SCIENTIFIC REPORTS

OPEN

Discrimination of contagious and environmental strains of *Streptococcus uberis* in dairy herds by means of mass spectrometry and machine-learning

Received: 1 August 2018
Accepted: 9 November 2018
Published online: 30 November 2018

Necati Esener[1], Martin J. Green[1], Richard D. Emes[1,2], Benjamin Jowett[1], Peers L. Davies[1], Andrew J. Bradley[1,3] & Tania Dottorini[1,2]

Streptococcus uberis **is one of the most common pathogens of clinical mastitis in the dairy industry. Knowledge of pathogen transmission route is essential for the selection of the most suitable intervention. Here we show that spectral profiles acquired from clinical isolates using matrix-assisted laser desorption ionization/time of flight (MALDI-TOF) can be used to implement diagnostic classifiers based on machine learning for the successful discrimination of environmental and contagious *S. uberis* strains. Classifiers dedicated to individual farms achieved up to 97.81% accuracy at cross-validation when using a genetic algorithm, with Cohen's kappa coefficient of 0.94. This indicates the potential of the proposed methodology to successfully support screening at the herd level. A global classifier developed on merged data from 19 farms achieved 95.88% accuracy at cross-validation (kappa 0.93) and 70.67% accuracy at external validation (kappa 0.34), using data from another 10 farms left as holdout. This indicates that more work is needed to develop a screening solution successful at the population level. Significant MALDI-TOF spectral peaks were extracted from the trained classifiers. The peaks were found to correspond to bacteriocin and ribosomal proteins, suggesting that immunity, growth and competition over nutrients may be correlated to the different transmission routes.**

Clinical mastitis is one of the most important challenges facing the dairy industry, where it reduces productivity, profitability and cow welfare. Considerable progress has been made in understanding the epidemiology and microbiology of mastitis over the past four decades, identifying the physical origins of infection as contagious or environmental[1] and temporal origins of the infection e.g. dry period or lactation[2]. Coliform bacteria are almost always environmental, whilst other pathogens such as *Staphilocossus aureus* and *Streptococcus agalactiae* are typically contagious on the contrary, *S. uberis* can commonly manifest itself in both contagious and environmental forms[3], and the ability to diagnose the clinical mastitis transmission pattern (contagious or environmental) is an essential step to identify appropriate and effective management interventions for the control of the disease at a herd level. Limited financial and labour resources are typically available to farmers for mastitis control. A role of the clinician is to identify the mode of transmission as early as possible, in order to react appropriately before costly production losses have occurred. Currently, clinicians assess the most likely mode of transmission through analysis of historical data[4], visual observation of management practices (milking, cleaning, etc.) and knowledge of pathogens, although there is limited evidence that the latter two methods are useful in determining transmission patterns in the modern dairy herd. This leads to inevitable delays and associated losses before a diagnosis of a new, emerging disease pattern can be made. Prompt diagnosis of the likely transmission route in case of an outbreak

[1]University of Nottingham School of Veterinary Medicine and Science, College Road, Sutton Bonington, Leicestershire, LE12 5RD, UK. [2]Advanced Data Analysis Centre, University of Nottingham, College Road, Sutton Bonington, Leicestershire, LE12 5RD, UK. [3]Quality Milk Management Services ltd, Cedar Barn, Easton Hill, Easton, Wells, BA5 1DU, UK. Correspondence and requests for materials should be addressed to T.D. (email: tania.dottorini@nottingham.ac.uk)

would allow appropriate control interventions to be implemented earlier and reduce deleterious production and welfare consequences of additional clinical mastitis cases.

Previous studies in this field have used genomic epidemiological techniques to classify individual bacterial strains broadly as contagious or environmental according to their observed patterns of clinical disease within multiple independent herds[4]. These techniques are useful as research tools to understand the observed patterns but are too costly and laborious to be practical clinical tools for clinicians. The discriminatory ability for genomic techniques such as multilocus sequence typing (MLST) may also not be appropriate for the classification of bacterial isolates according to their clinical manifestation if those attributes which govern the transmission behaviour of an isolate are determined by epigenetic factors or conferred by mobile genetic elements which are rapidly exchanged between bacteria such as *S. uberis*[5].

Evidence of epigenetic strain variation and strain evolution within a bacterial species has been described by several mechanisms, such as differential methylation resulting in phase variation[6] as a means for isolates of commensal and pathogenic bacterial species to adapt to new or changing environments. In human cases of Salmonellosis, epigenetic strain variation in virulence and host-pathogen interaction could be demonstrated by proteomic analysis where genomic discrimination of strains was not possible[7]. When identifying the route of transmission of an individual bacterial isolate from a case of bovine mastitis, a rapid proteomic technology able to characterise variation in antigenic expression related to virulence, such as surface protein molecules, may be more discriminatory than existing genomic techniques[8].

Gel-based proteomic techniques for strain differentiation have been used for decades in many species and disease processes including bovine mastitis[9]. More recently, highly discriminatory techniques, such as Matrix Assisted Laser Desorption/Ionization- Time of Flight Mass Spectrometry (MALDI-TOF MS) have increased our ability to investigate the molecular epidemiology and host-pathogen interactions of many bacterial pathogens[10].

In contrast to genomic techniques, MALDI-TOF MS provides a rapid and economic means of identifying bacteria and has been shown to be capable of strain differentiation within a bacterial species such as *S. pneumoniae, Y. entrocolitica* and *M. pneumoniae*[11–13].

Comparison of MALDI-TOF MS proteomic profiles may allow discrimination between bacterial, isolate strains of a pathogen, such as environmental *S. uberis* and contagious *S. uberis* strains which have acquired or evolved enhanced survival or colonisation characteristics (genetic or epigenetic) that increase the risk of a cow to cow transmission.

The primary aim of this study was to investigate MALDI-TOF MS data with machine learning as a method to discriminate between *S. uberis* isolates with different modes of transmission; contagious and environmental. A secondary aim of the study was to compare strain differences within and between farms of the UK. The final aim of the study was to identify proteins related to the differentiating peaks between transmission routes with bioinformatics tools.

Results

Data source. This study used MALDI-TOF spectra and MLST results obtained during a previous study[4] carried out between March 2004 and May 2005 using selected farms located in England and Wales (see Fig. 1). In that study[4], *S. uberis* isolates had been cultured from milk samples from 52 herds, each from a different farm (52 farms). The herds were selected only if containing >35 cases of clinical mastitis per 100 cows during a period of 12 months. Within each herd, only cows with a positive diagnosis of mastitis were considered, and each was assigned a clinical classification of either contagious or environmental based on multi-locus sequence typing, as previously described by Davies *et al.*[4]. In such work, the median incidence of clinical mastitis cases had been reported as 66 (mean, 75) per 100 cows per 365 days, with a range of 16 to 146 cases. The mean percentage of clinical cases diagnosed as *S. uberis* had been 28%, with a range in individual herds from 7% to 64%.

In this work, we first assessed the geographical distribution of clinical cases. The results are shown in Fig. 1, indicating spread in England and Wales, with a higher concentration towards the south, and no herds in Scotland.

In order to construct predictors/classifiers, we focused on herds containing both environmental and contagious *S. uberis* isolates. Thus, by looking at the available MLST data, we selected for the study only the 29 farms/herds containing both. The 23 eliminated herds consisted of: 2 not containing any *S. uberis* isolates, 4 containing only contagious, 13 containing only environmental and 4 containing unclassified isolates (see Fig. 2). Amongst the 29 herds selected, 10 were reserved for external validation (holdout group) as they each featured <20 MALDI-TOF spectra, too few to be useful for the generation of effective classifiers. The remaining 19 farms (Fig. 1b) were considered suitable for the development of classification models.

Generation of MALDI-TOF peak lists and set-up of the classifiers. The original MALDI-TOF raw spectra had been obtained in the previous study using Bruker technology[3]. In this work, peak lists, i.e. lists of paired mass/charge (m/z) ratios and corresponding intensity values, were extracted from the raw spectra as specified in the Methods section.

To verify if MALDI-TOF peak lists associated with isolates could be used to predict their contagious or environmental nature, supervised machine learning technologies were used to implement classifiers, i.e. software systems that, once provided with a spectrum as the input, would respond by predicting the most likely class (i.e. environmental or contagious) for the isolate. Being based on supervised learning, all methods required the availability of training datasets for model building[14] (i.e. peak lists with known associated classification of environmental or contagious from MLST), and validation datasets for assessing the performance of the classifier.

