BMC Genomics

Phenotype and multi-omics comparison of *Staphylococcus* and *Streptococcus* uncovers pathogenic traits and predicts zoonotic potential

Niels A. Zondervan[1], Vitor A. P. Martins dos Santos[1,2], Maria Suarez-Diez[1†] and Edoardo Saccenti[1*†]

Abstract

Background: *Staphylococcus* and *Streptococcus* species can cause many different diseases, ranging from mild skin infections to life-threatening necrotizing fasciitis. Both genera consist of commensal species that colonize the skin and nose of humans and animals, and of which some can display a pathogenic phenotype.

Results: We compared 235 *Staphylococcus* and 315 *Streptococcus* genomes based on their protein domain content. We show the relationships between protein persistence and essentiality by integrating essentiality predictions from two metabolic models and essentiality measurements from six large-scale transposon mutagenesis experiments. We identified clusters of strains within species based on proteins associated to similar biological processes. We built Random Forest classifiers that predicted the zoonotic potential. Furthermore, we identified shared attributes between of *Staphylococcus aureus* and *Streptococcus pyogenes* that allow them to cause necrotizing fasciitis.

Conclusions: Differences observed in clustering of strains based on functional groups of proteins correlate with phenotypes such as host tropism, capability to infect multiple hosts and drug resistance. Our method provides a solid basis towards large-scale prediction of phenotypes based on genomic information.

Keywords: Staphylococcus, Streptococcus, Multi-omics, Comparison, Pathogenic, Traits, Prediction, Phenotype, Host-trophism, Zoonotic

Background

Species from the genera *Staphylococcus* and *Streptococcus* are mostly commensals that live as part of the microbiota of various animals and humans [1]. Some of them are opportunistic pathogens, displaying a pathogenic phenotype when the immune system of the host is compromised or the epithelial barrier is damaged [2–5].

Few comparative genomic studies have been performed to analyse the evolution and the pathogenesis of *Staphylococcus* and *Streptococcus* species: the comparisons of the genomes of 11 *Staphylococcus* species determined that horizontal gene transfer of virulence factors is an important factor in adaptation of *S. aureus* to humans [6]; another study showed that protein domain based metabolic diversity among *Streptococcus* species could be used to identify differences in the metabolism of the highly pathogenic serotype 2 *S. suis* compared to other *Streptococci* [7]. Another study confirmed these results and showed that metabolic capability predicted using genome scale models (GEMs)

* Correspondence: edoardo.saccenti@wur.nl
†Maria Suarez-Diez and Edoardo Saccenti these authors jointly supervised this work.
[1]Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Stippeneng 4, 6708WE Wageningen, Netherlands
Full list of author information is available at the end of the article

Zondervan *et al. BMC Genomics*   (2021) 22:102

Page 2 of 21

could be used to identify *Streptococcus* strain specific bio-markers and metabolic determinants of virulence [8].

Protein domains and protein-domain architectures have been shown to be a fast and efficient method to define groups of functionally equivalent proteins that were used for comparative genomic studies [9, 10], including *Staphylococcus* and *Streptococcus* [11–13]. However, at the best of our knowledge, no work exists focusing on similarities and differences within and between *Staphylococcus* and *Streptococcus* genomes.
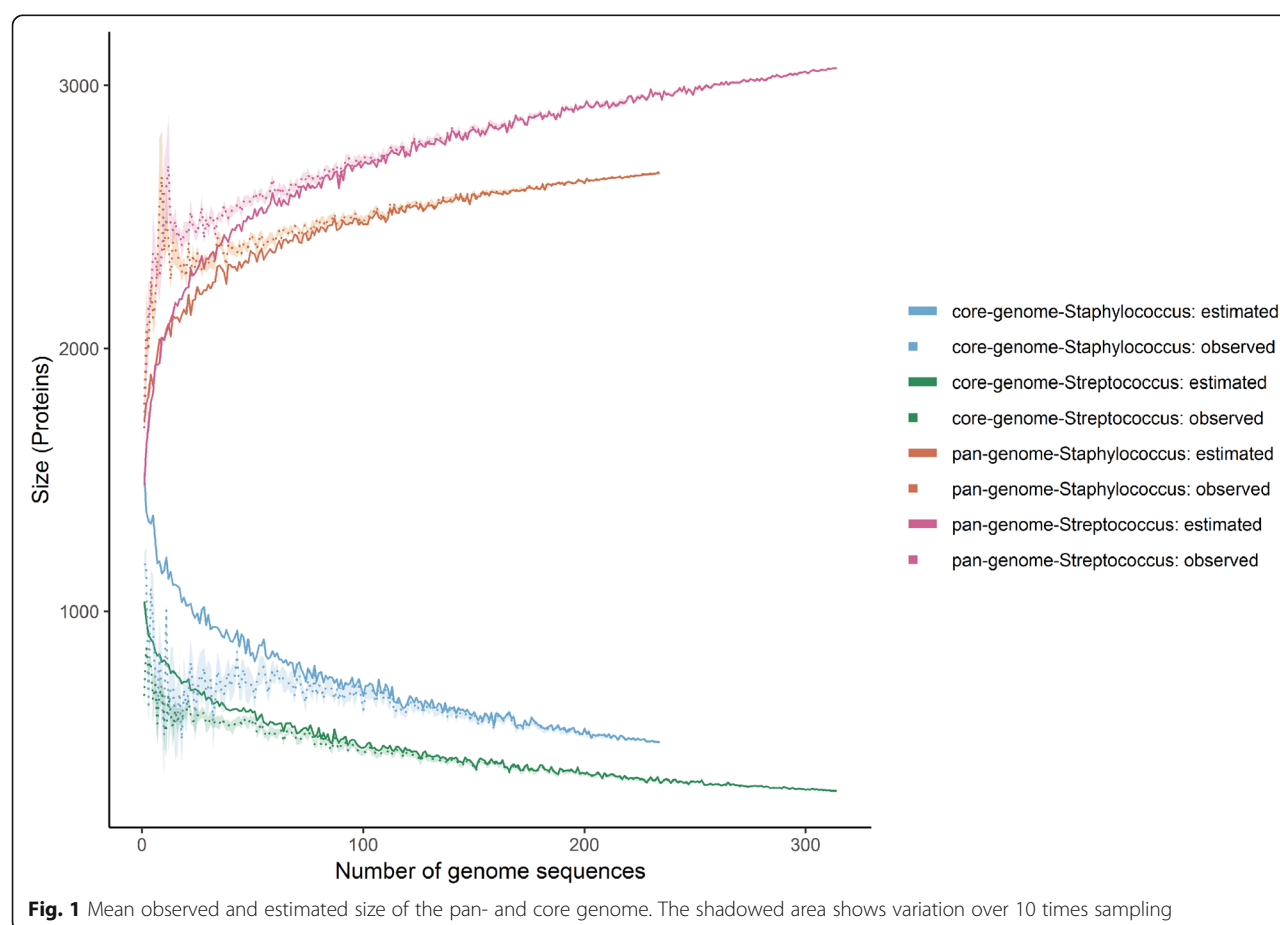
In this study we performed a comparative analysis of 235 and 315 fully sequenced *Staphylococci* and *Streptococci* genomes by annotating their proteins based on their domain content. We integrated this protein annotation with genome-scale metabolic-modelling predictions, transcriptomic and transposon-mutagenesis data sets to study gene essentiality and persistence. All annotation used in this paper as well as GO information is based on genomics annotation from databases based mainly on bacterial genomics studies. In this paper we compare within and between *Staphylococcus* and *Streptococcus* species with the objective to identify both difference and similarities in genomic properties as well as in specific combinations of genes that give rise to

pathogenic phenotypes. We compared the clustering of *Staphylococcus* and *Streptococcus* genomes based on proteins selected using on Gene Ontology (GO) terms associated with clinical phenotypes such drug resistance, pathogenesis, and tissue and host tropism. Furthermore, we used the functional grouping of proteins to predict zoonotic potential of *S. suis* and *S. agalactiae*, that is their ability to infect multiple hosts including humans. Finally, we compared *S. aureus* and *S. pyogenes* to identify the genomic basis for their shared ability to cause severe bacterial infections like necrotizing fasciitis. Our results are compared throughout the paper with findings from literature.

Results

Pan- and core genome analysis

The size of the pan- and core genomes of *Staphylococcus* and *Streptococcus* was determined based on protein domain content (Fig. 1). The pangenome contains all proteins present in the analysed genomes. The core genome contains only proteins that are present in all genomes and represents their genomic essence [14]. The ratio of the sizes of the core- and pan genome are 0.22 (557/

Fig. 1 Mean observed and estimated size of the pan- and core genome. The shadowed area shows variation over 10 times sampling

Zondervan *et al. BMC Genomics*   (2021) 22:102

Page 3 of 21

2563) for *Staphylococcus* and 0.17 (458/2725) for *Streptococcus*.

A Heaps' regression model was used to estimate the closedness of the pangenome [15]. The closedness of the pangenome represents how much the addition of more genome sequences is expected to increase the number of proteins in the pangenome. For both *Staphylococcus* ($\alpha = 1.10 \pm 0.02$) and *Streptococcus* ($\alpha = 1.12 \pm 0.01$) the pangenome was found to be closed (*i.e* few new genes are added as news strains are discovered/sequenced). Additional plots of the estimated pan- and core genomes size and the Heaps' regression model can be found in supplementary material (see Additional file 5).

