were designed to have similar physical characteristics, enabling simultaneous amplification under the same thermal cycling conditions and in multiplex reactions. Primer length (21-30 bp), GC content (40-60%), melting temperature (>68 ˚C if possible, but at least 60 ˚C), and expected amplicon size (100-1000 bp) were based on the manufacturer's recommendations for primer design using the Multiplex PCR *Plus* kit (Qiagen). Consistency between the positive detection of genetic markers and primer matches was investigated using BLASTn. Prior to ordering, all primers were queried against the NCBI nr nucleotide database to check for non-*S. suis* DNA matches. Primers were synthesised by Sigma-Aldrich (Haverhill, United Kingdom) and delivered in solution (TE buffer; 10 mM Tris-Cl, 1 mM EDTA [pH 8.0]) at a stock concentration of 100 µM; primers were used at a working stock concentration of 20 µM.

All mPCRs were performed using the Multiplex PCR *Plus* Kit (Qiagen), and unless otherwise stated contained the same reagents except for template DNA. The reaction mixture (50 µl) for each mPCR consisted of 25 µl 2x Multiplex PCR Master Mix, 5 µl 10x CoralLoad Dye, 10 µl RNase-free water, 0.2 µM (final concentration) of each primer, and 10 ng template DNA. The three-step thermal cycling program for all reactions was as follows: 95 ˚C for 5 min, followed by 35 cycles of (denaturation) 95 ˚C for 30 s, (annealing) 66 ˚C for 90 s, and (extension)

72 ˚C for 90 s; with a final extension of 68 ˚C for 10 min using a T100 Thermal Cycler (Bio-Rad).

PCR products were analysed by gel electrophoresis using 2% (wt/vol) UltraPure Agarose (Invitrogen) gels made with 1x TBE buffer, and contained 1x SYBR Safe DNA gel stain (Invitrogen). Running time was 60 min at a constant 100 V. Results were visualised using a GelDoc imager (BioRad). Where appropriate, mPCR products were purified using the QIAquick PCR Purification Kit (Qiagen) as per the manufacturer's instructions and Sanger sequenced using the Source Bioscience Lifesciences sequencing service. Returned sequencing data was aligned with reference sequences of the target protein-encoding sequence using CodonCode Aligner software (CodonCode Corporation).

The approximate limit of detection of the mPCR was estimated from 10-fold serial dilutions of *S. suis* genomic DNA of known concentration. DNA extracted from four isolates of the training collection representing invasive disease-associated (SS002 and SS004) and non-disease associated (LSS011 and LSS027) phenotypes/genotypes was mixed in equal quantities so that templates for each mPCR amplicon would be present in all reactions. A series of 10-fold dilutions were then performed to create mPCR templates of decreasing concentration. The limit of detection was considered to be the lowest concentration of template DNA from which all predicted mPCR amplicons, after 35 thermal cycles, were easily visible under UV transillumination.

To evaluate the specificity of the mPCR assay for *S. suis*, field isolates of Streptococcaceae commonly recovered from the upper respiratory tract of pigs on farms in England and Wales were used as a panel of negative controls. The collection included isolates of *Streptococcus gallolyticus*, *Streptococcus orisratti,*

297 *Streptococcus pneumoniae*, and *Streptococcus uberis*, sourced from BBSRC
298 research project: BB/L003902/1. In addition, commensal Pasteurellaceae
299 including *Actinobacillus indolicus, Actinobacillus minor, Actinobacillus porcinus,*
300 and *Haemophilus parasuis* (Nagasaki and SW140) were also included, as well as
301 DNA from an Alcaligenaceae isolate of *Bordetella bronchiseptica* RB50
302 (NC_002927) (50).

303 **Comparison of our pathotyping tool to existing methods used to**
304 **subtype disease-associated isolates of *S. suis*.** To compare our pathotyping
305 tool (GLM and mPCR) to published methods used to subtype disease-associated
306 isolates of *S. suis*, the molecular serotype, virulence-associated gene (*epf*, *mrp*,
307 and *sly*) profile, MLST, and MCG sequence type were all determined *in silico*. For
308 comparison of our pathotyping tool against each existing method the original
309 training collection was used to 'train' a model that was then applied to the out-
310 of-sample 'test' collection.

311 Traditional serotyping (by capillary precipitation) data was unavailable
312 for all *S. suis* isolates described in this study, therefore, molecular 'serotyping'
313 was performed using an adaptation (for *in silico* use) of the mPCR assays
314 described by Liu *et al.* (51). Primer sequences were used as BLASTn queries and
315 nucleotide level matches to >95% of the total length of each primer sequence
316 were considered hits. The distance between hits was compared to reported PCR
317 amplicon sizes. Isolates that could not be assigned to one of the 35 (1-34 & 1/2)
318 originally described *S. suis* serotypes were deemed non-serotypable (NT).
319 Differentiation of molecular 'serotypes' 1 from 14 and 2 from 1/2 was performed
320 using the published method described by Athey *et al.* (52). All isolates, in
321 particular those deemed to be NT, were confirmed to be *S. suis* using a

combination of i) biochemical profile (API 20 Strep), ii) MLST data, and iii) *recN* sequence homology (53).

Virulence-associated gene profiling was performed using an adaptation (for *in silico* use) of the method described by Silva *et al.* (54). Again mPCR and singleplex-PCR primer sequences were used as BLASTn queries and nucleotide level matches to >95% to the total length of each primer sequence were considered hits. The distance between hits compared to reported PCR amplicon sizes. Logistic regression (as described above) using the prevalence of *epf*, *mrp*, and/or *sly* as the GLM explanatory variables was used to classify all isolates as i) invasive disease-associated or ii) non-disease associated.

MLST was performed using the online software MLST version 2.0 (http://cge.cbs.dtu.dk) (55).

MCG sequence typing was performed using an adaptation (for *in silico* use) of the method described by Zheng *et al.* (37). Multiplex-PCR primer sequences were used as BLASTn queries and nucleotide level matches to >95% of the total length of each primer sequence were considered hits. The distance between hits compared to reported mPCR amplicon sizes. Nucleotide sequences between primer sequence matches were then extracted, aligned against the MCG typing reference strain GZ1 (GenBank: CP000837), and the 10 SNPs of interest called allowing isolates to be assigned to one of the seven reported MCG groups for *S. suis*.

McNemar's Chi-squared Test for Count Data, implemented in the R package: *stats* (42), was used to test for statistically significant differences in the sensitivities and specificities of two binary diagnostic tests in a paired study. The Weighted Generalised Score Statistic for Comparison of Predictive Values as

proposed by Kosinski (56), implemented in the R package: *DTComPair* (57), was used to test for significant differences in (negative and positive) predictive values of two binary diagnostic tests.