Classification methods. The following classification methods available in the Bruker mass spectrometry analysis software ClinProTools 3.0[15] were tested: Genetic Algorithm (GA), Supervised Neural Network (SNN) and Quick Classifier (QC).

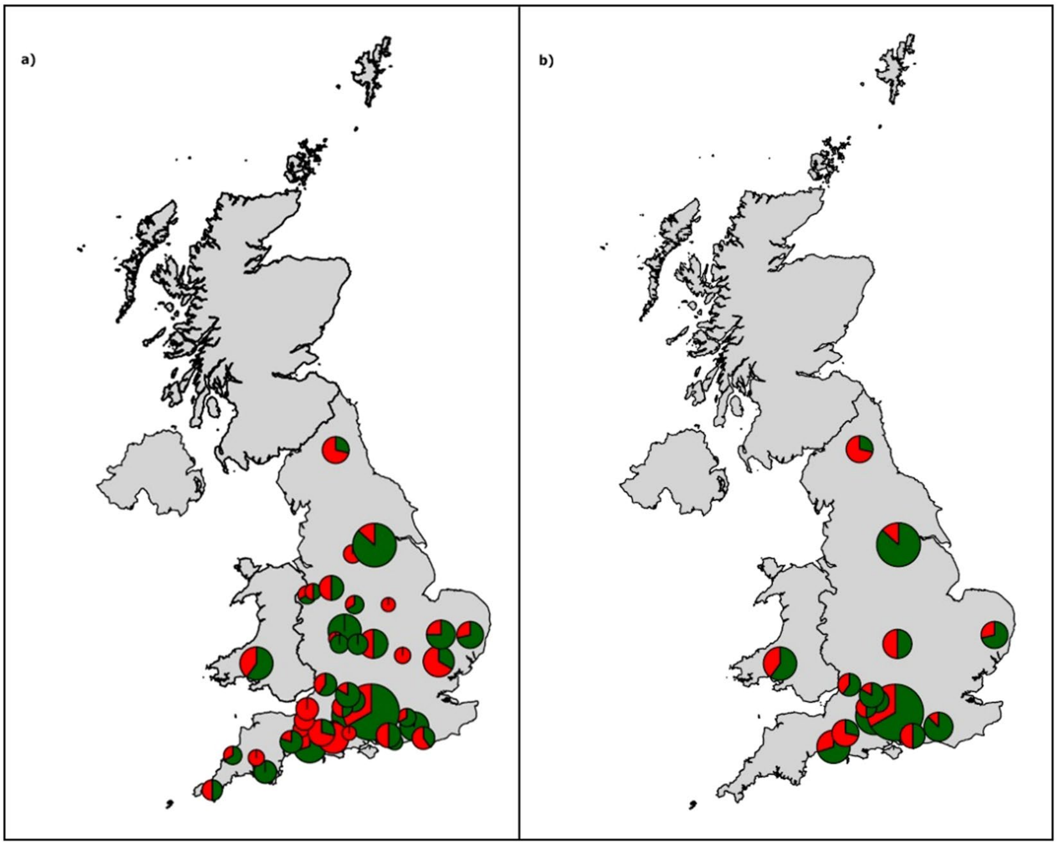

Figure 1. Location of the enrolled farms on the map of the United Kingdom. (**a**) The entire set of 52 farms (**b**) the 19 farms selected for building the model for intra-farm analysis. The red colour represents the environmental isolates of Streptococcus uberis while the green is for contagious ones. The size of the circle indicates the number of Streptococcus uberis isolates in the farms. Figure generated using open source R packages maps and mapdata available from CRAN[66].

The GA method uses the training datasets to identify a subset of peaks shared by all of the peak lists (referred to as "peak combination"), acting as the most effective discriminator between contagious and environmental isolates. The performance of each tentative combination in terms of discrimination effectiveness is assessed by evaluating the degree of separation of the clusters formed by the known contagious and environmental isolates, once that combination of peaks is considered. The identification of the best combination is treated as an optimisation problem and solved via a genetic algorithm[16]. When the best peak combination is found (end of training), new peak lists submitted to the classifier are predicted as being contagious or environmental by extracting the peak combination, and using it to determine which one of the two clusters is closest to the observation using the k-nearest neighbour (KNN) metric[15].

The SNN method is based on building a classifier powered by a neural network implementing a modified version of the supervised relevance neural gas algorithm[17]. The peak lists of the training set are investigated by the algorithm, to identify "prototype" lists suitable to act as representative for the corresponding class (contagious or environmental). Once trained, the network can be fed with any new peak list; the prediction will be based on understanding which prototypes are closer to the given list[15].

The QC method is based on grouping the peak lists into two classes (environmental and contagious), generating peak averages representative of each class, and ranking the relevance of the peaks when acting as discriminators based on statistical testing[15]. Any new observation is then tested on similarity against the weighted averages of each class, the most similar class being elected as the prediction result[15].

Prediction performance. The prediction performance of each classifier was evaluated by considering the following indicators, assuming P and N as the total number of positive (contagious) and negative (environmental) MLST results, and using T to indicate true (correct) and F to indicate false (wrong) predictions:

- sensitivity (TPR - true positive rate) = TP/P
- specificity (TNR - true negative rate) = TN/N
- precision (PPV - positive predictive value) = TP/(TP + FP)
- NPV - negative predictive value = TN/(TN + FN)
- accuracy (ACC) = (TP + TN)/(P + N)
- Cohen's kappa statistics = $(p_o - p_e)/(1 - p_e)$

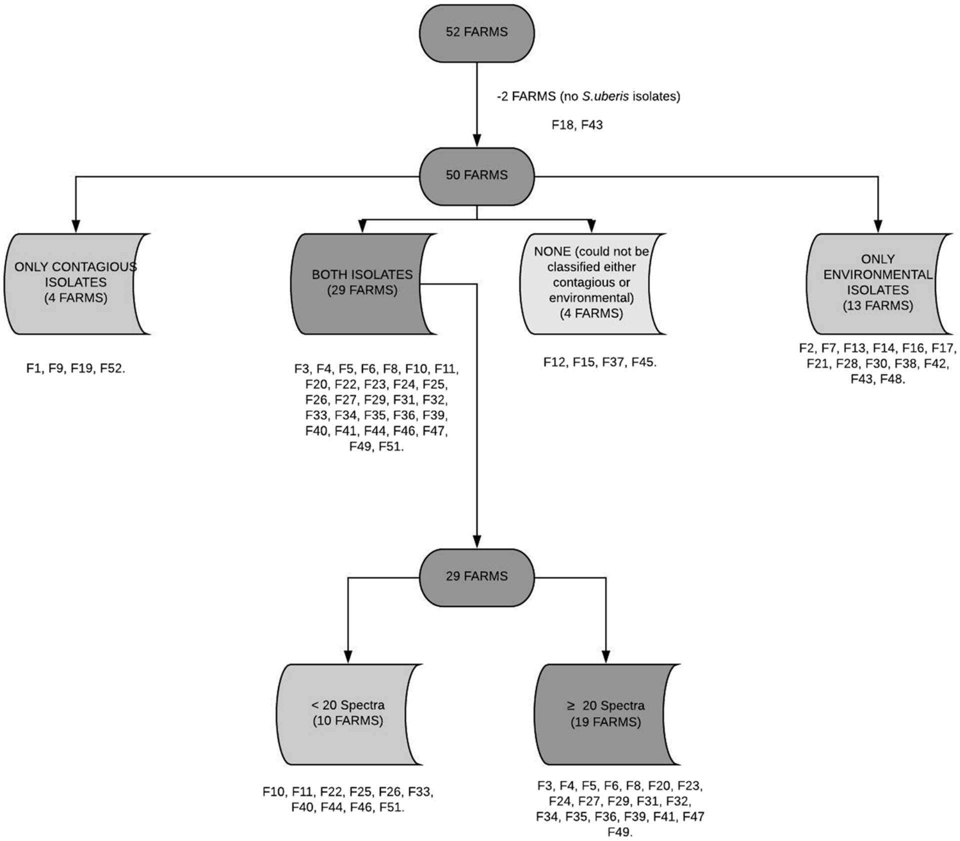
52 FARMS
-2 FARMS (no S.uberis isolates)
F18, F43
50 FARMS
ONLY CONTAGIOUS ISOLATES (4 FARMS)
F1, F9, F19, F52.
BOTH ISOLATES (29 FARMS)
F3, F4, F5, F6, F8, F10, F11, F20, F22, F23, F24, F25, F26, F27, F29, F31, F32, F33, F34, F35, F36, F39, F40, F41, F44, F46, F47, F49, F51.
NONE (could not be classified either contagious or environmental) (4 FARMS)
F12, F15, F37, F45.
ONLY ENVIRONMENTAL ISOLATES (13 FARMS)
F2, F7, F13, F14, F16, F17, F21, F28, F30, F38, F42, F43, F48.
29 FARMS
< 20 Spectra (10 FARMS)
F10, F11, F22, F25, F26, F33, F40, F44, F46, F51.
≥ 20 Spectra (19 FARMS)
F3, F4, F5, F6, F8, F20, F23, F24, F27, F29, F31, F32, F34, F35, F36, F39, F41, F47 F49.