Protein persistence and essentiality

Persistence of proteins over all *Staphylococcus* and all *Streptococcus* species was calculated. Protein persistence data was combined with model predictions of essentiality and experimentally determined essentiality data.Experimentally determined essentiality (labelled as EXP) is available for growth on rich media resembling in vivo conditions. GEMs predictions were made using minimal media conditions for all combinations of carbon, nitrogen, sulphur, and phosphorus sources. Simulations on rich media conditions were therefore indirectly performed since all rich media compounds are present in the models as exchange reactions and all combinations of these exchange reactions functioning as carbon, nitrogen, sulphur and phosphor sources were tested for essentiality. We used GEM to predict gene essentiality for *Staphylococcus aureus* NTCTC 8325 and *Streptococcus pyogenes* M49. The total number of medium combinations based on C, N, S, P sources was 12,432 for *Staphylococcus* and 714 for *Streptococcus*. The number of tested conditions for Staphylococcus is much larger than the Streptococcus model since the Staphylococcus model can use all amino acids as alternative nitrogen source through deamination, greatly increasing the number of minimal media combinations. The Staphylococcus model can use all amino acids as alternative nitrogen source through deamination, greatly increasing the number of minimal media combinations Protein persistence, in silico predictions of essential and in vitro essentiality data for *Staphylococcus* and *Streptococcus* were integrated based on their associated locus tags. Both GEM based and experimentally determined essentiality correlated with a high persistence, while essentiality by both criteria is associated with an even higher persistence (see Table 1 and Fig. 2). Proteins experimentally determined or GEM predicted to be essentiality are significantly different from the average protein persistence (Student's *t*-test, *p*-value = 5×10^{-14}) for both *Staphylococcus* and *Streptococcus*.

Table 1 Persistence of *Staphylococcus* (Staph.) and *Streptococcus* (Strep.) for all proteins, proteins associated to Genome Metabolic model (GEM) essential genes and experimentally (EXP) determined essential genes

Group	Avg persistence Staph.	Avg persistence Strep.
All	0.60 ± 0.44 ($N = 2655$)	0.42 ± 0.42 ($N = 3047$)
GEM-essential	0.94 ± 0.14 ($N = 153$)	0.98 ± 0.09 ($N = 225$)
Exp-Essential	0.97 ± 0.03 ($N = 411$)	0.97 ± 0.12 ($N = 254$)
EXP&GEM-Essential	0.94 ± 0.01 ($N = 46$)	0.98 ± 0.00 ($N = 113$)

Variability of gene expression and gene essentiality

Essentiality and domain persistence information for *Staphylococcus* was combined with the variability of transcription (measured by $\log_2$ fold changes). The variability in expression for experimentally determined essential and non-essential genes as well as for persistent and non-persistent genes were compared (Fig. 3). The fold change transcription levels of experimentally determined essential genes are significantly less variable than the transcription levels of non-essential genes (Student's *t*-test, *p*-value = 5×10^{-14}) as well as for persistent genes (Student's *t*-test, *p*-value = 0.000124)).

Functional analysis of pathogenesis and pathogenicity

For this analysis, we filtered proteins from *Staphylococcus* and *Streptococcus* on their association to 17 genome ontology (GO) biological process terms associated to pathogenesis and pathogenicity but cannot make predictions for rich media conditions were not all components are known or incorporated in the model. Filtering included all proteins associated to either the 17 main GO terms or any of their descendent terms. For all GO terms, proteins were found in both *Staphylococcus* and *Streptococcus* (Table 2). The ratio of proteins per GO function to the total number of proteins is similar for *Staphylococcus* and *Streptococcus* except for the group 'Biological adhesions' which has a larger fraction of proteins associated in *Streptococcus* than in *Staphylococcus*.

Functional trees, PCA and t-SNE plots were used to compare the (dis-)similarity in clustering of the genomes based on functional groups of proteins compared to clustering based on all proteins. Dissimilarity was calculated using the Euclidean distances of genomes in the functional trees and by scaling these distances to values between 0 and 1 to make them comparable.

Functional trees, PCA plots and t-SNE plots for *Staphylococcus* can be found in supplementary material [Additional files 6, 7 and 8]. Functional trees, PCA plots and t-SNE pots for *Streptococcus* can be found in supplementary material (Additional files 9, 10 and 11). PCA plots and t-SNE plots for *Staphylococcus* and *Streptococcus* species combined can be found in supplementary material (Additional files 12 and 13).
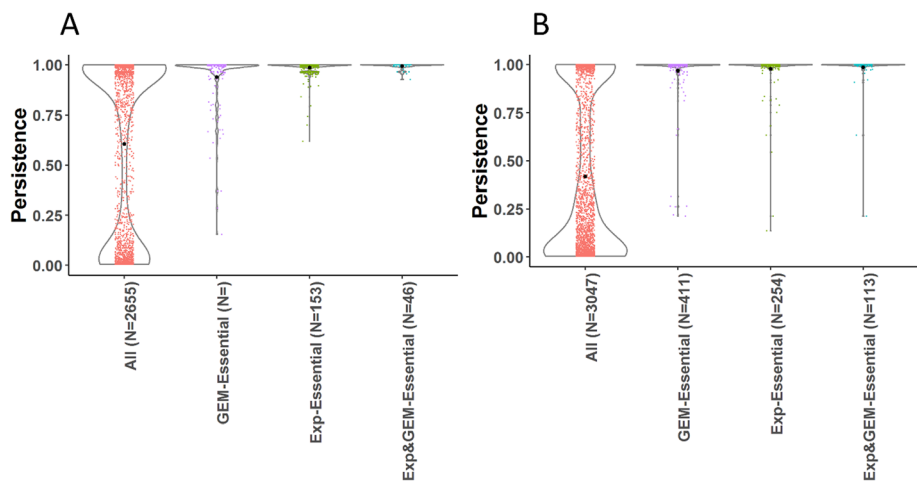
Fig. 2 Protein persistence. **a** Staphylococcus, **b** Streptococcus. Group labels: All = all proteins, GEM = in silico predicted to be essential using a Genome Scale Metabolic model, Exp-Essential = experimentally determined to be essential. Combined group strains

Heatmaps were used to investigate which proteins are absent for each species. Each of these analyses and visualization methods has their own strength and weaknesses in showing the differences in clustering. In the following we highlight some of the differences in clustering of *Staphylococcus* and *Streptococcus* genomes based on proteins annotated per GO term as compared to clustering based on all proteins.

Correlation between GO functional groups of proteins

We calculated the correlation between functional trees in order to compare the similarity in clustering per GO functional group of proteins (Fig. 4a-b). The correlation between functional trees is higher for children and parent

GO terms as well as for GO terms with similar functions such as 'drug metabolic process 'and 'response to drug'. In general, we see that functional trees based on fewer proteins have a lower correlation than functional trees based on many proteins. These results were expected since fewer proteins means less information to separate strains resulting in merging of branches in the tree. An interesting exception to this rule is the 'symbiont process' functional tree which has the lowest correlation with other functional trees for *Staphylococcus* even though there is a high number of proteins associated to this GO term.

There are some notable differences when comparing the correlation between functional trees for *Staphylococcus* and *Streptococcus*. For *Staphylococcus*, the 'pathogenesis'
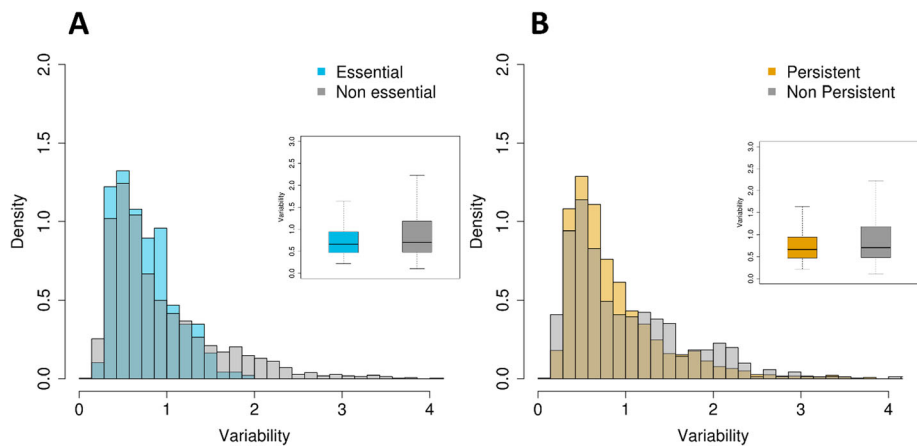
Fig. 3 a Transcriptional variability of essential and non-essential genes in *Staphylococcus*. Box plots show Variability values for both groups. Difference between mean values is significant (*p*-val < 0.01). **b** Transcriptional variability of persistent and non-persistent genes (genes with persistence lower or higher than 0.95, respectively). Box plots show Variability values for both groups. Difference between mean values is significant (*p*-val < 0.01)

Table 2 Number of proteins in the pangenome of *Staphylococcus* and the pangenome of *Streptococcus* per GO term. Root ontology terms, terms without a parent, are marked in their description with an asterixis (*). GO terms are order as such that descendent GO terms are shown below their parent

Filter	Description	Staph	Strep
	All proteins	2655	3047
GO:0008150	Biological process	1974	2222
GO:0008152	*Metabolic process	1661	1871
GO:0017144	Drug metabolic process	59	77
GO:0042493	Response to drug	56	60
GO:0023052	*Signalling	217	280
GO:0065007	*Biological regulation	823	929
GO:0022610	*Biological adhesion	70	147
GO:0044406	Adhesion of symbiont to host	2	3
GO:0051704	Multi-organism process	348	456
GO:0044419	Inter species interaction between organisms	210	309
GO:0042710	Biofilm formation	9	10
GO:0098743	Cell aggregation	7	10
GO:0044403	Symbiont process	150	202
GO:0009372	Quorum sensing	7	7
GO:0035821	Modification of morphology or physiology of other organism	45	74
GO:0009405	Pathogenesis	65	115

clusters together with the functional tree 'modification of morphology or physiology of other organisms'. For *Streptococcus,* the functional tree of 'pathogenesis' clusters together with the functional tree of 'biological adhesion'. Many 'modification of host morphology' proteins in *Staphylococcus* are also associated to the GO term 'pathogenesis' while many 'biological adhesion' proteins in *Streptococcus* are associated to the GO term 'pathogenesis'. These results could indicate that modification of host morphology is important for the pathology of *Staphylococcus* strains while biological adhesion is more important for the pathology of *Streptococcus.*

Horizontal gene transfer of proteins related to pathogenesis

The PCA plot based on all proteins combining *Staphylococcus* and *Streptococcus* genomes supplementary material (Additional files 12 and 13), shows genomes of the same species to cluster together as we would expect (Fig. 5a). The PCA plot based on presence/absence of proteins involved in Response to drug (GO:0042493) shows genomes are not always separated on the species level, however, there is a clear separation between *Staphylococcus* and *Streptococcus* genomes (Fig. 5b). However, both in the PCA (Fig. 5c) and in t-SNE plots based on proteins associated to the GO term 'Pathogenesis' proteins, *Staphylococcus* and *Streptococcus* species cluster together.