Results

Design of a pathotyping tool for *S. suis*. Genetic markers to pathotype *S. suis* were identified using positive detection data of 7261 putative protein-encoding sequences making up the *S. suis* accessory-genome (35). To do this, the output of two complementary genome-wide association studies were combined to define a preliminary list of 497 genetic markers associated with the observed clinical phenotypes i) invasive disease or ii) asymptomatic carriage on the palatine tonsils of pigs. A multistep process was used to reduce the preliminary list to a number suitable for logistic regression analysis, retaining genetic markers only if i) positively detected in >50% of invasive disease-associated and <50% of non-disease associated isolates (and vice versa <50% of invasive disease-associated and <50% of non-disease associated isolates; n=88 remaining), ii) protein-encoding sequence length was >500 bp (based on the manufacturer's recommendations for primer design using the Qiagen Multiplex PCR *Plus* kit; n=44 remaining), and iii) not predicted to be a mobile genetic element, such as a phage gene, integrase or transposon (based on Prokka annotations; n=14 remaining). A GLM with backwards-stepwise selection using penalised likelihood ratio tests was then used for the final selection of genetic markers, two associated with invasive disease and one associated with asymptomatic carriage (Table 1). A receiver operating characteristic (ROC) curve was used to visualise the GLM performance metrics true positive rate (sensitivity) and false positive rate (1-specificity), and select the cutoff threshold of 0.43 to be used to convert the real-valued output (fitted values) of the GLM into a binary class decision: invasive disease-associated/non-disease associated (Table S1). In comparison to the observed clinical metadata, considered the 'gold-standard' in this study, our

three genetic markers subtyped the 115 *S. suis* isolates of the training collection with a sensitivity of 0.91, specificity of 0.79, negative predictive value of 0.91, and positive predictive value of 0.79 (Table S3(a)).

At present, WGS is not readily available for routine surveillance studies in veterinary diagnostics laboratories, therefore, we designed a mPCR to target the three genetic markers selected to pathotype *S. suis*. In addition to genetic markers selected to differentiate *S. suis* isolates into i) invasive disease-associated and ii) non-disease associated groups, we also incorporated a *S. suis* species-specific marker into our mPCR assay. To do this, we first identified the most conserved protein-encoding sequences contributing to the *S. suis* core genome (i.e. present in all isolates) and selected SSU0577 as a novel *S. suis* species-specific marker, that had a minimum nucleotide sequence identity of 98.15% across the total length of the 918 bp protein-encoding sequence.

Evaluation of our pathotyping mPCR with the original training collection. Figure 1 shows an example of the mPCR amplicon patterns after gel electrophoresis on a 2% (wt/vol) agarose gel and photographed under UV transillumination. Amplicons of size 722 bp correspond to the *S. suis* species-specific marker (SSU0577), and were produced by all isolates of the training collection irrespective of invasive disease-associated/non-disease associated phenotype or genotype. Other amplicons, of size 211 bp and 347 bp correspond to the invasive disease-associated markers SSU0207 and SSU1589 respectively, and amplicons of size 892 bp correspond to the non-disease associated marker SSUST30534.

To determine the analytical sensitivity of the mPCR the approximate limit of detection was estimated from 10-fold serial dilutions of *S. suis* genomic DNA of

known concentration. The limit of detection was estimated to be ~0.0001 ng of *S. suis* genomic DNA (equivalent to ~45 genome copies), the lowest concentration of template DNA from which all predicted mPCR amplicons, after 35 thermal cycles, were easily visible under UV transillumination (data not shown).

To evaluate the specificity of our mPCR for *S. suis*, field isolates of Streptococcaceae, Pasteurellaceae, and Alcaligenaceae commonly recovered from the upper respiratory tract of pigs on farms in England and Wales were used as a panel of negative controls. No mPCR amplicons, after 35 thermal cycles and gel electrophoresis, were visible under UV transillumination for any of the panel of ten negative controls (data not shown).

Evaluation of our pathotyping tool with an out-of-sample collection. Further evaluation of our pathotyping tool (GLM and mPCR) was done using an out-of-sample test collection of 50 previously uncharacterised (genetically) *S. suis* isolates (23 invasive disease-associated and 27 non-disease associated). Template DNA extracted from each of the 50 isolates produced the 722 bp mPCR amplicon corresponding to the *S. suis* species-specific marker SSU0577. For each isolate, the presence/absence of mPCR amplicons was then input into the GLM and the cutoff threshold of 0.43 applied to the fitted-values to generate the binary classification decision. Table 2(a) summarises the classification of the out-of-sample test collection isolates in comparison to the observed clinical metadata, resulting in a sensitivity of 0.83, specificity of 1.00, negative predictive value of 0.87, and positive predictive value of 1.00.

Comparison of our pathotyping tool to existing methods used to subtype disease-associated isolates of *S. suis*. To compare our pathotyping

tool to the use of serotype as a proxy to predict the virulence potential of *S. suis* isolates, the serotypes most frequently recovered from diseased pigs (1-9, 14 & 1/2) were considered a marker of disease association and all other serotypes considered markers of non-disease association. Table 2(b) summarises the classification of the out-of-sample test collection isolates in comparison to the observed clinical metadata, and shows the use of molecular serotypes 1-9, 14 & 1/2 to predict disease-association performed with a sensitivity of 0.87 (n=3 type II errors), not statistically different from our new mPCR pathotyping tool (McNemar's Chi-squared test for count data *p*-value = 0.31731). Other performance metrics for the molecular serotype-based approach were a significantly worse positive predictive value of 0.77 (weighted generalised score statistic for comparison of predictive values *p*-value = 0.01149) and a significantly worse specificity of 0.78 (n=8 type I errors, McNemar's Chi-squared test for count data *p*-value = 0.01431); no statistically significant difference in negative predictive value was observed (weighted generalised score statistic for comparison of predictive values *p*-value = 0.90553).

To compare our pathotyping tool to the use of *epf*, *mrp* and/or *sly* for the identification of virulent *S. suis* strains, first a GLM was fitted to the prevalence data of these virulence-associated genes in the original 'training' collection of *S. suis* isolates and, using the same selection criteria as previously described for the pathotyping markers, a ROC curve used to select the cutoff of 0.12 to convert the GLM fitted values into a binary class decision. The predict function, implemented in the R package: *logistf* (47), was then used to generate fitted values for the isolates in the out-of-sample test collection (Table S1). Table 2(c) summarises the classification of the out-of-sample test collection isolates as

invasive disease-associated/non-disease associated based on the positive detection of *epf*, *mrp* and/or *sly* in comparison to the observed clinical phenotype. The combined virulence-associated markers performed with a sensitivity of 0.96 (n=1 type II errors), again not statistically different from our new mPCR pathotyping tool (*p*-value = 0.08326). Other performance metrics for the virulence-associated genotyping approach were a significantly worse positive predictive value of 0.46 (*p*-value = $2.97708e^{-7}$; incidentally performing no better than chance (Exact binomial test *p*-value = 1)), and a significantly worse specificity of 0.04 (*p*-value = $3.41417e^{-7}$). The negative predictive value was calculated to be 0.50, worse but not a statistically significant difference (*p*-value = 0.07853).

We compared our pathotyping tool to the use of the King *et al.* MLST scheme (38) as a proxy to predict the virulence potential of *S. suis* isolates. Sequence type (ST) 1 was assigned to 70% of disease-associated isolates and 3% of non-disease associated isolates of the training collection (Table S1). As ST1 is mostly associated with disease in both pigs and humans in Europe (12) we used assignment to ST1 as a binary classifier to indicate disease-association in comparison to the observed clinical metadata. Table 2(d) summarises the classification of the out-of-sample test collection isolates as invasive disease-associated/non-disease associated based on the assignment toST1 in comparison to the observed clinical phenotype. Assignment to ST1 performed with a sensitivity of 0.70 (n=7 type II errors), worse in comparison to our pathotyping tool but not a statistically significant difference (*p*-value = 0.08326). The negative predictive value was calculated to be 0.79, again worse but not a statistically significant difference (*p*-value = 0.08294). Other performance

metrics (specificity and positive predictive value) were found to be identical in comparison to our pathotyping tool.