Figure 2. Process of initial farm selection and farm codes. Categorisation was done according to type and presence of Streptococcus uberis (contagious and environmental) strains and number of MALDI-TOF spectra.

where:

$$p_o = (\mathrm{TP} + \mathrm{TN})/(\mathrm{P} + \mathrm{N}) \text{ and } \mathrm{p_e} = (\mathrm{P} \cdot (\mathrm{TP} + \mathrm{FN}) + \mathrm{N} \cdot (\mathrm{FP} + \mathrm{TN}))/(\mathrm{P} + \mathrm{N})^2 \tag{1}$$

Based on the above indicators, the following performance parameters were considered:

- recognition capability (RC): the accuracy obtained when the classifier is trained with the entirety of the dataset and tested on the same data.
- accuracy, kappa and other indicators from cross validation (CV): the dataset is split in a training subset (containing a percentage of the available spectra) and a testing subset. (containing the remaining spectra). The subsets are created by randomly drawing individuals from the same dataset. Accuracy, kappa and any other desired indicator resulting from the confusion matrix are computed. The entire process is repeated n times (training and testing n classifiers), then the final indicators are obtained as the arithmetic means of the values obtained on the n confusion matrices.
- accuracy, kappa and other indicators from external validation (EV): only one classifier is built, by using the entirety of the available dataset for training. Testing is performed using a dataset separate from the training dataset. The indicators are computed on the resulting confusion matrix.

Intra-farm analysis. In intra-farm analysis, classifiers were developed to operate exclusively within each individual farm. The spectra of the 19 farms selected as suitable for model building (Fig. 2) were used to implement 19 separate classifiers (one per farm). Only data pertaining the specific farm were used to implement and validate each classifier. RC and CV performance indicators were computed for each farm, but no external validation was performed.

The results of the intra-farm analysis on the 19 selected farms are shown in Fig. 3a. Classifiers were run with default settings as described in the Methods section. Recognition capability (RC), cross validation (CV) accuracy and kappa are shown as arithmetic means computed from the individual results of the 19 farms (GA: RC = 100.00%, CV accuracy = 97.81%; CV kappa = 93.72%, sensitivity = 97.13%, specificity = 96.26%, PPV = 98.03%, NPV = 96.19%. SNN: RC = 84.00%, CV accuracy = 82.17%, CV kappa = 60.20%, sensitivity = 87.22%, specificity = 72.65%, PPV = 80.09%, NPV = 84.25%. QC: RC = 97.32%, CV accuracy = 91.34%, CV kappa = 80.20%, sensitivity = 92.36%, specificity = 87.20%, PPV = 91.60%, NPV = 90.33%).

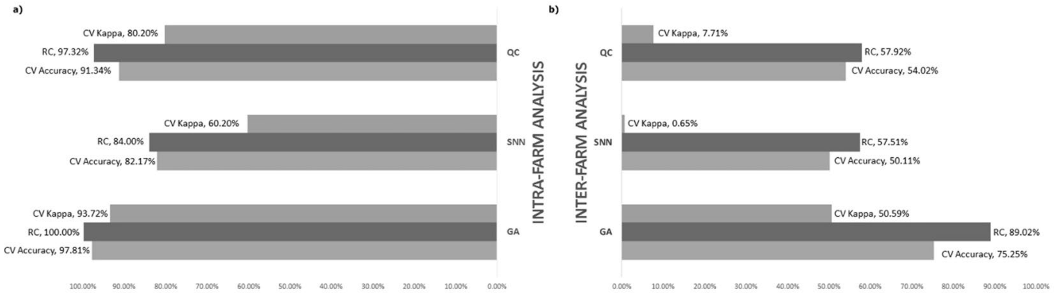

Figure 3. Comparison of intra-farm (**a**) and inter-farm (**b**) analysis results of 19 farms using Genetic Algorithm (GA), Supervised Neural Network (SNN) and Quick Classifier (QC). Inter-farm analysis results are the arithmetic mean of the results from nineteen classifiers (one per farm). All the results were obtained by adopting the default settings for the classifying methods.

Inter-farm analysis. In inter-farm analysis, a single classifier was developed and trained on the aggregated data from all the available 19 farms. RC, and performance indicators from cross validation were computed using the data from the 19 farms. The RC and CV results for the inter-farm analysis are shown in Fig. 3b. Classifiers were run with default settings (GA: RC = 89.02%, CV accuracy = 75.25% CV kappa = 50.59%, sensitivity = 81.75%, specificity = 69.16%, PPV = 82.98%, NPV = 67.32%. SNN: RC = 57.51%, CV accuracy = 50.11%, CV kappa = 0.65%, sensitivity = 64.09%, specificity = 36.55%, PPV = 57.98%, NPV = 42.70%. QC: RC = 57.92%, CV accuracy = 54.02%, CV kappa = 7.71%, sensitivity = 66.97%, specificity = 40.59%, PPV = 61.19%, NPV = 46.77%). The GA method was selected for further optimisation, due to better performance. With the optimised settings (see Methods), the performance of the GA classifier went from 89.02% to 99.09% for RC, from 75.25% to 95.88% for CV accuracy, and from 50.59% to 92.62% for CV kappa.

External validation was performed on the GA classifier, using the additional 10 farms in the holdout group. The following EV indicators were obtained: sensitivity 82.07%, specificity 50.00%, PPV 74.84%, NPV 60.73%, accuracy 70.67% and Cohen's kappa coefficient 33.80%.

The probability distributions of the performance indicators for intra-farm and inter-farm analysis are reported in Fig. 4.

Biomarker characterisation. The GA classifier for the intra-farm analysis identified a set of 19 peaks shared by all the isolates, providing optimal discrimination between environmental and contagious isolates.

The masses of these peaks were then compared with the molecular weights of *S. uberis* proteins in NCBI database and 7 out of 19 could be matched to 8 proteins from the proteome (in one case, the molecular weight of a peak was close to two different proteins). The entire procedure is described in the Methods section.

Five of the eight proteins have a known function (according to NCBI): two of them are ribosomal proteins, two of them bacteriocins and one is an ATP synthase protein. The remaining three proteins had unknown functions; two of these were hypothetical and one of them was of an unknown domain. Using the SMART database, the domains of the 8 proteins were found. The predicted three-dimensional models are shown in Fig. 5.

The analysis of the protein-protein interaction (PPI) network for the discriminant proteins in *S. uberis* (Fig. 6) showed that 5 out of 8 proteins (SUB14226, SUB0666, SUB0081, SUB1796 and SUB0585) share common first neighbour proteins with each other. Interestingly, SUB0666, SUB0081 and SUB1796 were also found to interact with each other. The ontology functions of these 158 proteins (5 of interest and 153 connected to at least two genes of interest) are shown in Fig. 7 as well as KEGG pathways they are involved in (FDR < 0.05).