Analysis of the presence and absence of proteins associated to 'pathogenesis' reveals that *Staphylococcus sciuri* GCA:002072755 and *Staphylococcus haemolytic us* GCA: 001611955 only contain one pathogenesis protein (PF04647) that is not present in any *Streptococcus* strain. This protein, PF04647 ArgB, is part of a quorum sensing system. Also *S. saprophyticus* GCA:002209265 only contains one protein not present in any *Streptococcus* strain. This protein, PF05480, is a haemolytic protein unique to *Staphylococcus.* Among *Streptococci, S. parauberis* six *S. iniae* and seven *S. thermophilus* strains lack any pathogenesis protein that separates them from *Staphylococcus.*

Domain shuffling of pathogenic proteins

The *Staphylococcus* pangenome contains 52 domains present in 65 proteins associated to the GO term 'pathogenesis' while the *Streptococcus* pangenome contains 88 domains present in 118 proteins associated to the GO term 'pathogenesis'. 20% of pathogenic proteins in *Staphylococcus* and 25% of the pathogenic proteins in *Streptococcus* consist of a few pathogenesis associated domains combined with domains not directly associated to pathogenesis. This implies that domain shuffling might be an important evolutionary factor for these pathogens. In *Staphylococcus* 46% (30/65) and in *Streptococcus* 72% (85/118) of the pathogenesis associated proteins contain multiple domains. This percentage is much higher than the average percentage of multi-domain proteins of 8.9 and 9.2% for *Staphylococcus* and
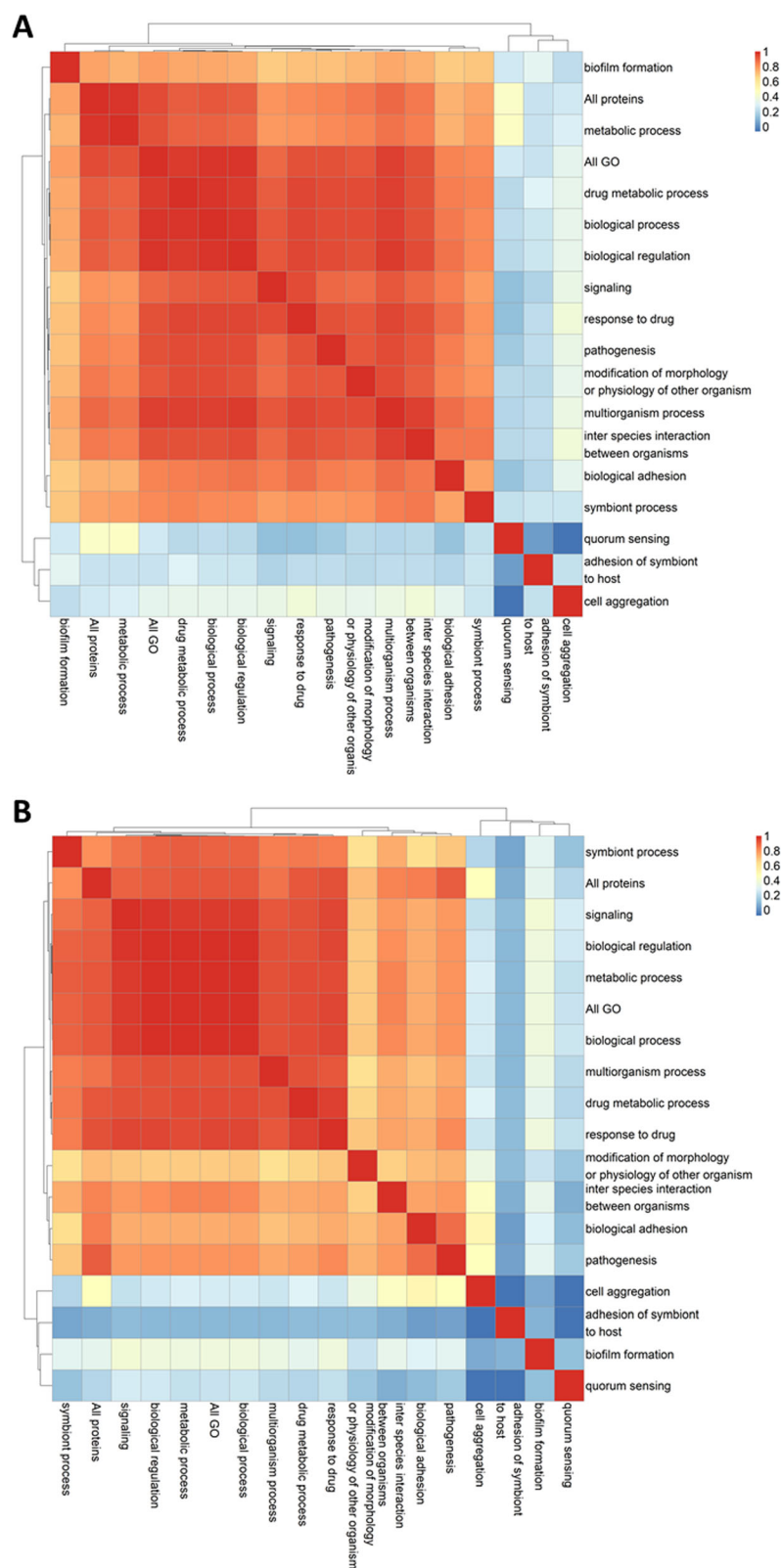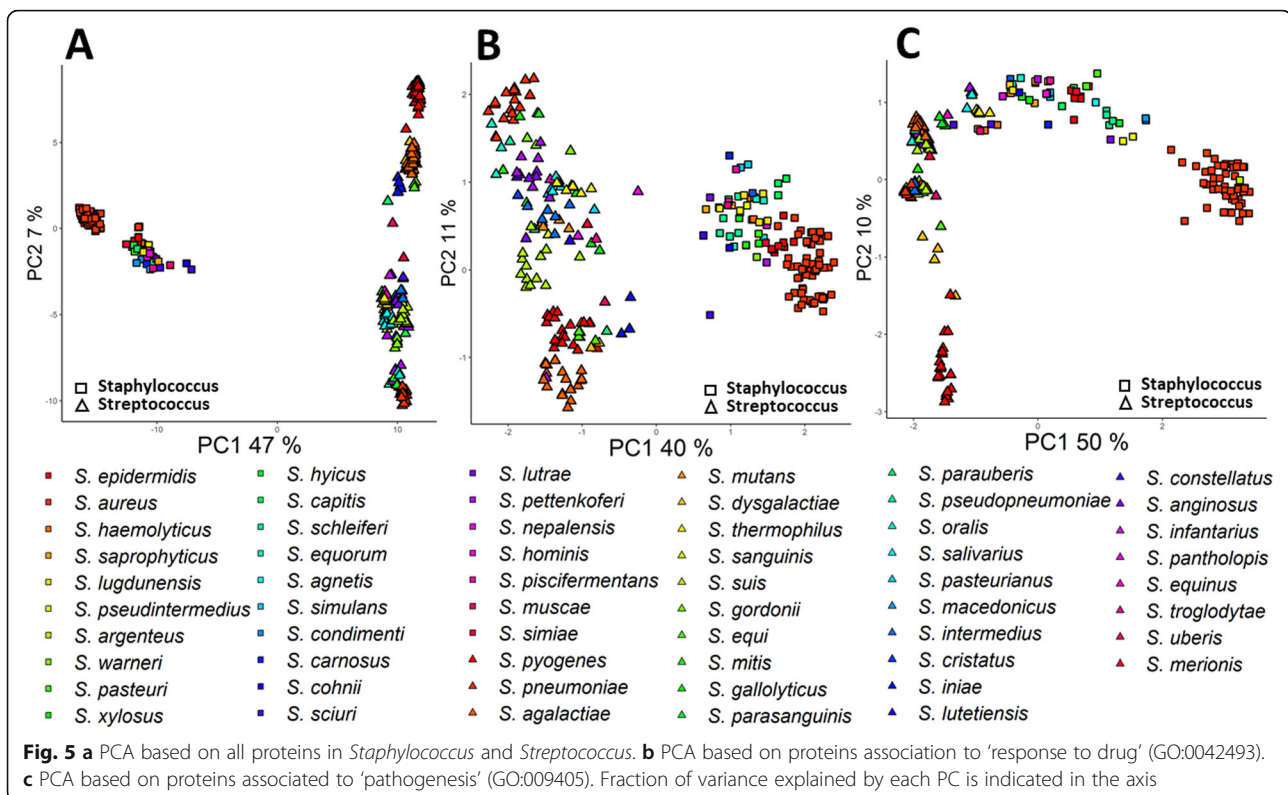
Fig. 4 a Heatmaps of the correlation between *Staphylococcus* functional trees, **b** Heatmaps of the correlation between *Streptococcus* functional trees

Fig. 5 a PCA based on all proteins in *Staphylococcus* and *Streptococcus*. **b** PCA based on proteins association to 'response to drug' (GO:0042493). **c** PCA based on proteins associated to 'pathogenesis' (GO:009405). Fraction of variance explained by each PC is indicated in the axis

Streptococcus respectively. It could be argued that proteins involved in pathogenesis would more often require multiple domains since many of them are cell-wall associated, secreted or contain multiple domains to facilitate interaction between host and the pathogen. The importance of cell wall associated proteins is reflected by the high percentage of 40% of pathogenesis proteins in *Staphylococcus* and 66% *Streptococcus* that contain LPXTG cell-wall anchor domain PF00746. The importance of this domain for pathogenesis was shown in a *S. aureus* mutant with a knockout of *srtA* coding for a class A sortase, which is required for secretion of proteins containing the LPXTG motif. This mutant was unable to form abscess lesions in organ tissues or cause lethal bacteraemia when inoculated in the blood stream of mice [16].