Finally, we compared our pathotyping tool to the use of the Zheng *et al.* MCG typing scheme (33, 37), one of the most recent typing schemes that exploits advances in next-generation sequencing to identify virulent *S. suis* strains. MCG group 1 was assigned to 77% of disease-associated isolates and 3% of non-disease associated isolates of the training collection (Table S1). Together with the report of MCG group 1 being assigned to all highly virulent isolates tested during design of the typing scheme (33), we used assignment to MCG group 1 as a binary classifier to indicate disease-association; performance in comparison to the observed clinical metadata is summarised in Table 2(e). Assignment to MCG group 1 performed with a sensitivity of 0.78 (n=5 type II errors), again worse in comparison to our pathotyping tool but not a statistically significant difference (p-value = 0.31731). The negative predictive value was calculated to be 0.84, also worse but not a statistically significant difference (p-value = 0.31725). Other performance metrics (specificity and positive predictive value) were found to be identical in comparison to our pathotyping tool.

.

Discussion

We have described the design of a pathotyping tool (GLM and mPCR) exploiting the identification of genetic markers in the *S. suis* accessory-genome (i.e. genes absent from one or more isolates or unique to a given isolate) associated with the observed clinical phenotypes i) invasive disease or ii) asymptomatic carriage on the palatine tonsils of pigs on UK farms. Initial analyses of the original training collection were unable to identify any single genetic marker of invasive disease prevalent in >95% of invasive disease-associated isolates and not positively identifiable in <5% of non-disease associated isolates. Furthermore, we found over half (n=40) of published putative "virulence-factors", extracted from the previous comprehensive review by Fittipaldi *et al.* (24) and present in P1.7, did not show a strong relationship with observed clinical phenotype as they were either i) positively detected in the *S. suis* core-genome (i.e. prevalent in all isolates; n=38) or ii) not detected by our methods in any of the 115 isolates of the training collection (n=2; data not shown). The reason for this is unclear, although could be an effect of previous studies being limited to small numbers of isolates often restricted to serotype 2 (58), and of varied and inconsistent animal models between research groups (25).

To avoid restricting our analyses to previously published reports and not taking full advantage of the statistical power of our WGS data set, we used two complementary genome-wide association studies and then logistic regression analysis for the final selection of genetic markers to pathotype *S. suis*. Using logistic regression analysis also allowed for the possibility that multiple genetic markers might best describe the *S. suis* pathotype. Our pathotyping markers assigned the 115 *S. suis* isolates of the original training collection to phenotypic

groups (disease-associated/non-disease associated) with a sensitivity of 0.91 i.e. the proportion of isolates recovered from systemic sites and predicted to be disease causing isolates. A specificity of 0.79 i.e. the proportion of isolates recovered from the upper respiratory tract of pigs without any typical signs of *S. suis* infection and predicted to be non-disease associated isolates. A negative predictive value of 0.91 i.e. the proportion of isolates predicted to be non-disease associated that were actually recovered from the upper respiratory tract of pigs without any typical signs of *S. suis* infection. As well as, a positive predictive value of 0.79 i.e. the proportion of isolates predicted to be associated with invasive disease that were actually recovered from a systemic site.

An important caveat of our pathotyping tool design is consideration of the observed clinical phenotype associated with each isolate as the 'gold standard' to characterise *S. suis* isolates as disease-associated or non-disease associated. In the absence of an agreed superior approach, clinical metadata was used to assign *S. suis* isolates to one of two phenotypic groups and it is acknowledged that such an approach is not perfect as not all additional factors can be accounted for, such as host-immune status, concurrent infections, or environmental conditions that could influence the susceptibility of a host to *S. suis*-associated disease. Indeed, reports of *in vivo* challenge studies can be readily found in the *S. suis* literature, although most describe data limited to a small number of isolates, often restricted to serotype 2 (58), and under very different conditions making the extrapolation of findings difficult to interpret. An ideal standard would require an agreed panel of isolates for which a series of consistently controlled experimental infection challenge studies had been undertaken using pigs of identical immune status and genetics. However, in order for this to happen

543 experts in the field must first agree on a suitable model and set of well-defined
544 criteria to score virulence (25, 59, 60).

545 Another important caveat of our pathotyping tool design is the source of
546 *S. suis* isolates of the original training collection that were deemed to be non-
547 disease associated. While all efforts were made to accurately define invasive
548 disease-associated and non-disease associated phenotypic groups it should be
549 acknowledged that non-disease associated isolates of the original training
550 collection were recovered from routine submissions to the APHA in 2010 and
551 that these pigs were not healthy, even though they did not show signs of typical
552 streptococcal disease; instead clinical features were consistent with different
553 non-infectious diseases or disease caused by other non-*S. suis* infectious agents.
554 Indeed, 13 isolates of the original training collection deemed to be non-disease
555 associated by phenotype were predicted by our pathotyping tool to have the
556 potential to cause invasive disease. These 13 type I errors (or 'false' positives) in
557 comparison to the observed clinical metadata could in fact be true predictions
558 and examples of *S. suis* strains with the potential to cause invasive disease being
559 carried in the upper respiratory tract of pigs on UK farms. Therefore, it is
560 possible that the mortality of these 13 pigs was due to clones of isolates
561 recovered from the palatine tonsils or trachea-bronchus yet was not identified as
562 so due to a concurrent or opportunistic infection presenting a more obvious
563 phenotype, such as diarrhoea. Such an observation is supported by evidence in
564 the literature reporting that virulent strains of *S. suis* can be isolated from the
565 tonsils of pigs without obvious streptococcal disease (61, 62), which is likely to
566 represent carriage of invasive disease-causing stains by pigs that have mounted
567 an effective immune response.

We deemed the false negative rate (1-sensitivity) to be the most valuable performance metric for a *S. suis* pathotyping tool in order to establish and maintain a pig population free of invasive disease-associated *S. suis* strains. During out-of-sample testing the false negative rate of 0.17 corresponded to four false negatives (or type II errors), where non-disease associated pathotyping tool predictions were made for isolates linked with invasive disease clinical metadata. It is interesting to speculate at the reasons for such observations. Often *S. suis* strains are described as opportunistic or secondary pathogens that without a weakened host immune status (due to stress or concurrent infection) would normally be carried asymptomatically, contributing to the normal oral microflora of pigs. This may be the explanation for the differences observed between our pathotyping tool prediction and the observed clinical phenotype, again emphasising the fallibility of the phenotype assigned when it is based on field sampling without carefully controlled infection challenge data.

Comparison to published methods revealed our molecular pathotyping tool to be a competitive method to subtype *S. suis* isolates, even though the necessarily small number of clinically phenotyped isolates in the out-of-sample collection limited the statistical power of the comparison. Comparing the commonly used performance metrics sensitivity, specificity, negative predictive value, and positive predictive value we found the use of i) serotypes 1-9, 14 & 1/2, ii) a GLM based on the positive detection of virulence-associated markers *epf*, *mrp* and/or *sly*, iii) assignment to MLST 1, and iv) assignment to MCG group 1 performed with statistically similar sensitivities in comparison to our pathotyping tool. However, the trade off for high sensitivities was significantly worse specificities and negative predictive values when using serotypes 1-9, 14

& 1/2 or the virulence-associated markers: *epf*, *mrp* and/or *sly*, in certain cases performing no better than chance (*p*-value =1) in comparison to our pathotyping tool. Over all, the performance of our pathotyping tool was at least statistically similar and competitive with, and in some cases, better than previously described methods for assessing the clinical significance of *S. suis* isolates. Similarly, performance of our newly proposed *S. suis* species-specific marker (SSU0577) was encouraging. An important part of our pathotyping tool, due to the presence of *S. suis*-like organisms such as *Streptococcus orisratti* in the pig upper respiratory tract, we acknowledge that the specificity of SSU0577 for *S. suis* and not *S. suis*-like organisms needs to be extended and studied further against markers such as *recN*.