Discussion

Our study shows that *S. uberis* isolates classified according to transmission route as either contagious or environmental can be discriminated by MALDI-TOF spectral profiles. The discriminatory power of MALDI-TOF appears to be greater in intra-farm analysis, in particular when using genetic algorithms, indicating the potential for the development of successful screening solutions at the herd level. Inter-farm classification does not work equally well, indicating that more work is needed to develop screening solutions applicable at the population level. Such limited classification performance may be due to divergent evolution of bacterial strains across farms. Contact between dairy farms within the UK is limited, and management practices to control environmental mastitis vary (e.g. different choices of bedding materials, different cleaning practices, different antibiotic treatment protocols). Thus, divergent evolution and hence the emergence of farm specific populations of *S. uberis* is entirely plausible. Diversity within common bacterial species has been demonstrated, such as for *Campylobacter* populations between broiler farms in Switzerland[18], and *Staphylococcus aureus* strains in bovines from different geographical regions in Argentina[19]. At a much larger scale, divergent evolution of *Helicobacter pylori* has been described in human populations migrated from a common origin to different geographical regions[20]. Because of such diversity, it is possible that the same combination of discriminant peaks may not work equally well for all the farms.

For the investigated data, classifiers based on the GA method showed better performance over SNN and QC both before and after being optimised.

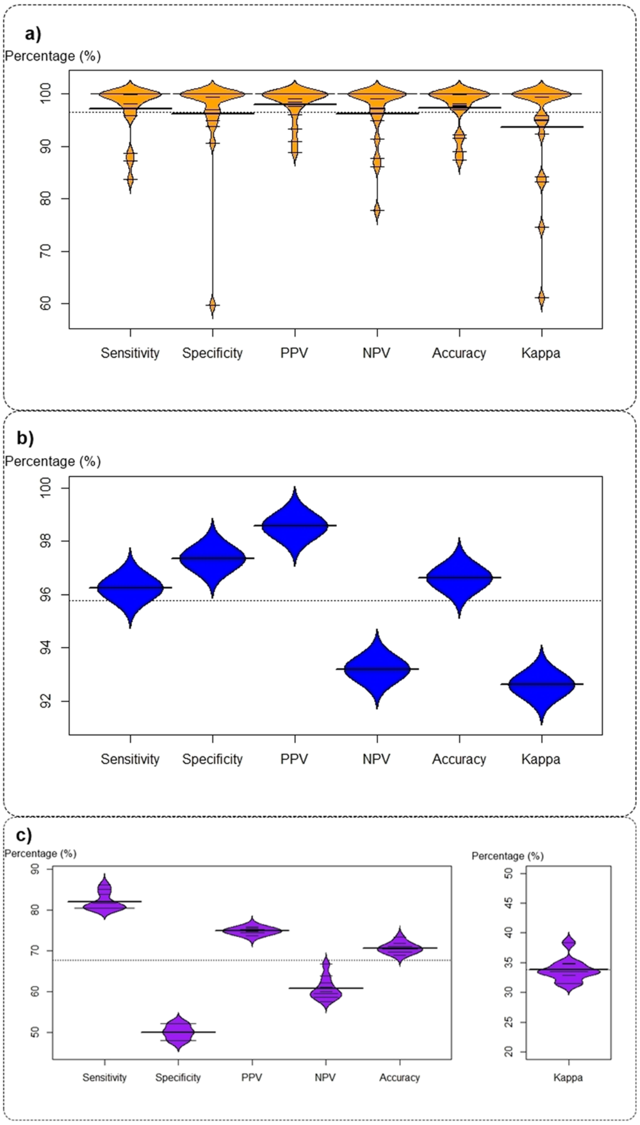

Figure 4. Distribution of the performance indicators for the classifiers/predictors. (**a**) intra-farm cross-validation; (**b**) inter-farm cross-validation; (**c**) inter-farm external validation. Data from 19 farms.

The results showed the presence of proteomic phenotypic differences between contagious and environmental strains of *S. uberis* along with previously demonstrated, different genotypic characteristics[4]. This would appear plausible since selection pressure in the mammary glands is likely to force changes in protein expressions.

In this study, seven of the 19 peaks identified by the GA as the most discriminant between environmental and contagious isolates, where found corresponding to the protein products of eight genes in the reference 0140 J genome. While the GA is treating the identification as a purely mathematical optimization problem, it is interesting to see if the mapped proteins have some functional meaning that might explain their differentiating power. This can be done by looking at gene ontology (GO), where protein functions are annotated according to the aspects of biological process, molecular function and cellular component (see Methods).

The SUB1796 protein is the ribosomal protein-L33 (RP-L33), a relatively small protein consisting of 49 to 66 amino acids, part of the large ribosomal subunit[21]. In GO, RP-L33 is annotated as follows: biological process – translation; molecular function – structural of constituent of ribosome; cellular component – intracellular and ribosome. RP-L33 was shown to have paralogs differing on whether it binds structural zinc or not, which helps bacteria survive in the case of zinc starvation[22]. This is an important consideration, as most of the ribosomal proteins are encoded by single and highly conserved genes. Zinc has a crucial role in preserving the immune status of the cow and mammary gland health; its deficiency can result in an increased mastitis incidence[23]. RP-L33 was discovered to be phylogenetically distinct amongst bacterial species and within single genomes such as that of *Bacillus subtilis*, *Lactococcus lactis*, *Mycoplasma pneumonia*, *Mycoplasma genitalium*, *Ureaplasma ueralyticum* and *Streptomyces coelicolor*[24]. Moreover, RP-L33 was demonstrated to be a core member of the minimal set of bacterial genes which is essential to maintain cell life[25]. RP-L33 was suggested as the putative drug target in three

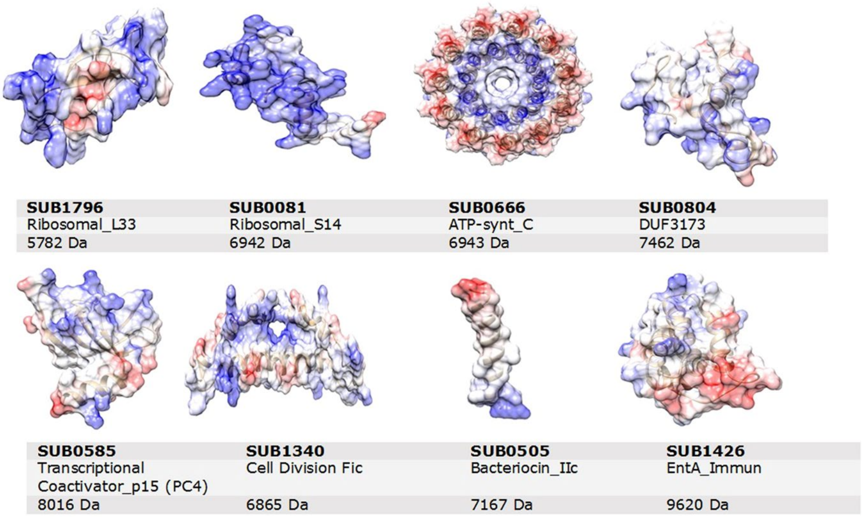

Figure 5. Selected proteins of Streptococcus uberis. Top to bottom: 3D protein structure, Protein ID, Domain of the protein and Molecular weight of the protein.

Streptococcus species (*S. agalactiae*, *S. pneumoniae*, and *S. pyogenes*)[26], two *Bordetella* species (*B. pertussis* and *B. parapertussis*)[27], *Helicobacter pylori*[28] and *Mycobacterium tuberculosis*[29].

The SUB0081 protein is the ribosomal protein S-14 (RpsN). The GO annotation is: biological process – translation; molecular function – metal ion binding, rRNA binding, and structural constituent of the ribosome; cellular component - ribosome. RpsN was shown to be essential for growth and deficiency of RpsN resulted in incomplete 30S subunits in *Bacillus subtilis*[30,31], *E. coli*[32] and *Staphylococcus aereus*[33].

The SUB0666 protein is an ATP synthase (subunit C, atpE). The GO annotation is: biological process - ATP hydrolysis and synthesis, and proton transport; molecular function - hydrogen ion transmembrane transporter, and lipid binding; cellular component – plasma membrane. The atpE gene was targeted by several drug studies such as R207910[34,35] and Bedaquilline[36] in mycobacterium species. Moreover, the C sub-unit was shown to be the target site of venturicidin in several *E. coli* studies[37,38].