Staphylococcus aureus multi-drug resistance

We investigated the clustering of *S. aureus* genomes in the functional tree associated to the terms "response to drug". We selected the genome of *S. aureus sub species aureus MRSA 252* (GCA:000011505), which is known to be a multiple drug resistant strain [17]. Next, we searched literature for information about drug resistance for eight genomes that cluster together with this strain in the functional tree response to drug. For seven of these strains (JH1, JH9, Mu50, Mu3, T0131, 04–20,981), evidence was found for these strains to be multi-drug resistant as well as identifying two pathogenicity islands as

the cause of their resistance [18–22]. For the last genome (GCA:001640885), no literature or other information could be retrieved. This genome has exactly the same proteins associated to response to drug as the seven strains for which multi drug resistance was reported in literature. Therefore, we can speculate that this strain may have the same multi drug resistance phenotype.

Streptococcus suis pathogenesis zoonotic potential

Large differences in clustering were observed for *S. suis* genomes in the functional trees relating to 'biological adhesion', 'modification of morphology or physiology of other organism' and 'pathogenesis', supplementary material (Additional file 9). *S. suis* genomes form two groups in the functional tree of biological adhesion, and three groups in the functional tree of pathogenesis and 'modification of morphology or physiology of other organism'.

Similarly, different groups can be distinguished in the PCA plot based on these three functional groups, as shown in Fig. 6a-c. We included information from literature on zoonotic species, namely *S. inae, S. agalactiae, S. dysgalactiae S. iniae and, S. equi zooepidemicus* and *S. suis* serotype 2 strains and serotype information and host isolation information for *S. suis* and *S. agalactiae* strains in the labels of Fig. 6a-c. Two *S. suis* clusters can be distinguished in the PCA score plot based on proteins
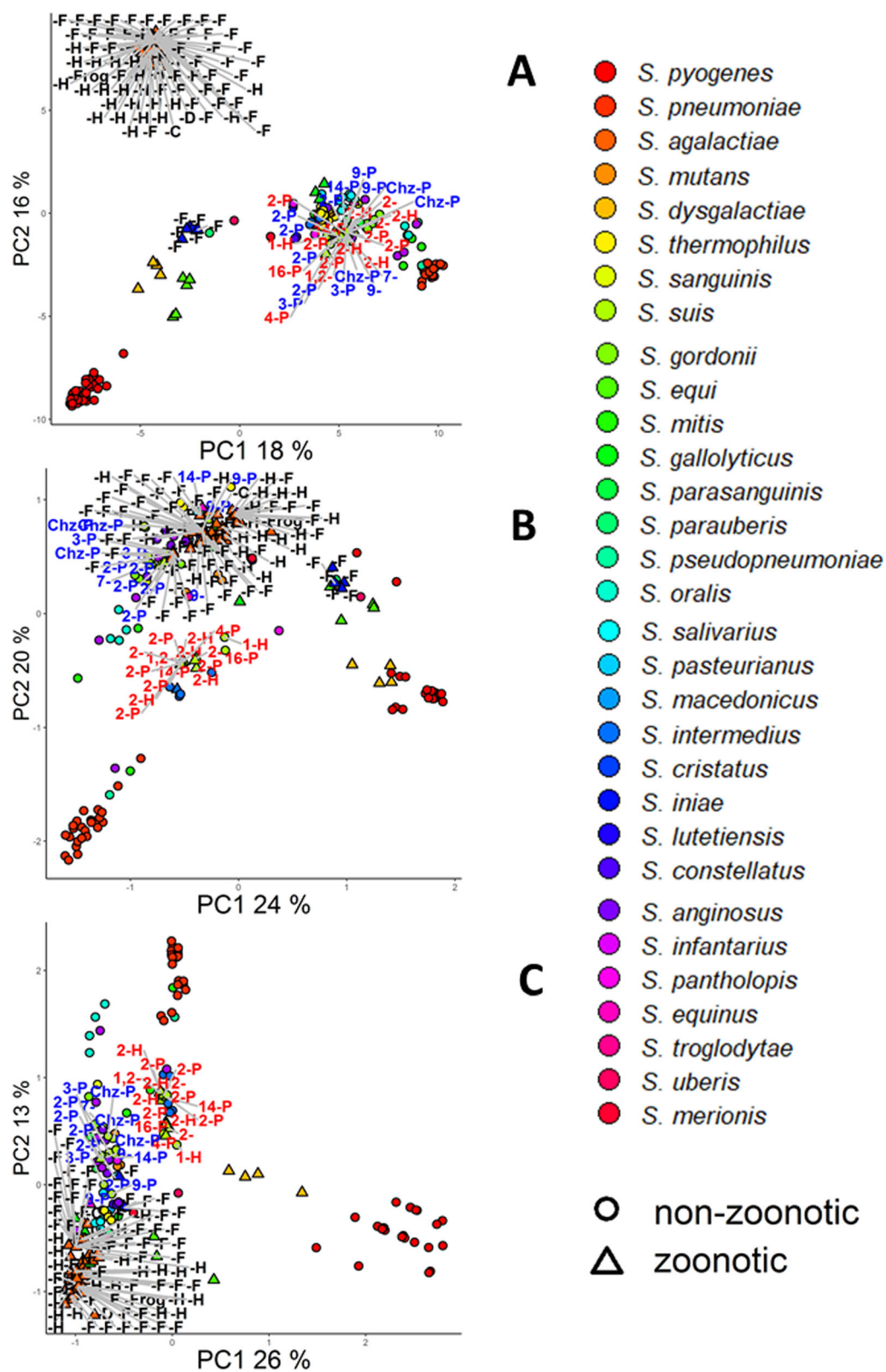
Fig. 6 (See legend on next page.)

Zondervan *et al. BMC Genomics*   (2021) 22:102

Page 9 of 21

(See figure on previous page.)

Fig. 6 PCA plot of *Streptococcus* strains based on all proteins (**a**), proteins filtered on 'modification of morphology or physiology of other organism' (**b**) and proteins filtered on 'pathogenesis' (**c**). *S. suis* serotypes are shown in the label, genomes from species mentioned in literature as having zoonotic capabilities are marked with a triangle and the isolation host is marked in the label with D = dog, F = fish, H = human, P = pig, T = toad. Genomes predicted in this study to have zoonotic potential are coloured red while strains in the cluster predicted not to have zoonotic potential are coloured blue. Fraction of variance explained by each PC is indicated in the axis

related to 'modification of morphology or physiology of other organism' (Fig. 6b) and the PCA based on 'pathogenesis' proteins (Fig. 6c): the first cluster contains 7 out of the 12 serotype 2 strains, as well as serotype 1, 1,2, 4, 16, while the second cluster contains 5 serotype 2 strains as well as strains with serotype 3, 7, 9 14 and Chz which were all isolated from pigs. The first group contains *S. suis* zoonotic strains of which some are isolated from pig and some from humans. The second group contains are non-zoonotic strains all isolated from pigs.

Streptococcus agalactiae zoonotic potential

Like *S. suis*, *S. agalactiae* forms two clusters when clustering on GO biological functional groups of proteins. Based on their isolation host, we can see that a cluster contains strains that are zoonotic while the other contains strains that are non-zoonotic. These two groups of *S. agalactiae* strains are better separated when using t-*t*-SNE plots based on all proteins and proteins involved in biological adhesion and pathogenesis (Fig. 7a-c) suggesting the existence of few proteins that are present in every genome in each group.

Identification of proteins that confer zoonotic potential

We used Random Forest, a machine learning approach, to investigate the association between genome content and phenotype using 75% of the data for training and 25% of the data for validation. Specifically, presence/absence of proteins filtered on association to GO Biological functions involved in pathogenesis to predict zoonotic potential *S. suis* and *S. agalactiae*, and we investigated which proteins are responsible for the zoonotic potential in these two species. We used functional groups of proteins that were shown to separate zoonotic and non-zoonotic strains for *S. suis* (Fig. 6b-c) and for *S. agalactiae* (Fig. 7b-c) to train a Random Forest classifiers. We investigated their overall importance for prediction as well as their contribution to predicting the class non-zoonotic, and the class zoonotic potential as shown in Fig. 8a-d. Where, the 'Impact' measure indicates the relevance of a protein of the prediction of given class. The 'importance' shows the proteins overall importance for the random forest classifier. Random Forest classifiers as well as the optimal hyper parameters can be found in (see Additional file 14).