At present the role in pathogenesis of our newly defined pathotyping markers is unknown. Based on predicted biological functions (Table 1) we speculate that marker SSU0207, predicted to be a copper exporting ATPase, might allow *S. suis* to avoid copper toxicity inside phagocytes as copper homeostasis has been shown to be important in many bacterial species (63-65). The marker SSU1589 is annotated as a Type I restriction-modification (RM) system S protein in *S. suis* strain P1/7. Ubiquitous among prokaryotes, Type I RM systems are large multifunctional protein complexes thought to defend host bacterium from foreign DNA borne by bacteriophages, and have recently been described in P1/7 and *S. suis* strains isolated in the Netherlands (66, 67). Considered primitive immune systems in bacteria, it has been proposed that the range of functions RM systems may have should be expanded to include stabilising mobile genetic elements or gene regulation, potentially providing evolutionary fitness advantages and virulence under certain conditions (68).

618 Indeed, the proposed role in protection against foreign DNA may merely be a
619 coincidental benefit of these functions (69). In fact, a Type I RM system in
620 *Streptococcus pneumoniae* which can undergo genetic recombination with
621 truncated variants of the same gene to generate alternative variants with
622 different methylation specificities could control global changes in gene
623 expression (70). In *Streptococcus pneumoniae* there is a selection for variants of
624 this genetic switching *in vivo*, indicating a role in systemic disease.

625 Our third genetic marker (SSUST30534), a putative sugar ABC
626 transporter, was positively associated with the non-disease associated
627 phenotype (asymptomatic commensal-like carriage). The practical application of
628 the genetic marker positively associated with asymptomatic carriage might not
629 be immediately obvious but its statistical significance in the GLM is noteworthy.
630 Indeed, gene loss (of so-called 'antivirulence genes') in the evolution of bacterial
631 pathogens from non-pathogenic commensal strains could be a mechanism of fine
632 tuning pathogen genomes for maximal fitness in new host environments; in
633 short when regulation of invasion, replication and transmission processes is
634 altered, virulence can emerge (71). Indeed, genome reduction via gene loss and
635 pseudogenisation associated with enhanced pathogenicity has been described in
636 other bacteria, such as *Rickettsia* spp., *Shigella* spp. and *Yersinia* spp. (71).
637 Genome reduction through the loss of genes, potentially interfering with host
638 infection, has also been proposed in *S. suis* (35). Therefore, as the elimination of
639 the genetic marker associated with asymptomatic carriage from the GLM could
640 not be done without a statistically significant loss of fit it was retained and its
641 usefulness evaluated.

642 In conclusion, we foresee a useful clinical application of our pathotyping
643 tool in preventative programs aimed at monitoring the health status of pigs and
644 identification of subclinical carriers of invasive disease-associated *S. suis* strains
645 in the upper respiratory tract. Our approach can easily be updated to incorporate
646 global strain collections (such as, from North America and Southeast Asia) to
647 identify geographically-dependent phenotypes. This could contribute to a lower
648 prevalence of disease attributed to *S. suis* among pig populations and
649 consequently a reduction in the usage of antibiotics in the swine industry, as well
650 as a reduction in zoonotic transmission of this pathogen through improved
651 surveillance of pig populations.
652

Acknowledgements

We acknowledge Dr Trevelyan J. McKinley for helpful conversations and assistance with statistical analyses. We also thank Dr Andrew Preston for the genomic DNA of *Bordetella bronchiseptica* reference strain RB50, and Brian Hunt and Jon Rogers of the APHA for the collection of *S. suis* isolates.

This work was supported by a Biotechnology and Biological Sciences Research Council (BBSRC) Knowledge Transfer Network CASE studentship co-funded by Zoetis (previously Pfizer Animal Health UK) and with significant contribution from BQP Ltd (Award Reference: BB/L502479/1). Funding bodies provided scholarship support but had no part in study design, data collection, analysis and interpretation of data or in writing the manuscript. AWT is supported by a BBSRC Longer and Larger (LoLa) grant (Award Reference: BB/G019274/1). LAW is supported by a Dorothy Hodgkin Fellowship funded by the Royal Society (Grant Number: DH140195) and a Sir Henry Dale Fellowship co-funded by the Royal Society and Wellcome Trust (Grant Number: 109385/Z/15/Z).

Members of the Bacterial Respiratory Diseases of Pigs-1 Technology (BRaDP1T) Consortium include: (Imperial College London) Janine T. Bossé, Paul R. Langford, Yanwen Li; (London School of Hygiene & Tropical Medicine) Jon Cuccui, Vanessa S. Terra, Brendan W. Wren; (The Royal Veterinary College) Jessica Beddow, Gareth A. Maglennon, Andrew N. Rycroft; (University of Cambridge) Roy R. Chaudhuri, Duncan J. Maskell, Sarah E. Peters, Alexander W. Tucker, Jinhong Wang, and Lucy A. Weinert.

References

1. **Staats JJ, Feder I, Okwumabua O, Chengappa MM.** 1997. Streptococcus suis: past and present. Vet Res Commun **21:**381-407.
2. **Gottschalk M.** 2012. Streptococcosis, p 841-855. *In* Zimmerman J, Karriker L, Ramirez A, Schwartz K, Stevenson G (ed), Diseases of Swine, Tenth Edition ed. John Wiley & Sons, Inc.
3. **Dutkiewicz J, Zajac V, Sroka J, Wasinski B, Cisak E, Sawczyn A, Kloc A, Wojcik-Fatla A.** 2018. Streptococcus suis: a re-emerging pathogen associated with occupational exposure to pigs or pork products. Part II - Pathogenesis. Ann Agric Environ Med **25:**186-203.
4. **Perch B, Kristjansen P, Skadhauge K.** 1968. Group R streptococci pathogenic for man. Two cases of meningitis and one fatal case of sepsis. Acta Pathol Microbiol Scand **74:**69-76.
5. **Gottschalk M, Xu J, Calzas C, Segura M.** 2010. Streptococcus suis: a new emerging or an old neglected zoonotic pathogen? Future Microbiol **5:**371-391.
6. **Mai NT, Hoa NT, Nga TV, Linh le D, Chau TT, Sinh DX, Phu NH, Chuong LV, Diep TS, Campbell J, Nghia HD, Minh TN, Chau NV, de Jong MD, Chinh NT, Hien TT, Farrar J, Schultsz C.** 2008. Streptococcus suis meningitis in adults in Vietnam. Clin Infect Dis **46:**659-667.
7. **Hoa NT, Chieu TT, Nghia HD, Mai NT, Anh PH, Wolbers M, Baker S, Campbell JI, Chau NV, Hien TT, Farrar J, Schultsz C.** 2011. The antimicrobial resistance patterns and associated determinants in Streptococcus suis isolated from humans in southern Vietnam, 1997-2008. BMC Infect Dis **11:**6.