The SUB0585 protein is the transcriptional coactivator p15 (PC4). The GO annotation is: biological process - regulation of transcription, DNA-templated; molecular function - DNA binding and transcription coactivator; cellular component – no terms assigned. PC4 is also believed to have an important role in DNA repair of bacterial species since studies with *E. coli*[39] and *Leptospira* species[40] revealed that PC4 protected the DNA during oxidative stress.

The SUB1340 protein (GO annotation: no terms assigned) contains the Fic domain. This domain is often found in pathogenic and non-pathogenic bacteria with different structures, where some families may contain conserved regulatory functions[41]. It is suggested that pathogenic bacteria secrete Fic proteins, thus showing similar functionality to toxins in terms of fulfilling some duties in the host cell such as interfering with cytoskeletal, signalling and translation pathways[41]. Fic proteins participate to cell division and have been shown to synthesize folate in *E. coli*[42]. In turn, folate is involved in the secretion of enzymes controlling the bacterial pathogenesis in cattle pathogenic bacteria, *Histophilus somni*[43].

The SUB0505 protein is a bacteriocin. It has the following GO annotation: biological process - defense, response to bacterium; molecular function – no terms assigned; cellular component – no terms assigned. Many bacteria produce tiny peptides called bacteriocins for antimicrobial activity to compete with intra and interspecies over limited nutrients in the environment[44]. In the study by Ward, *et al.*[45], six bacteriocin proteins including SUB0505 were found in the 0140 J strain, where this redundancy was interpreted as the result of mutations. The study done by Hossain, *et al.*[46] revealed the absence of bacteriocin genes, including SUB0505, in the EF20 strain, which may be correlated to the non-virulent status of the EF20 strains. The EF20 strain of *S. uberis* was shown to be susceptible to phagocytosis by bovine neutrophils in the presence of serum[47], and mammary gland macrophages were reported to have capability of killing the EF20 strain in the media containing 50% skimmed milk as the source of opsonin or 10% pooled bovine serum[48]. Moreover, a comparison of the EF20 strain with the reference strain 0140 J showed EF20 growing relatively slowly in raw skimmed milk[49]. In another study[50], EF20 strains of *S. uberis* performed a high amount of bounding plasmin activity following growth while bovine plasminogen was present in the media. Bacteriocin immunity proteins prevent the bacteria from the toxic effect of its own bacteriocins by forming a stable compound with the receptors[51,52]. The EntA immune protein SUB1426 (no GO annotation) was discovered to guard the particular bacteria against its own class II bacteriocins[53]. The studies in *Streptococcus* species revealed that immunity proteins play a significant role in antimicrobial sensitivity by regulating quorum-sensing[54,55]. In summary, the literature shows that bacteriocins feature high levels of differentiation depending on environment and host immunity response. This may justify why we found SUB0505 as a discriminating peak between environmental and contagious strains of *S. uberis*.

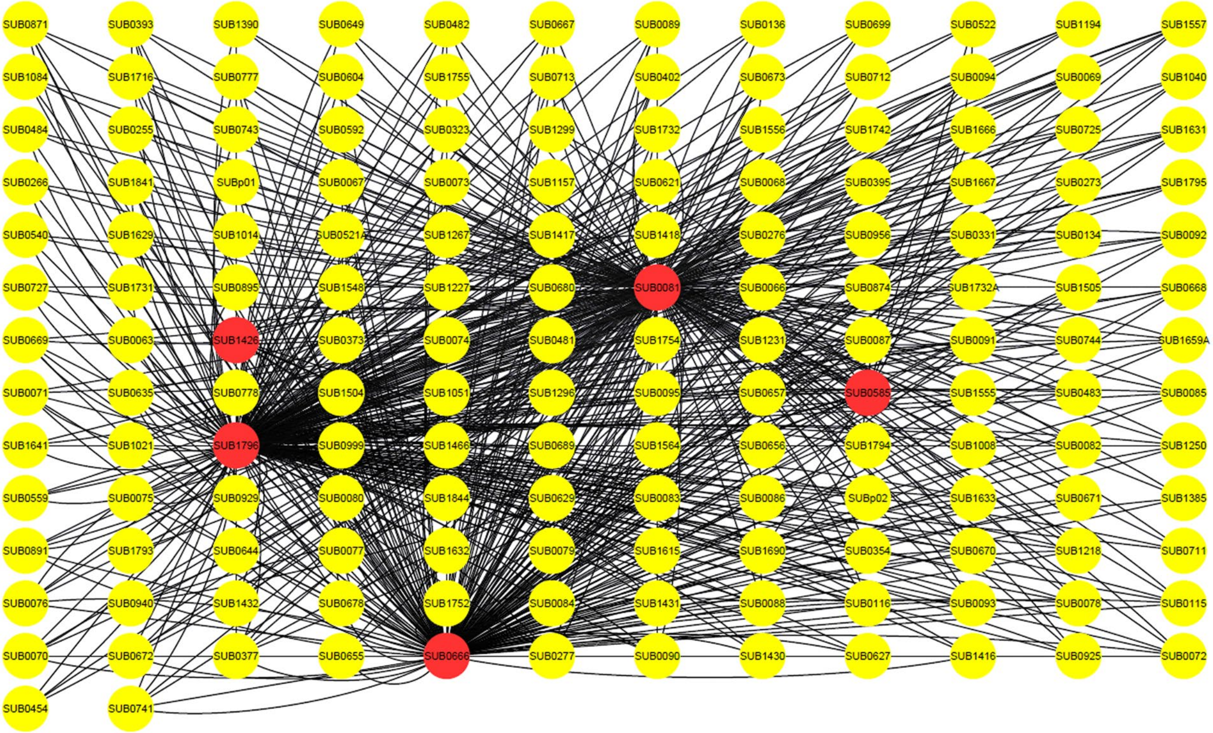

Figure 6. The protein-protein interaction (PPI) network showing 153 Streptococcus uberis proteins (yellow) interacting with the 5 discriminant proteins (red).

Interestingly, the protein network analysis showed that three of the identified proteins (SUB0081, SUB0666 and SUB1796) interact with one another. This may suggest that the functions (ribosome, oxidative phosphorylation and bacterial secretion system) and importantly the expressions of key proteins participating in the differentiation of transmission routes of *S. uberis*, are co-ordinated. The co-ordination implies that the regulatory changes acting on these genes are accumulated over time across strains. Such stringency in constraining expression variance shifts may play an evolutionary role in the definition of different phenotypically-related traits.

The results of this study suggest that MALDI-TOF spectral analysis of clinical mastitis isolates could provide a rapid means of diagnosing the likely route of transmission at the early stages of a mastitis outbreak, as previously suggested by Archer, *et al.*[56]. Given that potentially contagious transmission of *S. uberis* has been identified in two-thirds of commercial herds and it has been found as the dominant transmission route in a third of UK herds[56], there is a clear need for diagnostic tools capable to discriminate between contagious and environmentally acquired infections. Tools based on MALDI-TOF spectral analysis would enable clinicians to identify the most appropriate control measures promptly, during an outbreak of disease. A diagnosis based on MALDI-TOF spectra has the potential to reduce the incidence of clinical disease, reduce associated production losses, reduce the costs associated with treatment of clinical mastitis and improve the efficiency of labour and resource allocation on farm.

In conclusion, the analysis of MALDI-TOF spectral profiles through solutions powered by machine learning, and in particular genetic algorithms, was shown to be useful to predict the contagious or environmental nature of *S. uberis* mastitis. Classifiers developed to target individual farms achieved 97.81% CV accuracy (mean over 19 farms), with a mean Cohen's kappa coefficient of 94%, clearly indicating the possibility to deploy effective diagnostic solutions capable to distinguish between environmental and contagious *S. uberis* strains within a farm. Prediction performance was still high at cross-validation for a classifier trained and tested on the aggregation of the data available from the same 19 farms (CV accuracy 95.88%, kappa 92.62%) but dropped to accuracy 70.67% (kappa 33.80%) when the predictor was externally validated with data from ten additional farms left as holdout. It is unclear at the moment if such degradation may be due inherent proteomic diversity in *S. uberis* populations between herds, and if performance may be improved by simply increasing the amount of data available to train the predictors, for example by including a larger number of farms, or by focusing on ensuring more variation within the training sets. In any case, elucidating the role of specific proteins that have been found discriminatory between contagious and environmental transmission, may provide insights into the underpinning biology of the pathogen. The protein network analysis has also shown the presence of a protein functional network suggesting the existence of a constrained co-evolution of functional pathways and protein expression in participating in differentiating transmission routes of *S. uberis*.