The protein domain content of the five most important proteins for *S. suis* classification based on 'modification of morphology or physiology of other organism' proteins are: 1) PF01289 a thiol-activated cytolysin, 2) PF17440 thiol-activated cytolysin beta sandwich domain, 3) PF00910 replication initiation protein involved in viral RNA duplication 3) PF00078;PF08388; PF13655 group II intron reverse transcriptase/maturase, 4) PF03432 a relaxase involved in transfer of plasmids, 5) PF00665 Prokaryotic N-terminal methylation motif often found in pilins and other proteins involved in secretion (Fig. 8a).The most important proteins for *S. suis* classification based on 'pathogenesis' proteins are 1) PF01289 thiol-activated cytolysin, 2) PF17440 a thiol-activated cytolysin beta sandwich domain, 3) PF07564 hypothetical protein containing a domain of unknown function, 4) PF00092; PF00746 chemotaxin protein 5) PF00746;PF08363;PF16364 a glucan binding protein (Fig. 8b).

The *S. suis* classifiers based on 'modification of morphology or physiology of other organism' proteins as well the classifier based on 'pathogenesis' proteins, predict *S. suis* zoonotic potential with 100% accuracy solely based on the presence of either PF01289, a thiol-activated cytolysin or PF17440, a thiol-activated cytolysin beta sandwich Fig. 8a-b).

The most important features for *S. agalactiae* classification based on 'biological adhesion' proteins are 1) PF02872 a 5′-nucleotidase-C 2) PF00746;PF08017 Fibrinogen binding protein A, 3) PF00746; PF8428; PF08829; PF174802 surface protein Rib and 4) PF00746; PF16555; PF16569; PF16570 pilus complex 5) PF00746; PF11966 a cell wall anchored linked to a ySIRK signal domain (Fig. 8c). The most important features for *S. agalactiae* classification based on ´pathogenesis' proteins are 1) PF00746;PF08017 Fibrinogen binding protein, 2) PF00746;PF8428;PF08829;PF17480 2) surface protein Rib and 3) PF00746;PF16555;PF16569;PF16570 pilus complex 4) PF00746;PF11966 a cell wall anchored linked to a ySIRK signal domain 5) PF02225;PF07554, a serine protease (Fig. 8d).

For *S. agalactiae* classification, 5′-nucleotidase-C can predict training data with 100% accuracy and test data with 94% accuracy.

Virulence factors of necrotising fasciitis

We compared *Staphylococcus aureus* and *Streptococcus pyogenes* since both are a major cause of (monomicrobial) necrotizing fasciitis. Both *S. aureus* [23–25] and *S.*
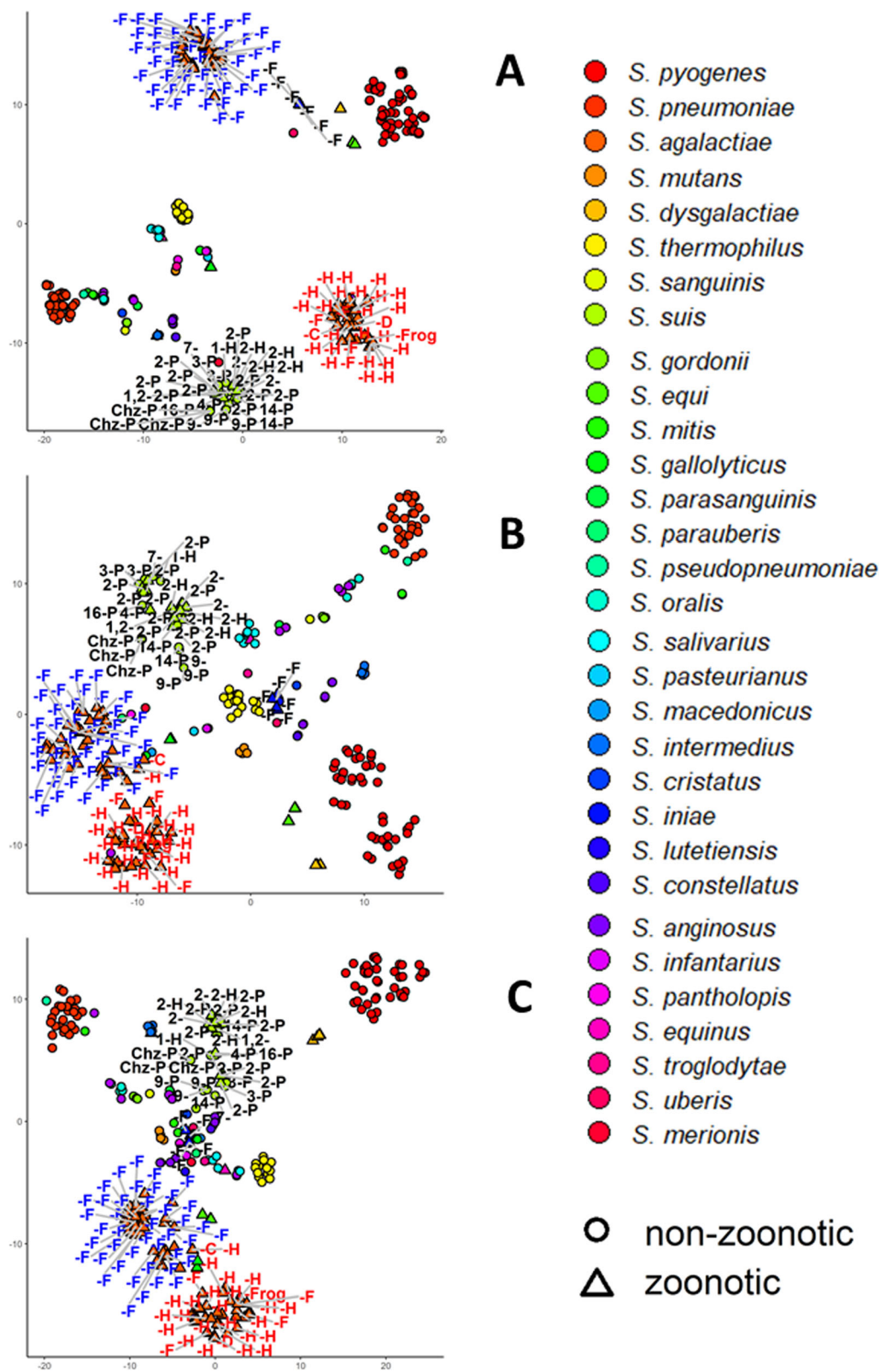
Fig. 7 (See legend on next page.)

(See figure on previous page.)

Fig. 7 t-SNE plots of Streptococcus strains based on proteins all proteins (**a**), proteins filtered on 'biological adhesion' (**b**) and 'pathogenesis' (**c**). t-SNE is a technique for dimensional reduction and visualization, so that similar objects appear as nearby objects in the two-dimensional plots here presented. *S. suis* serotypes are shown in the label, genomes from species mentioned in literature as having zoonotic capabilities are marked with a triangle and the isolation host is marked in the label with D = dog, F = fish, H = human, P = pig, T = toad. Genomes predicted in this study to be part of the zoonotic potential cluster are coloured red while strains in the cluster predicted not to have zoonotic potential are coloured blue. Fraction of variance explained by each PC is indicated in the axis

pyogenes [26, 27] fully lyse red blood cells, induce toxic shock syndrome as well as bind and invade epithelial host cells. Based on their GO term association *S. aureus* has 21 proteins associated to pathogenesis that occur in nearly all *S. aureus* genomes and only rarely in any other *Staphylococcus* specie (Table 3). Some of these pathogenesis proteins can directly be linked to pathogenesis proteins reported in literature for *S. aureus* [31] and for *S. pyogenes* [32]. An exact match with proteins reported in literature is however not always possible due to differences in annotation. When considering virulence factors that are not unique to *S. aureus* or *S. pyogenes* the number of virulence factors is about 1.5 times as many as reported in literature [33].

We looked at shared proteins as well as functional alternatives to find the molecular basis for necrotising fasciitis and we used the PFAM description of protein domains as well as description of proteins based on the locus tags associated to these proteins. We found that both *S. aureus* and *S. pyogenes* contain proteins involved in fibronectin binding, wound invasion, haemolysis, cell adhesion, IgA and IgG binding, multiple (super-)toxins as well as proteins involved in resisting phagocytosis and invading host cells (Table 3, Fig. 5). For example,
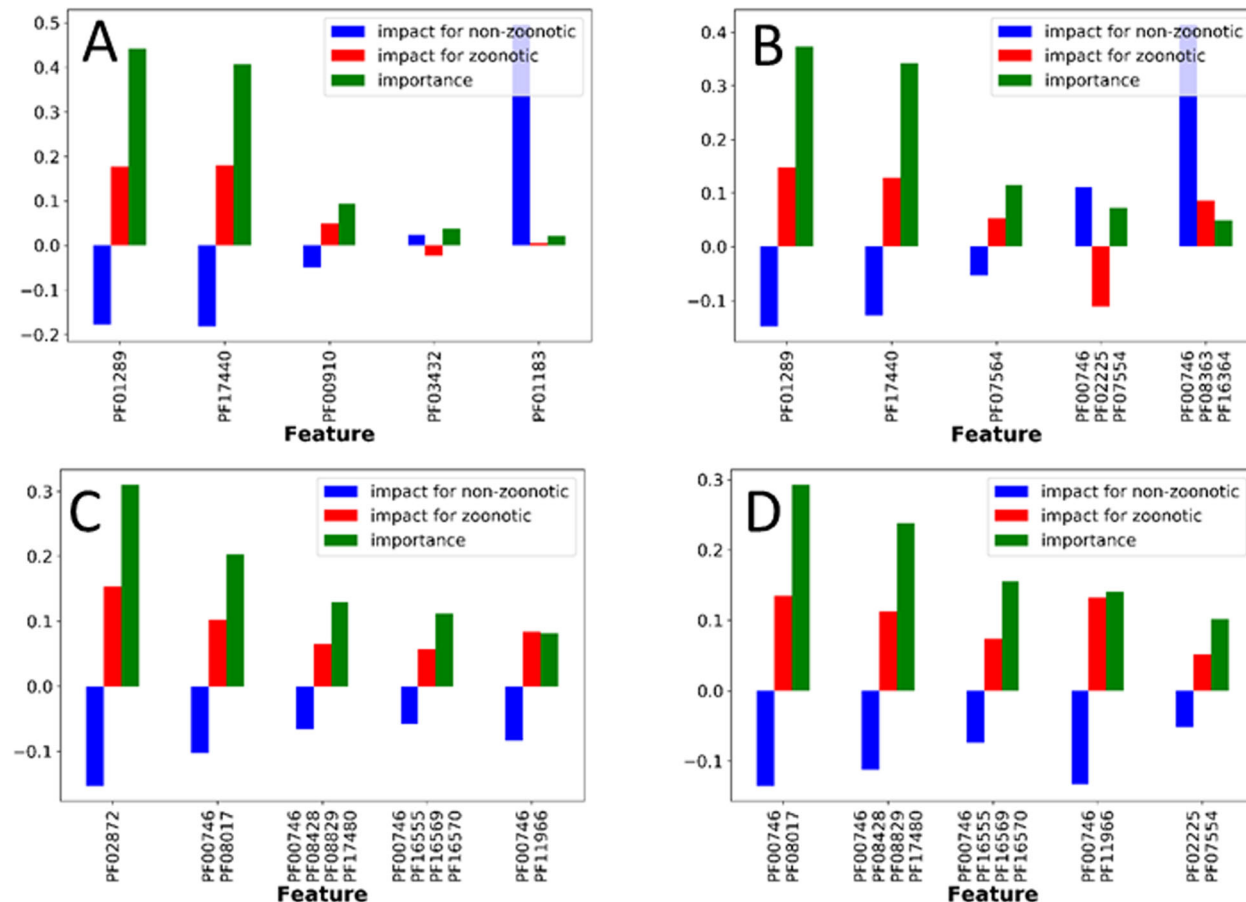