8. **Praphasiri P, Owusu JT, Thammathitiwat S, Ditsungnoen D, Boonmongkon P, Sangwichian O, Prasert K, Srihapanya S, Sornwong K, Kerdsin A, Dejsirilert S, Baggett HC, Olsen SJ.** 2015. Streptococcus suis infection in hospitalized patients, Nakhon Phanom Province, Thailand. Emerg Infect Dis **21:**345-348.

9. **Chen FL, Hsueh PR, Ou TY, Hsieh TC, Lee WS.** 2016. A cluster of Streptococcus suis meningitis in a family who traveled to Taiwan from Southern Vietnam. J Microbiol Immunol Infect **49:**468-469.

10. **Clifton-Hadley FA, Alexander TJ.** 1980. The carrier site and carrier rate of Streptococcus suis type II in pigs. Vet Rec **107:**40-41.

11. **Xia X, Wang X, Wei X, Jiang J, Hu J.** 2018. Methods for the detection and characterization of Streptococcus suis: from conventional bacterial culture methods to immunosensors. Antonie Van Leeuwenhoek doi:10.1007/s10482-018-1116-7.

12. **Goyette-Desjardins G, Auger J-P, Xu J, Segura M, Gottschalk M.** 2014. Streptococcus suis, an important pig pathogen and emerging zoonotic agent—an update on the worldwide distribution based on serotyping and sequence typing. Emerging Microbes & Infections **3:**e45.

13. **Kerdsin A, Akeda Y, Hatrongjit R, Detchawna U, Sekizaki T, Hamada S, Gottschalk M, Oishi K.** 2014. Streptococcus suis serotyping by a new multiplex PCR. J Med Microbiol **63:**824-830.

14. **Fittipaldi N, Fuller TE, Teel JF, Wilson TL, Wolfram TJ, Lowery DE, Gottschalk M.** 2009. Serotype distribution and production of muramidase-released protein, extracellular factor and suilysin by field

Journal of Clinical Microbiology

725 strains of Streptococcus suis isolated in the United States. Vet Microbiol
726 **139:**310-317.
727 15. **Messier S, Lacouture S, Gottschalk M, Groupe de Recherche sur les**
728 **Maladies Infectieuses d, Porc, Centre de Recherche en Infectiologie**
729 **P.** 2008. Distribution of Streptococcus suis capsular types from 2001 to
730 2007. Can Vet J **49:**461-462.
731 16. **Heath PJ, Hunt BW.** 2001. Streptococcus suis serotypes 3 to 28
732 associated with disease in pigs. Veterinary Record **148:**207-208.
733 17. **Wisselink HJ, Smith HE, Stockhofe-Zurwieden N, Peperkamp K, Vecht**
734 **U.** 2000. Distribution of capsular types and production of muramidase-
735 released protein (MRP) and extracellular factor (EF) of Streptococcus suis
736 strains isolated from diseased pigs in seven European countries. Vet
737 Microbiol **74:**237-248.
738 18. **Perch B, Pedersen KB, Henrichsen J.** 1983. Serology of capsulated
739 streptococci pathogenic for pigs: six new serotypes of Streptococcus suis.
740 J Clin Microbiol **17:**993-996.
741 19. **Vecht U, Arends JP, Vandermolen EJ, Vanleengoed LAMG.** 1989.
742 Differences in Virulence between 2 Strains of Streptococcus-Suis Type-II
743 after Experimentally Induced Infection of Newborn Germ-Free Pigs.
744 American Journal of Veterinary Research **50:**1037-1043.
745 20. **Vecht U, Wisselink HJ, van Dijk JE, Smith HE.** 1992. Virulence of
746 Streptococcus suis type 2 strains in newborn germfree pigs depends on
747 phenotype. Infect Immun **60:**550-556.

21. **Si Y, Yuan F, Chang H, Liu X, Li H, Cai K, Xu Z, Huang Q, Bei W, Chen H.** 2009. Contribution of glutamine synthetase to the virulence of Streptococcus suis serotype 2. Vet Microbiol **139:**80-88.

22. **Baums CG, Kaim U, Fulde M, Ramachandran G, Goethe R, Valentin-Weigand P.** 2006. Identification of a novel virulence determinant with serum opacification activity in Streptococcus suis. Infect Immun **74:**6154-6162.

23. **Zhang H, Fan H, Lu C.** 2010. Identification of a novel virulence-related gene in Streptococcus suis type 2 strains. Curr Microbiol **61:**494-499.

24. **Fittipaldi N, Segura M, Grenier D, Gottschalk M.** 2012. Virulence factors involved in the pathogenesis of the infection caused by the swine pathogen and zoonotic agent Streptococcus suis. Future Microbiol **7:**259-279.

25. **Segura M, Fittipaldi N, Calzas C, Gottschalk M.** 2017. Critical Streptococcus suis Virulence Factors: Are They All Really Critical? Trends Microbiol **25:**585-599.

26. **Smith HE, Reek FH, Vecht U, Gielkens AL, Smits MA.** 1993. Repeats in an extracellular protein of weakly pathogenic strains of Streptococcus suis type 2 are absent in pathogenic strains. Infect Immun **61:**3318-3326.

27. **Vecht U, Wisselink HJ, Jellema ML, Smith HE.** 1991. Identification of Two Proteins Associated with Virulence of Streptococcus suis Type 2. Infection and Immunity **59:**3156-3162.

28. **Jacobs AA, Loeffen PL, van den Berg AJ, Storm PK.** 1994. Identification, purification, and characterization of a thiol-activated hemolysin (suilysin) of Streptococcus suis. Infect Immun **62:**1742-1748.

Journal of Clinical Microbiology

29. **Jacobs AAC, Vandenberg AJG, Baars JC, Nielsen B, Johannsen LW.** 1995. Production of Suilysin, the Thiol-Activated Hemolysin of Streptococcus-Suis, by Field Isolates from Diseased Pigs. Veterinary Record **137:**295-296.

30. **Gottschalk M, Lebrun A, Wisselink H, Dubreuil JD, Smith H, Vecht U.** 1998. Production of virulence-related proteins by Canadian strains of Streptococcus suis capsular type 2. Can J Vet Res **62:**75-79.

31. **Smith HE, Vecht U, Wisselink HJ, Stockhofe-Zurwieden N, Biermann Y, Smits MA.** 1996. Mutants of Streptococcus suis types 1 and 2 impaired in expression of muramidase-released protein and extracellular protein induce disease in newborn germfree pigs. Infect Immun **64:**4409-4412.

32. **Holden MT, Hauser H, Sanders M, Ngo TH, Cherevach I, Cronin A, Goodhead I, Mungall K, Quail MA, Price C, Rabbinowitsch E, Sharp S, Croucher NJ, Chieu TB, Mai NT, Diep TS, Chinh NT, Kehoe M, Leigh JA, Ward PN, Dowson CG, Whatmore AM, Chanter N, Iversen P, Gottschalk M, Slater JD, Smith HE, Spratt BG, Xu J, Ye C, Bentley S, Barrell BG, Schultsz C, Maskell DJ, Parkhill J.** 2009. Rapid evolution of virulence and drug resistance in the emerging zoonotic pathogen Streptococcus suis. PLoS One **4:**e6072.

33. **Chen C, Zhang W, Zheng H, Lan R, Wang H, Du P, Bai X, Ji S, Meng Q, Jin D, Liu K, Jing H, Ye C, Gao GF, Wang L, Gottschalk M, Xu J.** 2013. Minimum core genome sequence typing of bacterial pathogens: a unified approach for clinical and public health microbiology. J Clin Microbiol **51:**2582-2591.