As future endeavour, it may be interesting to investigate whether similar solutions based on the analysis of MALDI-TOF spectra by means of machine learning may be used to develop screening tools to identify early signs of mastitis, or related risk factors. Such a research goal has not been covered yet by our studies, as a significantly different approach is required both in terms of planning and executing data collection and for validating the

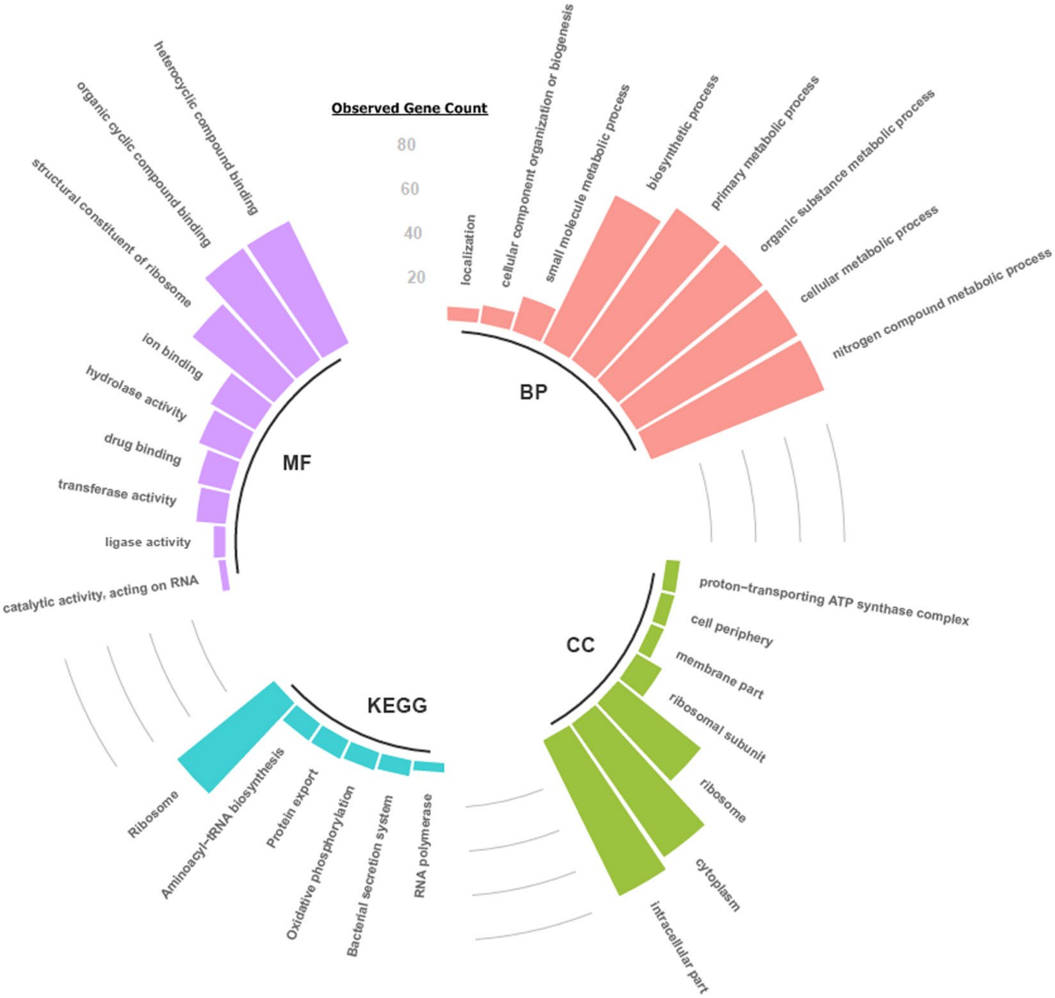

Figure 7. Functional annotation of 158 proteins (5 of interest and 153 interacting with at least two genes of interest) in Streptococcus uberis, based on Gene Ontology and KEGG Pathway.

results. Nevertheless, the analysis of MALDI-TOF peaks has proven successful at discriminating between contagious and environmental strains of infected animals, in particular within individual herds, and one may wonder if a discrimination between healthy and infected individuals may be carried out by looking at similar sets of peaks.

Methods

MLST and MALDI-TOF datasets.

In order to better appreciate the nature of the available data, the following information is provided from previous work[4].

MLST: for the classification of environmental and contagious *S. uberis* strains, samples were cultured by a commercial milk laboratory. Putative *S. uberis* colonies were stored on beads at −80 °C. The gDNA was extracted for MLST sequencing in accordance with the protocol described in previous literature[57]. Clinical cases attributed to isolates of the same multi-locus sequence type (MLST) occurring in different cows in the same herd within a 42 day time period were classified as contagious, whereas cases attributed to isolates occurring only once in any cow of any herd were classified as environmental.

MALDI-TOF spectra: all isolates classified as contagious or environmental were incubated on blood agar at 37 °C for 18 to 24 hours. Protein extraction was conducted as previously described[58] by means of a Time-of-flight (TOF) MALDI mass spectrometer (Bruker Daltonics, Billerica, MA). For each *S. uberis* isolate, 6 technical replicate profiles were generated from 40 desorptions per replicate. Spectra were compared visually using Biotyper 3.1 (Bruker Daltonics). Those with insufficient resolution, low intensity, or substantial background noise were removed. Technical replicates were further compared using composite correlation indices (CCI) to remove dissimilar spectra with CCI < 0.99[59].

In this work, to extract each peak list, the following steps were applied in ClinProTools 3.0[15]:

a) baseline subtraction: using the Top Hat baseline (minimal baseline width: 10%)[60];
b) normalisation: to the total ion count, leading to spectral intensities in the [0–1] range[15];
c) recalibration of the m/z values: using as reference masses those appearing in at least 30% of the spectra and setting 1000 ppm as the maximal peak shift[15];

d) total average spectrum calculation: using weighted contributions from the available replicates[15,61];
e) average peak list calculation: the calculation was applied on the total average spectrum rather than on each single spectrum[15];
f) peak picking on the total average spectrum: using resolution: 800 for spectrum smoothing, signal to noise threshold: 5.00, and 0.000% relative threshold base peak to include all peaks[15];
g) peak normalization: to give the same relevance (weight) to all peaks within the classification/prediction models[15];
h) mass range filter: the mass range of the spectra was limited to 4–10 KDa.

Parameters used for the classification methods. The GA default parameters (as suggested by the ClinProTools software) were as follows: initial number of peak combinations (INPC): automatic detection; maximal number of peaks (MNP, maximum number of peaks to be included in the combination): 5; maximum number of generations (MNG, number of GA iterations to identify the optimal result): 50; mutation rate (MR): 0.2; crossover rate (CR): 0.5 (mutation and crossover control how new candidate combinations are created starting from those tested in the current generation); number of neighbours for KNN (KNN, number of neighbours considered by the KNN method to determine the distance of a new observation from an existing class): 50. The optimised set of GA parameter was: INPC: 125, MNP: 19, MNG: 50, MR: 0.2, CR: 0.5 and KNN: 3.

The SNN parameters were: maximum number of peaks (MNP): automatic detection; upper limit of cycles (ULC): 1000; number of prototypes (NP): automatic detection.

The QC method had only one controllable parameter: the maximum number of peaks (MNP), which was set to automatic detection.

Methods for cross and external validation. For intra-farm analysis, cross validation of each classifier as performed using 80% of the available spectra for training and the remaining 20% for validation. The procedure was repeated 10 times, each time randomising the extraction of spectra for the training set.

For inter-farm analysis cross validation of the global classifier was performed using 80% of the spectra obtained by aggregation of data from the 19 farms for training and the remaining 20% for validation. The procedure was repeated 10 times, each time randomising the extraction of spectra for the training set. For external validation, 100% of the available aggregated spectra was used for training, and the spectra from the 10 holdout farms were used for validation. The entire procedure was repeated 10 times for the GA classifier, because of the random components present in the method.