Fig. 8 Protein feature contribution to predict the class 'non-zoonotic' and 'zoonotic' as well as the overall importance of the protein feature for classification. **a** The five most important 'modification of morphology or physiology of other organism' proteins used to classify *S. suis*. **b** The five most important 'pathogenesis' proteins used to classify *S. suis*. **c** The five most important 'biological adhesion' proteins used to classify *S. agalactiae*. **d** The five most important 'pathogenesis' proteins used to classify *S. agalactiae*

Table 3 Proteins associated to *S. aureus* Pathogenesis (GO: GO:0009405). Domains that are shared between *S. aureus* and *S. pyogenes* are underlined. Proteins that are shared have are written in bold

Protein	Description
PF00746;PF07501; PF17041	LPXTG cell wall anchor; G5 domain, suggested adhesion, in peptide that cleaves IgA; E domain, rod like structure
PF00746; PF17210	LPXTG cell wall anchor, SdrD B-like domain, involved in adhesion to nose squamous cells [28]
PF13545	Crp-like helix-turn-helix domain, possibly cAMP interaction
PF05543	Staphopain peptidase C47, secreted cysteine protease
PF14731	Staphopain proregion
PF03373	Octapeptide repeat, part of SpA virulence factor frequently used to type *S. aureus* strains [29]
PF07968	Haemolysin, part of the Leukocidin/Hemolysin toxin family
PF09199	Staphylococcal superantigen-like OB-fold domain, interact with IgA, inhibits the end stage of complement activation and IgA binding to Fc-α-R [30]
PF02216	SpAB protein domain, immunoglobin binding domain
PF02876	Staphylococcal/Streptococcal toxin, beta-grasp domain
PF11621	C3 binding domain 4 of IgG-bind protein SB
PF01123	Enterotoxin type B, supertoxin, involved in food poisoning, causing the immune system to release a large number of cytokines that lead to significant inflammation
PF03642	MAP domain, major histocompatibility complex class II analog
PF00746;PF01476	LPXTG cell wall anchor motif; LysM domain found in many receptors, peptidoglycan-binding protein [12]
PF00746;PF02986	LPXTG cell wall anchor motif; Fibronectin binding repeat, enables uptake by host cell
PF00746;PF05031	LPXTG cell wall anchor motif; Iron Transport-associated domain, heme and/or hemoprotein-binding
PF07564	Domain of Unknown Function (DUF1542), several proteins containing this domain are involved in antibiotic resistance and/or cell adhesion
PF09023	Staphostatin B inhibits the cysteine protease Staphopain B
PF02821	Streptokinase (SK) is a thrombolytic medication and enzyme, breaks down blood cloths
PF01468	GA module, GA modules may promote bacterial growth and virulence in mammalian hosts by scavenging albumin-bound nutrients and camouflaging the bacteria
PF07554;PF07564;PF08428	FIVAR domain, likely binds fibronectin or more specifically N-acetyl glucosamine, occurs in proteins involved in methicillin resistance; Domain of Unknown Function (DUF1542); Rib/alpha-like repeat. Occurs in some Rib, a thought to confer protective immunity. Occurs in some *Streptococcus* surface proteins. Extracellular matrix-binding protein Ebh

PF01123 and PF02876 toxin β-grasp domain together form Enterotoxin type 2 which is important for causing the toxic shock [4, 34]. Enterotoxin type 2 antibodies are currently in clinical trials tested and have shown potential in treating necrotizing fasciitis [35].

S. pyogenes has 10 proteins that present in all *S. pyogenes* species and only occur separately in a few other *Streptococcus* strains (Table 4, Fig. 5). Five of these proteins are associated to the ability of *S. pyogenes* to bind to and break down fibrin in blood cloths [27]. Other proteins include toxin and enterotoxin, involved in over-activation of the immune response [38, 39], fibrin binding proteins, involved in adhesion and intracellular access of host cells, as well as proteases, involved in resistance to phagocytosis [40]. *S. pyogenes* has fewer proteins that are unique to this species compared to *S. aureus* since many other *Streptococcus* strains produce some of the pathogenic proteins present in *S. pyogenes* [41].

Only three proteins associated to pathogenesis are shared between *S. aureus* and *S. pyogenes,* Enterotoxin B

C-terminal domain (PF02876), Enterotoxin B N-terminal beta-grasp domain (PF01123) and Fibronectin binding protein (PF00746;PF02986). Both fibronectin binding protein A (FnbpA) and B (FnbpB) are expressed during infection conditions and were shown to be complexly regulated by a large number of regulators such as sigma factors and two component systems by *Mader* et al. [42]. Among these proteins identified in our study are potential biomarkers. *FnbpA* was found not to be essential in KO studies [43], but was found to be essential in a rapid shotgun antisense RNA method to identify essential genes in *S. aureus* [44].. No essentiality information is available for FnbpB. *S. aureus* fibronectin binding protein A (FnbpA) is called fibronectin binding protein X (SfbX) in *S. pyogenes*. For *S. aureus,* FnbpA was found to be essential for entry in the host cells [45]. FnbpA has functional homologs in other species such as *S. epidermidis*, however all homologs lack the C-terminal multiple fibronectin binding repeats variants present in FnbpA, of which at least one high affinity binding repeat

Table 4 Proteins associated to *S. pyogenes* Pathogenesis (GO: GO:0009405). Domains that are shared between *S. aureus* and *S. pyogenes* are underlined. Proteins that are shared are written in bold

Protein	Description
PF02821	Streptokinase, breaks down blood cloths
PF01640	Peptidase C10 family
PF01123	Enterotoxin type B, super antigen involved in food poisoning
PF02876	Staphylococcal/Streptococcal toxin, beta-grasp domain
PF03734	L,D-transpeptidase catalytic domain, peptidoglycan binding
PF00746; PF01391	LPXTG cell wall anchor motif; Collagen helix, rod like structure, coagulation-fibrinolytic binding in blood, Scl1 adhesin specifically recognizes the wound microenvironment [36]
PF00746; PF02370	LPXTG cell wall anchor motif; M protein repeat, binds IgA, major virulence factor involved in host cell invasion and resistance to phagocytosis [37]
PF00746; PF08017	LPXTG cell wall anchor motif; Fibronogen binding protein, members of this family include the fibrinogen receptor, FbsA which mediates platelet aggregation
PF00746; PF02986	LPXTG cell wall anchor motif; Fibronectin binding repeat, mediate adherence to host cells, enable the colonisation of wound tissue and blood clots
PF00092; PF00746; PF02986	Von Willebrand factor type A domain domains participate in numerous biological events (e.g. cell adhesion, migration, homing, pattern formation, and signal transduction); LPXTG cell wall anchor motif; Fibronectin binding repeat

is needed for host cell uptake [16, 34, 46]. A SfbX knockout mutants was shown to be only minimally affect *S. pyogenes* ability to infect epithelial host cells [47]. Enterotoxin type B, is a super toxin involved in over-activation of the immune response and interferes with phagocytosis by supressing the generation of myeloid-derived suppressor cells [48–50].

Discussion

The *Staphylococcus* and *Streptococcus* genera were compared on their genomic properties. Both genera have a similar ratio of their pan and core genome size. It should be considered that this analysis has been done with all fully assembled genomes data that were available at the time of the study. In our study we do not separate between pathogenic and non-pathogenic species since there are several ways to infect humans and animals. Instead, we use the underlying annotation of proteins marked as being involved in GO functions associated to pathogenesis to investigate patterns in pathogenesis. The choice not to define before-hand if species are pathogenic is deliberate since we recognize there are many forms of pathogenesis which depend on both species as well as the infection site as we discuss in the sections "Streptococcus suis pathogenesis zoonotic potential" and in the section "Streptococcus agalactiae zoonotic potential". We do however recognize the selected population affect our results as can be seen for in the ratio of their pan and core genome size found in this study. The alpha value of 1.12 found in our study for *Streptococcus* is higher than the 0.87 values reported by Koehorst et al. [12]. This difference can be explained by the number of genomes analysed which was 314 in our study opposed

to 60 in the study by Koehorst et al. Additionally, we allowed a maximum of one genome per species to be selected in our sampling approach to avoid population bias introduced by species with many sequenced genomes such as *S. aureus* which was not the case in the analysis of by Koehorst et al. [12].