34. **Athey TB, Auger JP, Teatero S, Dumesnil A, Takamatsu D, Wasserscheid J, Dewar K, Gottschalk M, Fittipaldi N.** 2015. Complex Population Structure and Virulence Differences among Serotype 2 Streptococcus suis Strains Belonging to Sequence Type 28. PLoS One **10:**e0137760.

35. **Weinert LA, Chaudhuri RR, Wang J, Peters SE, Corander J, Jombart T, Baig A, Howell KJ, Vehkala M, Valimaki N, Harris D, Chieu TT, Van Vinh Chau N, Campbell J, Schultsz C, Parkhill J, Bentley SD, Langford PR, Rycroft AN, Wren BW, Farrar J, Baker S, Hoa NT, Holden MT, Tucker AW, Maskell DJ, Consortium BRT.** 2015. Genomic signatures of human and animal disease in the zoonotic pathogen Streptococcus suis. Nat Commun **6:**6740.

36. **Athey TB, Teatero S, Takamatsu D, Wasserscheid J, Dewar K, Gottschalk M, Fittipaldi N.** 2016. Population Structure and Antimicrobial Resistance Profiles of Streptococcus suis Serotype 2 Sequence Type 25 Strains. PLoS One **11:**e0150908.

37. **Zheng H, Ji S, Lan R, Liu Z, Bai X, Zhang W, Gottschalk M, Xu J.** 2014. Population analysis of Streptococcus suis isolates from slaughtered swine by use of minimum core genome sequence typing. J Clin Microbiol **52:**3568-3572.

38. **King SJ, Leigh JA, Heath PJ, Luque I, Tarradas C, Dowson CG, Whatmore AM.** 2002. Development of a Multilocus Sequence Typing Scheme for the Pig Pathogen Streptococcus suis: Identification of Virulent Clones and Potential Capsular Serotype Exchange. Journal of Clinical Microbiology **40:**3671-3680.

822 39. **Falush D, Bowden R.** 2006. Genome-wide association mapping in
823 bacteria? Trends Microbiol **14:**353-355.
824 40. **Read TD, Massey RC.** 2014. Characterizing the genetic basis of bacterial
825 phenotypes using genome-wide association studies: a new direction for
826 bacteriology. Genome Med **6:**109.
827 41. **Clifton-Hadley FA.** 1981. Studies of *Streptococcus suis* type 2 infection in
828 pigs: University of Cambridge.
829 42. **Team RDC.** 2008. R: A language and environment for statistical
830 computing. R Foundation for Statistical Computing.
831 43. **Jombart T, Pontier D, Dufour AB.** 2009. Genetic markers in the
832 playground of multivariate analysis. Heredity (Edinb) **102:**330-341.
833 44. **Jombart T, Devillard S, Balloux F.** 2010. Discriminant analysis of
834 principal components: a new method for the analysis of genetically
835 structured populations. Bmc Genetics **11**.
836 45. **Jombart T.** 2008. adegenet: a R package for the multivariate analysis of
837 genetic markers. Bioinformatics **24:**1403-1405.
838 46. **Jombart T, Ahmed I.** 2011. adegenet 1.3-1: new tools for the analysis of
839 genome-wide SNP data. Bioinformatics **27:**3070-3071.
840 47. **Heinze G, Ploner M, Dunkler D, Southworth H.** 2013. logistf: Firth's
841 bias reduced logistic regression.
842 48. **Sing T, Sander O, Beerenwinkel N, Lengauer T.** 2005. ROCR: visualizing
843 classifier performance in R. Bioinformatics **21:**3940-3941.
844 49. **R Core Team.** 2013. R: A language and environment for statistical
845 computing.

50. **Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N, Harris DE, Holden MT, Churcher CM, Bentley SD, Mungall KL, Cerdeno-Tarraga AM, Temple L, James K, Harris B, Quail MA, Achtman M, Atkin R, Baker S, Basham D, Bason N, Cherevach I, Chillingworth T, Collins M, Cronin A, Davis P, Doggett J, Feltwell T, Goble A, Hamlin N, Hauser H, Holroyd S, Jagels K, Leather S, Moule S, Norberczak H, O'Neil S, Ormond D, Price C, Rabbinowitsch E, Rutter S, Sanders M, Saunders D, Seeger K, Sharp S, Simmonds M, Skelton J, Squares R, Squares S, Stevens K, Unwin L, et al.** 2003. Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica. Nat Genet **35:**32-40.

51. **Liu Z, Zheng H, Gottschalk M, Bai X, Lan R, Ji S, Liu H, Xu J.** 2013. Development of multiplex PCR assays for the identification of the 33 serotypes of Streptococcus suis. PLoS One **8:**e72070.

52. **Athey TB, Teatero S, Lacouture S, Takamatsu D, Gottschalk M, Fittipaldi N.** 2016. Determining Streptococcus suis serotype from short-read whole-genome sequencing data. BMC Microbiol **16:**162.

53. **Ishida S, Tien le HT, Osawa R, Tohya M, Nomoto R, Kawamura Y, Takahashi T, Kikuchi N, Kikuchi K, Sekizaki T.** 2014. Development of an appropriate PCR system for the reclassification of Streptococcus suis. J Microbiol Methods **107:**66-70.

54. **Silva LM, Baums CG, Rehm T, Wisselink HJ, Goethe R, Valentin-Weigand P.** 2006. Virulence-associated gene profiling of Streptococcus suis isolates by PCR. Vet Microbiol **115:**117-127.

Journal of Clinical Microbiology

55. **Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-Ponten T, Ussery DW, Aarestrup FM, Lund O.** 2012. Multilocus sequence typing of total-genome-sequenced bacteria. J Clin Microbiol **50:**1355-1361.

56. **Kosinski AS.** 2013. A weighted generalized score statistic for comparison of predictive values of diagnostic tests. Stat Med **32:**964-977.

57. **Stock C, Hielscher T.** 2014. DTComPair: comparison of binary diagnostic tests in a paired study design. http://CRAN.R-project.org/package=DTComPair. Accessed

58. **Gottschalk M, Segura M.** 2000. The pathogenesis of the meningitis caused by Streptococcus suis: the unresolved questions. Vet Microbiol **76:**259-272.

59. **Segura M, Zheng H, de Greeff A, Gao GF, Grenier D, Jiang Y, Lu C, Maskell D, Oishi K, Okura M, Osawa R, Schultsz C, Schwerk C, Sekizaki T, Smith H, Srimanote P, Takamatsu D, Tang J, Tenenbaum T, Tharavichitkul P, Hoa NT, Valentin-Weigand P, Wells JM, Wertheim H, Zhu B, Gottschalk M, Xu J.** 2014. Latest developments on Streptococcus suis: an emerging zoonotic pathogen: part 1. Future Microbiol **9:**441-444.

60. **Segura M, Zheng H, Greeff A, Gao GF, Grenier D, Jiang Y, Lu C, Maskell D, Oishi K, Okura M, Osawa R, Schultsz C, Schwerk C, Sekizaki T, Smith H, Srimanote P, Takamatsu D, Tang J, Tenenbaum T, Tharavichitkul P, Hoa NT, Valentin-Weigand P, Wells JM, Wertheim H, Zhu B, Xu J, Gottschalk M.** 2014. Latest developments on

Streptococcus suis: an emerging zoonotic pathogen: part 2. Future Microbiol **9:**587-591.