Methods for biomarker characterisation. To find correspondences between the individual peaks of the peak lists and actual proteins of *S. uberis, the* NCBI protein database was used, using a maximum of 0.2% difference in mass as threshold for the successful identification of a correspondence.

The Simple Modular Architecture Research Tool (SMART) was used to find the domains of the proteins. The sequence of hypothetical proteins was pasted into the SMART database search panel. Outlier homologues, as well as homologues of known structure, PFAM domains, signal peptides and internal repeats, were checked. The next step was identifying the homology of these hypothetical proteins in Basic Local Alignment Search Tool (BLASTp). To further investigate the function of the identified proteins (peaks that were cross-referenced with the NCBI *S. uberis* protein database) we studied their protein networks. This was done by analysing the protein-protein interaction (PPI) dataset of *S. uberis*, obtained from STRING database[62] with Cytoscape 3.6.1[63]. Only genes showing at least two interactions were kept and their functions were detailed along the aspects of biological process, molecular function, cellular component and KEGG pathway. The three dimensional homology models of *S. uberis* were obtained by retrieving the aminoacidic FASTA sequence from Uniprot (https://www.uniprot.org/uniprot/B9DSW3) and using Swiss-Model[64] to search for templates and build the models. The visualisation shown in Fig. 5 was created using UCSF Chimera[65].

References

1. Todhunter, D. A., Smith, K. L. & Hogan, J. S. Environmental Streptococcal Intramammary Infections of the Bovine Mammary Gland1. *Journal of dairy science* **78**, 2366–2374 (1995).
2. Green, M. J., Leach, K. A., Breen, J. E., Green, L. E. & Bradley, A. J. National intervention study of mastitis control in dairy herds in England and Wales. *The Veterinary Record* **160**, 287–293 (2007).
3. Zadoks, R. N. *et al.* Clinical, epidemiological and molecular characteristics of Streptococcus uberis infections in dairy herds. *Epidemiol Infect* **130**, 335–349 (2003).
4. Davies, P. L. *et al.* Molecular epidemiology of Streptococcus uberis clinical mastitis in dairy herds: strain heterogeneity and transmission. *Journal of clinical microbiology* **54**, 68–74 (2016).
5. Casadesús, J. & Low, D. Epigenetic gene regulation in the bacterial world. *Microbiology and molecular biology reviews* **70**, 830–856 (2006).
6. Seib, K. L. *et al.* Specificity of the ModA11, ModA12 and ModD1 epigenetic regulator N6-adenine DNA methyltransferases of Neisseria meningitidis. *Nucleic acids research* **43**, 4150–4162 (2015).
7. Badie, G., Heithoff, D. M., Sinsheimer, R. L. & Mahan, M. J. Altered levels of Salmonella DNA adenine methylase are associated with defects in gene expression, motility, flagellar synthesis, and bile resistance in the pathogenic strain 14028 but not in the laboratory strain LT2. *Journal of bacteriology* **189**, 1556–1564 (2007).
8. Baseggio, N., Mansell, P. D., Browning, J. W. & Browning, G. F. Strain differentiation of isolates of streptococci from bovine mastitis by pulsed-field gel electrophoresis. *Molecular and cellular probes* **11**, 349–354 (1997).
9. Kallow, W. *et al.* Microbial strain characterisation by MALDI-TOF MS-possibilities and limits (2006).
10. Nakano, S. *et al.* Development and evaluation of MALDI-TOF MS-based serotyping for Streptococcus pneumoniae. *European Journal of Clinical Microbiology & Infectious Diseases* **34**, 2191–2198 (2015).

11. Xiao, D., Zhao, F., Zhang, H., Meng, F. & Zhang, J. Novel strategy for typing Mycoplasma pneumoniae isolates by use of matrix-assisted laser desorption ionization–time of flight mass spectrometry coupled with ClinProTools. *Journal of clinical microbiology* **52**, 3038–3043 (2014).
12. Rizzardi, K., Wahab, T. & Jernberg, C. Rapid subtyping of Yersinia enterocolitica by matrix-assisted laser desorption ionization–time of flight mass spectrometry (MALDI-TOF MS) for diagnostics and surveillance. *Journal of clinical microbiology* **51**, 4200–4203 (2013).
13. Barreiro, J. R. *et al*. Identification of subclinical cow mastitis pathogens in milk by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Journal of dairy science* **93**, 5661–5667 (2010).
14. Russell, S. J. & Norvig, P. *Artificial Intelligence: A Modern Approach*. Third Edition edn (Prentice Hall, 2010).
15. ClinProTools 3.0 User Manual (MA, Bremen, 2011).
16. Holland, J. H. Adaptation in natural and artificial systems. An introductory analysis with application to biology, control, and artificial intelligence. *Ann Arbor, MI: University of Michigan Press*, 439–444 (1975).
17. Hammer, B., Strickert, M. & Villmann, T. Supervised neural gas with general similarity measure. *Neural Processing Letters* **21**, 21–44 (2005).
18. Wittwer, M. *et al*. Genetic diversity and antibiotic resistance patterns in a Campylobacter population isolated from poultry farms in Switzerland. *Applied and Environmental Microbiology* **71**, 2840–2847 (2005).
19. Sordelli, D. O. *et al*. Capsule Expression by Bovine Isolates of Staphylococcus aureus from Argentina: Genetic and Epidemiologic Analyses. *Journal of Clinical Microbiology* **38**, 846–850 (2000).
20. Falush, D. *et al*. Traces of human migrations in Helicobacter pylori populations. *science* **299**, 1582–1585 (2003).
21. Sharp, P. M. Identification of genes encoding ribosomal protein L33 from Bacillus licheniformis, Thermus thermophilus and Thermotoga maritima. *Gene* **139**, 135–136, https://doi.org/10.1016/0378-1119(94)90537-1 (1994).
22. Panina, E. M., Mironov, A. A. & Gelfand, M. S. Comparative genomics of bacterial zinc regulons: Enhanced ion transport, pathogenesis, and rearrangement of ribosomal proteins. *Proceedings of the National Academy of Sciences* **100**, 9912–9917, https://doi.org/10.1073/pnas.1733691100 (2003).
23. O'Rourke, D. Nutrition and udder health in dairy cows: a review. *Irish Veterinary Journal* **62**, S15 (2009).
24. Makarova, K. S., Ponomarev, V. A. & Koonin, E. V. Two C or not two C: recurrent disruption of Zn-ribbons, gene duplication, lineage-specific gene loss, and horizontal gene transfer in evolution of bacterial ribosomal proteins. *Genome Biology* **2**, research0033.0031, https://doi.org/10.1186/gb-2001-2-9-research0033 (2001).
25. Gil, R., Silva, F. J., Peretó, J. & Moya, A. Determination of the core of a minimal bacterial gene set. *Microbiology and Molecular Biology Reviews* **68**, 518–537 (2004).
26. Georrge, J. J. & Umrania, V. V. Subtractive Genomics Approach to Identify Putative Drug Targets and Identification of Drug-like Molecules for Beta Subunit of DNA Polymerase III in Streptococcus Species. *Applied Biochemistry and Biotechnology* **167**, 1377–1395, https://doi.org/10.1007/s12010-012-9620-0 (2012).
27. Zhu, Y.-Z. *et al*. Gene expression profiling-based in silico approach to identify potential vaccine candidates and drug targets against B. pertussis and B. parapertussis. *OMICS A Journal of Integrative Biology* **12**, 161–169 (2008).
28. Rao Reddy Neelapu, N. & Pavani, T. Identification of novel drug targets in HpB38, HpP12, HpG27, Hpshi470, HpSJM180 strains of Helicobacter pylori: an in silico approach for therapeutic intervention. *Current drug targets* **14**, 601–611 (2013).
29. Fan, X., Tang, X., Yan, J. & Xie, J. Identification of idiosyncratic Mycobacterium tuberculosis ribosomal protein subunits with implications in extraribosomal function, persistence, and drug resistance based on transcriptome data. *Journal of Biomolecular Structure and Dynamics* **32**, 1546–1551, https://doi.org/10.1080/07391102.2013.826143 (2014).
30. Natori, Y. *et al*. A fail-safe system for the ribosome under zinc-limiting conditions in Bacillus subtilis. *Molecular Microbiology* **63**, 294–307, https://doi.org/10.1111/j.1365-2958.2006.05513.x (2007).
31. Nanamiya, H. & Kawamura, F. Towards an Elucidation of the Roles of the Ribosome during Different Growth Phases in *Bacillus*. *Bioscience, Biotechnology, and Biochemistry* **74**, 451–461, https://doi.org/10.1271/bbb.90859 (2010).
32. Shoji, S., Dambacher, C. M., Shajani, Z., Williamson, J. R. & Schultz, P. G. Systematic Chromosomal Deletion of Bacterial Ribosomal Protein Genes. *Journal of Molecular Biology* **413**, 751–761, https://doi.org/10.1016/j.jmb.2011.09.004 (2011).
33. Forsyth, R. *et al*. A genome-wide strategy for the identification of essential genes in Staphylococcus aureus. *Molecular microbiology* **43**, 1387–1400 (2002).
34. Koul, A. *et al*. Diarylquinolines target subunit c of mycobacterial ATP synthase. *Nature chemical biology* **3**, 323 (2007).
35. Petrella, S. *et al*. Genetic basis for natural and acquired resistance to the diarylquinoline R207910 in mycobacteria. *Antimicrobial agents and chemotherapy* **50**, 2853–2856 (2006).
36. Aguilar-Ayala, D. A. *et al*. *In vitro* activity of bedaquiline against rapidly growing nontuberculous mycobacteria. *Journal of Medical Microbiology* **66**, 1140–1143, https://doi.org/10.1099/jmm.0.000537 (2017).
37. Sambongi, Y. *et al*. Mechanical rotation of the c subunit oligomer in ATP synthase (F0F1): direct observation. *Science* **286**, 1722–1724 (1999).
38. Fillingame, R. H., Oldenburg, M. & Fraga, D. Mutation of alanine 24 to serine in subunit c of the Escherichia coli F1F0-ATP synthase reduces reactivity of aspartyl 61 with dicyclohexylcarbodiimide. *Journal of Biological Chemistry* **266**, 20934–20939 (1991).
39. Wang, J.-Y., Sarker, A. H., Cooper, P. K. & Volkert, M. R. The single-strand DNA binding activity of human PC4 prevents mutagenesis and killing by oxidative DNA damage. *Molecular and cellular biology* **24**, 6084–6093 (2004).
40. Nascimento, A. *et al*. Comparative genomics of two Leptospira interrogans serovars reveals novel insights into physiology and pathogenesis. *Journal of bacteriology* **186**, 2164–2172 (2004).
41. Roy, C. R. & Cherfils, J. Structure and function of Fic proteins. *Nature Reviews Microbiology* **13**, 631, https://doi.org/10.1038/nrmicro3520 (2015).
42. Komano, T., Utsumi, R. & Kawamukai, M. Functional analysis of the fic gene involved in regulation of cell division. *Research in Microbiology* **142**, 269–277, https://doi.org/10.1016/0923-2508(91)90040-H (1991).
43. Worby, C. A. *et al*. The Fic Domain: Regulation of Cell Signaling by Adenylylation. *Molecular Cell* **34**, 93–103, https://doi.org/10.1016/j.molcel.2009.03.008 (2009).
44. Eijsink, V. G. *et al*. Production of class II bacteriocins by lactic acid bacteria; an example of biological warfare and communication. *Antonie Van Leeuwenhoek* **81**, 639–654 (2002).
45. Ward, P. N. *et al*. Evidence for niche adaptation in the genome of the bovine pathogen Streptococcus uberis. *BMC Genomics* **10**, 54, https://doi.org/10.1186/1471-2164-10-54 (2009).
46. Hossain, M. *et al*. Virulence related sequences; insights provided by comparative genomics of Streptococcus uberis of differing virulence. *BMC genomics* **16**, 334 (2015).
47. Leigh, J. A. & Field, T. R. Killing of Streptococcus uberis by bovine neutrophils following growth in chemically defined media. *Veterinary research communications* **15**, 1–6 (1991).
48. Grant, R. G. & Finch, J. M. Phagocytosis of Streptococcus uberis by bovine mammary gland macrophages. *Research in veterinary science* **62**, 74–78 (1997).
49. Leigh, J. A., Field, T. R. & Williams, M. R. Two strains of Streptococcus uberis, of differing ability to cause clinical mastitis, differ in their ability to resist some host defence factors. *Research in veterinary science* **49**, 85–87 (1990).