Similar to what was found for *Pseudomonas* [9], gene expression variability of essential genes was found to be less than the expression variability of non-essential genes in both *S. aureus* and *S. pyogenes*.

Combination of experimentally determined essentiality and GEM based essentiality prediction were shown to be associated to a higher protein persistence than each of them individually. These results are to be expected since In vitro essentiality measurements are often only available for one condition, while GEM can easily be used to predict essentiality over multiple media conditions. Our GEM analysis predicted 153 *Staphylococcus aureus* genes to be essential in 90% of the minimal medium combination tested, while 163 genes were found to be essential for growth on rich medium and minimal medium. For *Streptococcus*, 196 genes were found to be essential in 90% of the minimal medium combinations tested, while no genes were found to be essential on rich medium. The *Streptococcus* model contains exchange reactions for all nutrients necessary for growth, meaning only the biomass reaction was found to be essential. Since we know from experimental results that there are several essential genes in *S. pyogenes*, we chose our method of testing all minimal medium compounds to best balance false positive and false negative results while keeping a unified method for our GEM essentiality analysis in both *S. aureus* and *S. pyogenes*. Similar to what has been

experimentally observed and was shown by previous published GEM simulations [8], our simulations show that *Staphylococcus* can use amino acids as alternative carbon source for survival in the host [51]. A possible limitation of this approach is that GEM predictions can only be made for metabolic (and their associated) proteins. Although many highly persistent genes tend to be essential, not all are highly persistent. This indicated alternatives in essentiality exist [52]. Similarly, many non-essential genes do have a high persistence, indicating they might be essential for *Staphylococcus* or *Streptococcus* specific functions such as survival and growth in non-lab conditions such as those found in the host.

Differences in pathogenesis, essentiality as well as other properties such as drug resistance, arise from different selection pressures for individual species within genera [53]. For example, similar to what was found in this study, a recent study shows that although streptococcal virulence factors have no clear patterns among species groups, some virulence factors were shown to be congruous with the evolution of species groups [13]. Core genes together with accessory genes form a complex network that comprise the molecular basis of virulence in *Staphylococcus* and *Streptococcus* [54, 55]. Within some individual species, strong selective pressure exist as was shown for *S. aureus* MRSA resistant species where there is an interplay of two strong evolutionary selective pressures: 1) the host type and 2) the antibiotics used in treatment which varies between humans, pets and livestock [56, 57].

We compared the similarity and differences between *Staphylococcus* and *Streptococcus* based on the clustering of species in GO functional trees. Some notable differences where observed between *Staphylococcus* and *Streptoccus*. Many 'modification of host morphology' proteins in *Staphylococcus* are also associated to the GO term 'pathogenesis' while many 'biological adhesion' proteins in *Streptococcus* are associated to the GO term 'pathogenesis'. These results could indicate that modification of host morphology is important for the pathology of *Staphylococcus* strains while biological adhesion is more important for the pathology of *Streptococcus*.

Next, we looked at which pathogenesis proteins separate *Staphylococcus* from *Streptococcus* species. Analysis of the presence and absence of proteins associated to 'pathogenesis' reveals that *Staphylococcus sciuri* GCA:002072755 and *Staphylococcus haemolytic us* GCA:001611955 only contain one pathogenesis protein (PF04647) that is not present in any *Streptococcus* strain. These results could indicate that horizontal gene transfer of pathogenic proteins occurred between *Staphylococcus* and *Streptococcus* or that they only carry pathogenesis proteins derived from a common ancestor.

Additionally, some pathogenic proteins only occur in one or a few genomes, indicating horizontal gene transfer from species outside the *Staphylococcus* and *Streptococcus* genus. Horizontal gene transfer is known to be a driving factor in the development of pathogenesis in *Staphylococcus* and *Streptococcus* [2, 6, 58–60]. For example, fibronectin binding domain PF02986 has been acquired by *Staphylococcus* and *Streptococcus* from an animal host, further spread among different *Streptococci* and *Staphylococci* through horizontal gene transfer, and further evolved through domain shuffling [41, 61].

Clustering of *Streptococcus* and *Staphylococcus* species based on different GO functional groups revealed sub cluster to be present for *S. suis* based on GO functional groups 'modification of host morphorlogy' and 'pathogenesis', and revealed a sub cluster to be present for *S. agalactiae* based on GO functional groups 'biological adhesion' and 'pathogenesis'. What is more, these sub cluster coincides with the potential to infect multiple hosts. It is known that predominantly *S. suis* serotype 2 strains are associated to zoonotic potential [62, 63]. However, as we could see in Fig. 6a-c, serotype information is not able to separate zoonotic and non-zoonotic *S. suis* strains.

Based on their isolation host we can see that the first group are *S. suis* strains are zoonotic, while the second group are non-zoonotic strains. Furthermore, human infections with strains for all serotypes in the first cluster have been reported [64, 65]: these results show that these functional groups of proteins can be used to predict the *S. suis* zoonotic potential. Interestingly, the zoonotic group of *S. suis* strains clusters with the human and dog oral commensal *S. intermedius* which can cause meningitis through brain abscesses as well as liver abscesses and in some rare cases endocarditis [66]. Since *S. suis* and *S. intermedius* are distantly related, this clustering is specific for proteins with functions in modification of host morphology and pathogenesis.

The similarity of phenotypes such as causing meningitis and tropism for brain and liver, suggests these traits may be caused by 'modification of morphology or physiology of other organism' and 'pathogenesis' proteins, and suggest a causal relationship between the proteins associated to these GO terms and the observed phenotype.

Investigation of proteins required to predict *S. suis* zoonotic potential using a random forest classifier revealed PF01289 thiol-activated cytolysin, 2) PF17440 a thiol-activated cytolysin beta sandwich domain to be the two most important factors associated to zoonotic potential. In support of these findings, it was found that a *S. suis* cytolysin knockout mutant made the strain non-haemolytic and non-cytotoxic for cultured macrophage-like cells [67] while increased secretion of thiol activated cytolysins was shown to directly cause epithelial cell

Zondervan *et al. BMC Genomics*   (2021) 22:102

Page 15 of 21

damage in humans, allowing *S. suis* to spread into deeper tissues [68]. Based on these studies it appears these two cytolysins are involved in formation of a pore-forming complex in cholesterol containing host membranes, which explains their importance for conferring zoonotic potential.

It has been suggested that *S. agalactiae* may have jumped from animals to humans in a certain moment of the evolution although it is still debatable if this zoonotic potential remains nowadays [69]. Here, we show that based on their genomic content *S. agalactiae* can be separated in two groups, one that is zoonotic and infects humans, fish and dog, and one group that only infects fish. This separation can be made based on all proteins, indicating zoonotic and non-zoonotic species are likely to have separated some time ago. In the *t*-SNE plots of biological adhesion and pathogenesis a third group can be seen of strains that infects mainly fish but also cow and human. This third cluster contains *S. agalactiae* strains that infect human most likely originate from this cluster and have further adapted to their human host by acquisition of proteins involved in biological adhesion and pathogenesis. The strains in this group appear to retain zoonotic potential since the cluster contains isolates from humans, fish and dog.

Investigation of proteins required to predict *S. agalactiae* zoonotic potential using a random forest classifier revealed multiple proteins to be important for classifying species as zoonotic. *S. agalactiae* 5′-nucleotidase-C is present in two proteins: Trifunctional nucleotide phospho-esterase protein YfkN precursor and Endonuclease YhcR precursor. Secreted nucleases play a role in evasion of the human innate immune response via destruction of extracellular traps and interference with phagocytosis signals [70]. Fibrinogen binding protein A allows *S. agalactiae* to attach to fibrinogen and to aggregate platelets [71]. Rib protein contains a Rib domain that confers protective immunity and an alpha C and alpha N protein domains involved in invasion and translocation along human epithelial cells according to their PFAM description. The pilus complex contains Pillin D1, Pillin B and Pillin D3 domains and contributes to the initial attachment and invasion of lung and cervical epithelial cells [71]. PF02225;PF07554 CspA serine protease breaks down three chemokines that attract and activate neutrophils [72].

In summary, all proteins important for classification of zoonotic potential appear to be causal to the zoonotic potential phenotype. Of these proteins, nucleosidases YfkN, YhcR and fibrinogen binding protein appear to be the most important factors for *S. agalactiae* zoonotic potential.

Conclusions
In this study we dissected *Staphylococcus* and *Streptococcus* pathogenesis through the systematic and integrated analysis of genomic, functional, metabolic, and expression data. Both genera were found to have a closed pan-genome and lower expression variation for essential and highly persistent genes than for non-essential and low persistent genes.

The study of functional groups of proteins in the pangenome of *Staphylococcus* and *Streptococcus* involved in pathogenesis, indicates that domain shuffling and horizontal gene transfer have played an important role in the development and acquisition of pathogenesis proteins of *Staphylococcus* and *Streptococcus* species.

The analysis of bacterial clusters based on functional groups of proteins involved in pathogenesis shows that clustering of strains correlates with phenotypes such as zoonotic potential.

Comparison between *S. aureus* and *S. pyogenes* indicate three proteins, Enterotoxin B C-terminal domain, Enterotoxin B N-terminal beta-grasp domain together with several functionally equivalent proteins allow *Staphylococcus aureus* and *Streptococcus pyogenes* to cause necrotizing fasciitis.