61. **Marois C, Bougeard S, Gottschalk M, Kobisch M.** 2004. Multiplex PCR assay for detection of Streptococcus suis species and serotypes 2 and 1/2 in tonsils of live and dead pigs. J Clin Microbiol **42:**3169-3175.

62. **Marois C, Le Devendec L, Gottschalk M, Kobisch M.** 2007. Detection and molecular typing of Streptococcus suis in tonsils from live pigs in France. Can J Vet Res **71:**14-22.

63. **Samanovic MI, Ding C, Thiele DJ, Darwin KH.** 2012. Copper in microbial pathogenesis: meddling with the metal. Cell Host Microbe **11:**106-115.

64. **Fu Y, Tsui HC, Bruce KE, Sham LT, Higgins KA, Lisher JP, Kazmierczak KM, Maroney MJ, Dann CE, 3rd, Winkler ME, Giedroc DP.** 2013. A new structural paradigm in copper resistance in Streptococcus pneumoniae. Nat Chem Biol **9:**177-183.

65. **Ladomersky E, Petris MJ.** 2015. Copper tolerance and virulence in bacteria. Metallomics **7:**957-964.

66. **Willemse N, Schultsz C.** 2016. Distribution of Type I Restriction-Modification Systems in Streptococcus suis: An Outlook. Pathogens **5**.

67. **Willemse N, Howell KJ, Weinert LA, Heuvelink A, Pannekoek Y, Wagenaar JA, Smith HE, van der Ende A, Schultsz C.** 2016. An emerging zoonotic clone in the Netherlands provides clues to virulence and zoonotic potential of Streptococcus suis. Sci Rep **6:**28984.

68. **Vasu K, Nagaraja V.** 2013. Diverse functions of restriction-modification systems in addition to cellular defense. Microbiol Mol Biol Rev **77:**53-72.

69. **Blow MJ, Clark TA, Daum CG, Deutschbauer AM, Fomenkov A, Fries R, Froula J, Kang DD, Malmstrom RR, Morgan RD, Posfai J, Singh K, Visel A, Wetmore K, Zhao Z, Rubin EM, Korlach J, Pennacchio LA, Roberts RJ.** 2016. The Epigenomic Landscape of Prokaryotes. PLoS Genet **12:**e1005854.

70. **Manso AS, Chai MH, Atack JM, Furi L, De Ste Croix M, Haigh R, Trappetti C, Ogunniyi AD, Shewell LK, Boitano M, Clark TA, Korlach J, Blades M, Mirkes E, Gorban AN, Paton JC, Jennings MP, Oggioni MR.** 2014. A random six-phase switch regulates pneumococcal virulence via global epigenetic changes. Nat Commun **5:**5055.

71. **Merhej V, Georgiades K, Raoult D.** 2013. Postgenomic analysis of bacterial pathogens repertoire reveals genome reduction rather than virulence factors. Brief Funct Genomics **12:**291-304.

Figure 1. Agarose gel showing the expected amplicon sizes of our three genetic markers with the *Streptococcus suis*-specific marker from 14 isolates of the training collection. Agarose gel containing multiplex-PCR amplicons produced from genomic DNA of eight invasive disease-associated, and six non-disease associated isolates of *S. suis* recovered from pigs on farms in England and Wales. PCR amplicons were electrophoresed on a 2% (wt/vol) agarose gel containing 1x SYBR Safe DNA gel stain for 60 minutes at a constant 100 V and photographed under UV transillumination. Multiplex-PCR amplicon patterns matched anticipated amplicon patterns based on *in silico* analyses for all isolates described in this study. Isolate names are indicated above lanes. Lane M contains 1x Bioline HyperLadder 100 bp Plus DNA ladder with sizes indicated on the left (bp). Multiplex-PCR amplicon sizes are indicated on the right (bp).

945 **Table 1. Multiplex-PCR primer details.** Multiplex-PCR primers were
946 designed using the online software primer3 (version 4.0.0, http://primer3.ut.ee)
947 and designed to target conserved regions within the protein-encoding sequence
948 of genetic markers (as opposed to flanking regions). Primers were designed to
949 have similar physical characteristics, enabling simultaneous amplification under
950 the same thermal cycling conditions and in multiplex reactions. GenBank
951 identifier prefixes "SSU" and "SSUST3" correspond to *Streptococcus suis* P1/7
952 (NC_012925) (32) and *Streptococcus suis* ST3 (NC_015433) (71) respectively.
953

954

Primer name	Primer sequence (5' – 3')	Marker of	Multiplex-PCR amplicon size (bp)	Predicted biological function (Interpro)
SSU0207_0735F	TTACAAGAACAGGGCAAGACAGTCGCC	Disease-association	211	Copper exporting ATPase 1
SSU0207_0945R	GCTGCTTTATAATCTGGGTGTTCGTTG			
SSU1589_0460F	CCTTTAATGCAGGGGACAAAAGTGAGCTC	Disease-association	347	Type I restriction-modification (RM) system S protein
SSU1589_0806R	CCCATAATCTTACAGTTAACTTCCTTGC			
SSUST30534_0368F	ATCCCCTCCCAATAAAAGATTTGGATGC	Non-disease association	892	Putative sugar ABC transporter
SSUST30534_1259R	TTTTCGAGCTCTCCATACACTGCTTCTG			
SSU0577_0086F	CAGGTAGTTTGGGCTTAGCTTCATCAGG	*Streptococcus suis* sp.	722	Sporulation regulator (WhiA)
SSU0577_0807R	TGGATGCTGAATTCGCAACTGGGCAATC			

Table 2. Contingency tables used to calculate the performance metrics summarising the classification of *Streptococcus suis* isolates in the out-of-sample test collection (n=50). Contingency tables used to calculate and summarise the performance metrics of two binary diagnostic tests. Each table compares the observed clinical phenotype (considered the 'gold-standard' in this study) to the use of the **a**) newly described pathotyping markers, **b**) serotypes: 1-9 and 1/2, **c**) Virulence-associated markers: *epf, mrp*, and/or *sly*, **d**) Multilocus sequence type (MLST): 1, and **e**) assignment to Minimum Core Genome (MCG) sequence type: 1 as markers of invasive disease.

a)

		Phenotype			
	50 Total population	23 Phenotype positive	27 Phenotype negative		
mPCR	19 mPCR positive	19 True positive	0 False positive	1.00 Positive predictive rate	0.00 False discovery rate
	31 mPCR negative	4 False negative	27 True negative	0.13 False omission rate	0.87 Negative predictive rate
		0.83 True positive rate	0.00 False positive rate	0.90 F_1 score	
		0.17 False negative rate	1.00 True negative rate		

966 **b)**

		Phenotype			
	50 Total population	23 Phenotype positive	27 Phenotype negative		
Serotype: 1-9, 14 & 1/2	26 Positive	20 True positive	6 False positive	0.77 Positive predictive rate	0.23 False discovery rate
	24 Negative	3 False negative	21 True negative	0.13 False omission rate	0.88 Negative predictive rate
		0.87 True positive rate	0.22 False positive rate	0.82 F_1 score	
		0.13 False negative rate	0.78 True negative rate		

967

968 **c)**

		Phenotype			
	50 Total population	23 Phenotype positive	27 Phenotype negative		
Virulence-associated markers: *epf*, *mrp* & *sly*	48 Positive	22 True positive	26 False positive	0.46 Positive predictive rate	0.54 False discovery rate
	2 Negative	1 False negative	1 True negative	0.50 False omission rate	0.50 Negative predictive rate
		0.96 True positive rate	0.96 False positive rate	0.62 F_1 score	
		0.04 False negative rate	0.04 True negative rate		