50. Leigh, J. A. & Lincoln, R. A. Streptococcus uberis acquires plasmin activity following growth in the presence of bovine plasminogen through the action of its specific plasminogen activator. *FEMS Microbiology Letters* **154**, 123–129, https://doi.org/10.1111/j.1574-6968.1997.tb12633.x (1997).
51. Chang, C. *et al*. The structure of pyogenecin immunity protein, a novel bacteriocin-like immunity protein from Streptococcus pyogenes. *BMC structural biology* **9**, 75 (2009).
52. Kjos, M., Salehian, Z., Nes, I. F. & Diep, D. B. An extracellular loop of the mannose phosphotransferase system component IIC is responsible for specific targeting by class IIa bacteriocins. *Journal of bacteriology* **192**, 5906–5913 (2010).
53. Johnsen, L., Fimland, G. & Nissen-Meyer, J. The C-terminal domain of pediocin-like antimicrobial peptides (class IIa bacteriocins) is involved in specific recognition of the C-terminal part of cognate immunity proteins and in determining the antimicrobial spectrum. *Journal of Biological Chemistry* **280**, 9243–9250 (2005).
54. Wang, W.-L., Liu, J., Huo, Y.-B. & Ling, J.-Q. Bacteriocin immunity proteins play a role in quorum-sensing system regulated antimicrobial sensitivity of Streptococcus mutans UA159. *Archives of Oral Biology* **58**, 384–390, https://doi.org/10.1016/j.archoralbio.2012.09.001 (2013).
55. Matsumoto-Nakano, M. & Kuramitsu, H. K. Role of Bacteriocin Immunity Proteins in the Antimicrobial Sensitivity of Streptococcus mutans. *Journal of Bacteriology* **188**, 8095–8102, https://doi.org/10.1128/jb.00908-06 (2006).
56. Archer, S. C., Bradley, A. J., Cooper, S., Davies, P. L. & Green, M. J. Prediction of Streptococcus uberis clinical mastitis risk using Matrix-assisted laser desorption ionization time of flight mass spectrometry (MALDI-TOF MS) in dairy herds. *Preventive veterinary medicine* **144**, 1–6 (2017).
57. Leigh, J. A., Egan, S. A., Ward, P. N., Field, T. R. & Coffey, T. J. Sortase anchored proteins of Streptococcus uberis play major roles in the pathogenesis of bovine mastitis in dairy cattle. *Veterinary research* **41**, 63 (2010).
58. Ruppitsch, W. *et al*. Defining and evaluating a core genome multilocus sequence typing scheme for whole-genome sequence-based typing of Listeria monocytogenes. *Journal of clinical microbiology* **53**, 2869–2876 (2015).
59. Arnold, R. J. & Reilly, J. P. Fingerprint matching of E. coli strains with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry of whole cells using a modified correlation approach. *Rapid Communications in Mass Spectrometry* **12**, 630–636 (1998).
60. Serra, J. *Image analysis and mathematical morphology*. (Academic Press, Inc., 1983).
61. Morris, J. S., Coombes, K. R., Koomen, J., Baggerly, K. A. & Kobayashi, R. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics* **21**, 1764–1775 (2005).
62. Szklarczyk, D. *et al*. STRINGv10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, gku1003 (2014).
63. Shannon, P. *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498–2504 (2003).
64. Waterhouse, A. *et al*. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research* **46**, 296–303 (2018).
65. Pettersen, E. F. *et al*. UCSF Chimera–a visualization system for exploratory research and analysis. *Journal of Computational Chemistry* **25**, 1605–1612 (2004).
66. R Core Team *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, https://www.R-project.org (2017).

Author Contributions

M.J.G. and A.J.B. provided the original data. T.D., M.J.G., R.D.E. and A.J.B. conceived and designed the data analysis procedures. N.E., R.D.E., T.D. and B.J. carried on the data analysis. T.D., N.E. and P.L.D. wrote the manuscript. All authors reviewed the manuscript.

Additional Information

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.