969

970 **d)**

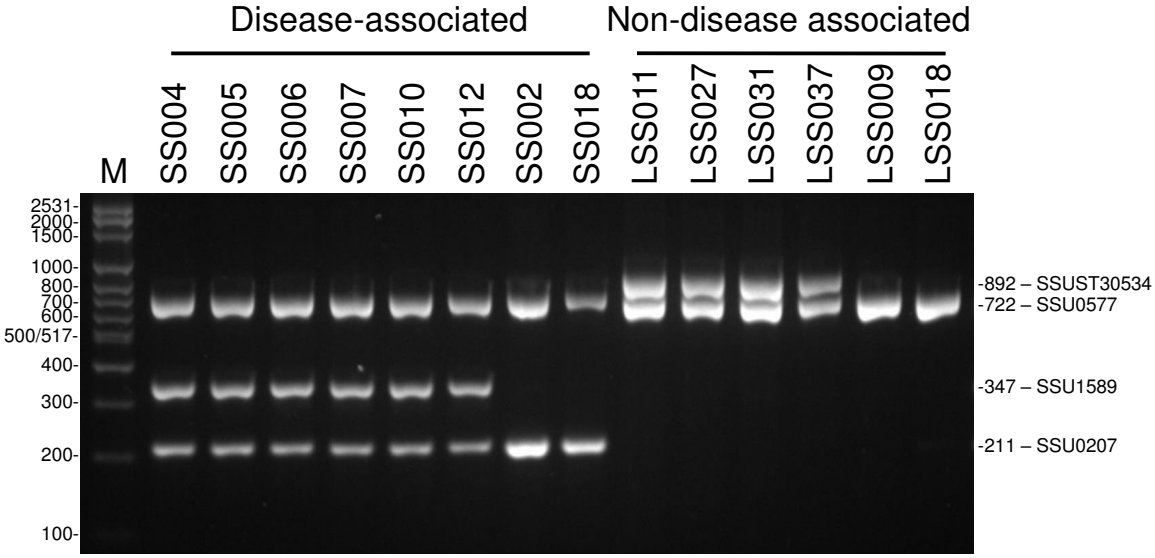
		Phenotype			
	50 Total population	23 Phenotype positive	27 Phenotype negative		
Multilocus Sequence Type: 1	16 Positive	16 True positive	0 False positive	1.00 Positive predictive rate	0.00 False discovery rate
	34 Negative	7 False negative	27 True negative	0.21 False omission rate	0.79 Negative predictive rate
		0.70 True positive rate	0.00 False positive rate	0.82 F_1 score	
		0.30 False negative rate	1.00 True negative rate		

971

972 **e)**

		Phenotype			
	50 Total population	23 Phenotype positive	27 Phenotype negative		
Minimum Core Genome group: 1	18 Positive	18 True positive	0 False positive	1.00 Positive predictive rate	0.00 False discovery rate
	32 Negative	5 False negative	27 True negative	0.16 False omission rate	0.84 Negative predictive rate
		0.78 True positive rate	0.00 False positive rate	0. F_1 score	
		0.22 False negative rate	1.00 True negative rate		

973



Primer name	Primer sequence (5' – 3')	Marker of	Multiplex-PCR amplicon size (bp)	Predicted biological function (Interpro)
SSU0207_0735F	TTACAAGAACAGGGCAAGACAGTCGCC	Disease-association	211	Copper exporting ATPase 1
SSU0207_0945R	GCTGCTTTATAATCTGGGTGTTCGTTG			
SSU1589_0460F	CCTTTAATGCAGGGGACAAAAGTGAGCTC	Disease-association	347	Type I restriction-modification (RM) system S protein
SSU1589_0806R	CCCATAATCTTACAGTTAACTTCCTTGC			
SSUST30534_0368F	ATCCCCTCCCAATAAAAGATTTGGATGC	Non-disease association	892	Putative sugar ABC transporter
SSUST30534_1259R	TTTTCGAGCTCTCCATACACTGCTTCTG			
SSU0577_0086F	CAGGTAGTTTGGGCTTAGCTTCATCAGG	*Streptococcus suis* sp.	722	Sporulation regulator (WhiA)
SSU0577_0807R	TGGATGCTGAATTCGCAACTGGGCAATC			

Table 2. Contingency tables used to calculate the performance metrics summarising the classification of *Streptococcus suis* isolates in the out-of-sample test collection (n=50). Contingency tables used to calculate and summarise the performance metrics of two binary diagnostic tests. Each table compares the observed clinical phenotype (considered the 'gold-standard' in this study) to the use of the **a**) newly described pathotyping markers, **b**) serotypes: 1-9 and 1/2, **c**) Virulence-associated markers: *epf, mrp*, and/or *sly*, **d**) Multilocus sequence type (MLST): 1, and **e**) assignment to Minimum Core Genome (MCG) sequence type: 1 as markers of invasive disease.

a)

		Phenotype			
	50 Total population	23 Phenotype positive	27 Phenotype negative		
mPCR	19 mPCR positive	19 True positive	0 False positive	1.00 Positive predictive rate	0.00 False discovery rate
mPCR	31 mPCR negative	4 False negative	27 True negative	0.13 False omission rate	0.87 Negative predictive rate
		0.83 True positive rate	0.00 False positive rate	0.90 F_1 score	
		0.17 False negative rate	1.00 True negative rate		

b)

		Phenotype			
	50 Total population	23 Phenotype positive	27 Phenotype negative		
Serotype: 1-9, 14 & 1/2	26 Positive	20 True positive	6 False positive	0.77 Positive predictive rate	0.23 False discovery rate
	24 Negative	3 False negative	21 True negative	0.13 False omission rate	0.88 Negative predictive rate
		0.87 True positive rate	0.22 False positive rate	0.82 F_1 score	
		0.13 False negative rate	0.78 True negative rate		

c)

		Phenotype			
	50 Total population	23 Phenotype positive	27 Phenotype negative		
Virulence-associated markers: *epf, mrp & sly*	48 Positive	22 True positive	26 False positive	0.46 Positive predictive rate	0.54 False discovery rate
	2 Negative	1 False negative	1 True negative	0.50 False omission rate	0.50 Negative predictive rate
		0.96 True positive rate	0.96 False positive rate	0.62 F_1 score	
		0.04 False negative rate	0.04 True negative rate		

d)

		Phenotype			
	50 Total population	23 Phenotype positive	27 Phenotype negative		
Multilocus Sequence Type: 1	16 Positive	16 True positive	0 False positive	1.00 Positive predictive rate	0.00 False discovery rate
	34 Negative	7 False negative	27 True negative	0.21 False omission rate	0.79 Negative predictive rate
		0.70 True positive rate	0.00 False positive rate	0.82 F_1 score	
		0.30 False negative rate	1.00 True negative rate		

e)

		Phenotype			
	50 Total population	23 Phenotype positive	27 Phenotype negative		
Minimum Core Genome group: 1	18 Positive	18 True positive	0 False positive	1.00 Positive predictive rate	0.00 False discovery rate
	32 Negative	5 False negative	27 True negative	0.16 False omission rate	0.84 Negative predictive rate
		0.78 True positive rate	0.00 False positive rate	0. F_1 score	
		0.22 False negative rate	1.00 True negative rate		