We have also shown that prediction of the phenotype zoonotic potential only requires information about a few proteins, suggesting a direct causal relationship with zoonotic potential.

These findings will enable further research in each of the areas addressed, whereas the approaches and methods herein deployed provide a solid basis towards large-scale prediction of phenotypes based on genomic information.

Methods
Genome retrieval and annotation
All available completely assembled genomes of 235 Staphylococcus and 315 Streptococcus were downloaded as EMBL files from EBI-ENA using the Python EnaBrowserTool [73]. Lists of these genomes accession number, name and taxon ID can be found in supplementary material [Additional file 1 and 2]. Genome EMBL files were converted to RDF and de novo annotation was performed storing the results in a graph file per genome using SAPP, a Semantic Annotation Platform with Provenance [74] and the GBOL ontology [75]. Gene calling was performed using Prodigal with codon Table 5 [76]. Annotation was performed using InterProScan version 5.25 [77]. Protein domains were identified by InterProScan by their Pfam identifier [78]. The GNU "parallel" package was used to perform all of the above steps in parallel [79].

The graph files were loaded in GraphDB Free version 8.4.1 in order to query the annotated genomes. Additionally, taxonomic information from UniProt was downloaded in RDF format and loaded in GraphDB. The GraphDB SPARQL endpoint was queried using the

Table 5 Gene Ontology (GO) terms used to select proteins based on their domain content for functional trees, PCA and t-SNA analysis. GO terms that are direct children of the 'Biological process' GO term are marked with an asterisk (*)

GO ID	Description
GO:0008150	Biological process
GO:0008152	*Metabolic process
GO:0017144	Drug metabolic process
GO:0042493	Response to drug
GO:0023052	*Signalling
GO:0065007	*Biological regulation
GO:0022610	*Biological adhesion
GO:0044406	Adhesion of symbiont to host
GO:0051704	Multi-organism process
GO:0044419	Inter species interaction between organisms
GO:0042710	Biofilm formation
GO:0098743	Cell aggregation
GO:0044403	Symbiont process
GO:0009372	Quorum sensing
GO:0035821	Modification of morphology or physiology of other organism
GO:0009405	Pathogenesis

Python SPARQLWrapper [80] package and the R Curl package [81] to retrieve information and store them as matrixes given in supplementary material [see Additional file 3 and 4]. These files were used for all subsequent analyses.

Estimation of the size of the pan- and core genome
Proteins were compared based on their Pfam domain content. We defined protein domain content as the alphabetical order of all unique domains associated with a given protein. A matrix was built to collect information on the presence or absence of proteins in each genome. Two sampling approaches were used: 1) genomes were randomly selected from all genomes in the analysed genera and 2) a maximum of one genome per species was selected to avoid bias introduced by species with many sequenced genomes. One up to the total number of genomes were sampled and analysed using the micropan R package [24] to investigate the effect of the number of genomes on the estimation of the size of the pan- and core genome. Additionally, these samples we used to estimate the sizes of the pan- and core genome using a binomial mixture model using the micropan BinomixEstimate function with 5000 permutations and a core detect probability of 1. The process was repeated 10 times to estimate the variance of the estimated size of the pan- and core genomes. The Heaps' function was used to fit a

Heaps' regression model; $\alpha > 1$ indicates convergence of the size of the pan-genome and that it is closed.

Variability of gene expression and its association to persistence
Gene variability was calculated based on 156 *S. aureus* RNA samples from 44 conditions ranging from laboratory to conditions mimicking infection, measured by Tiling arrays [42]. These 44 conditions can be categorized in four groups:

1) rich medium (TSB), 2) minimal medium (CDM), 3) cell culture media (RPMI, pMEM) 4) in human plasma (plasma), 5) growth with human bronchial epithelial cell line S9 and the human monocyte cell line THP-1.

Samples were taken at different time points and for infection simulations oxygen availability was limited at later time points. For a complete description of the conditions we refer to S1 Data in the original paper by U. Mader et al. [42]. For every gene we considered its expression profile over all samples and a variability value was calculated as the ratio between the standard deviation and the mean expression value using the same approach as in Koehorst et al. [9].

Protein persistence and essentiality
We defined the persistence of a gene as

$$Persistence = \frac{N\,(orth)}{N}$$

where $N\,(orth)$ is the number of genomes carrying a given orthologue and N is the number of genomes searched [82]. Orthologue genes were identified as genes with identical protein domain content. Locus tags associated to the genes were inferred from the original annotation and used to integrated genome wide gene essentiality data from transposon mutagenesis studies for *Staphylococcus* strains S0385 grown on whole porcine blood [83], NCTC8325 Newman grown on BHI broth [43] and JE2 grown on Handke mannitol medium [84] and *Streptococcus* strains *S. pyogenes* M1T1 strain 5448 and M49 strain NZ131 grown in rich Todd-Hewitt Yeast (THY) medium [85].

GEM-based predictions of essentiality
Gene essentiality analysis based on genome scale modelling was performed using the genome-scale, constraint-based metabolic model (GEM) of *S. aureus* NTCTC 8325 [8] and the GEM model of *S. pyogenes* M49 [86]. First, a minimal medium was determined using the 'cobrapy minimal_medium function'. All carbon, nitrogen, sulphur and phosphor sources from the medium that could support growth were detected by substituting the default carbon, nitrogen, sulphur and phosphor

sources. All combinations of minimal media containing these carbon, nitrogen, sulphur and phosphor sources were generated.

Gene essentiality for all combinations of minimal media containing these carbon, nitrogen, sulphur and phosphor were tested by performing single gene deletions followed by flux balance analysis optimizing for growth. If a gene knock-out reduced predicted growth for the media compositions below 1% the gene was considered conditionally essential. Genes predicted to be conditionally essential in at least 90% of the in-silico media compositions were marked as essential.

All optimizations were performed using the Gurobi optimizer 8.1 [87] with COBRApy 0.13.4 [88] and Python 3.6.

Functional analysis

Genome information was retrieved from associated literature and from the Biosample database [89], including serotype information and zoonotic potential and isolation-host. Zoonotic classification was derived from literature for *S. inae* [90–93], *S. agalactiae* [69], *S. dysgalactiae, S. equi* [94–96] and at the serotype level for *S. suis* [97–99]. For all zoonotic *Streptococcus* species data about the isolation host was retrieved from the Biosamples databases [100] or literature [63–65, 101–108]. Additional Gene Ontology (GO) annotation from the GODM (GO Domain Miner) database [33] was added to proteins based on their domain content, increasing the number of GO terms by approximately 10-fold compared to GO term annotation retrieved from the Inter-Pro database. Literature was used to select 17 GO terms in the Biological process ontology with known or suspected association to pathogenesis [109–113] (Table 5).

The presence/absence matrix of proteins was filtered on proteins annotated with any of the 17 GO terms (Table 5) or their descendent GO terms using the R GO.db package [114]. The filtered matrix was used to calculate the Euclidean distance between genomes. Hierarchical complete-linkage clustering was used to generate dendrograms. These GO-specific dendrograms were compared to a reference dendrogram based on all proteins.

Because these dendrograms are based on annotation of proteins for a specific function, we will refer to them as 'functional trees'. Euclidean distances of genomes in the functional trees and the reference tree were calculated and scaled to values between 0 and 1 using the R scale function using the minimum value for centring, and (min – max) for scaling. Scaled values were used to calculate similarity scores for the position of genomes in each functional tree compared to the reference. These similarity scores were calculated as the Pearson correlation between the scaled Euclidean distances of genomes

in the functional tree and the scaled Euclidean distance in the reference tree. Interactive heatmap were generated showing the presence and absence of proteins per genome, while showing the similarity in the side column to highlight differences compared to the reference tree. These interactive graphs were generated using the dendextend and heatmaply packages [115, 116]. Similarity scores for functional trees were calculated using the dendextend cor_cophenetic function.

Matrix manipulations, Principal Component Analysis (PCA), t-Distributed Stochastic Neighbour Embedding (t-SNE) and graphs were performed using R 3.6.1 [117], the prcomp command, and the Rtsne 0.15 [118] and ggplot2 2.3.2.1 [119] packages. t-SNE was performed with default parameters.

Random Forest classification

Proteins belonging to GO categories 'pathogenesis', 'modification of morphology or physiology of other organisms' and 'biological adhesions', were used to train Random Forests classifiers for *S. suis* and *S. agalactiae* strains to predict whether they belong to the class 'zoonotic potential' on 'non-zoonotic potential'. This classification was based on the clustering of in PCA and t-SNE plots which revealed the presence of a zoonotic and a non-zoonotic group of strains. Data was split in 75% training data and 25% validation data.

Data was loaded using Python 3.6, pandas 0.24.2. Skicit-learn 0.20.3 used to load data and train Random Forest classifiers. Treeinterpreter 0.1.0 was used to interpret feature (protein) importance for classification in general and feature contribution to predict specific classes. Grid search for 300 combinations of parameters was performed optimizing the parameters n_estimators, max_features, max_depth, min_samples_split, and the min_samples. Iterative rounds of feature reduction, that is removal of the protein which least contribute to the classification, followed by hyper parameter optimization, was used to find the minimal set of features (proteins) needed to classify both training and test data. Feature importance and contribution were plotted using matplotlib 3.0.3.

Supplementary Information

Additional file 1. List of Staphylococcus genomes

Additional file 2. List of Streptococcus genomes

Additional file 3. Staphylococcus All proteins

Additional file 4. Streptococcus All proteins

Additional file 5. Estimated Pan and Core genome size, Heaps Analysis

Additional file 6. Staphylococcus Functional trees

Additional file 7. Staphylococcus PCA

Abbreviations
Exp: Experimentally; GEM: Genome scale metabolic model; GO: Genome Ontology; IgA: Immunoglobulin A; IgG: Immunoglobulin G; t-SNE: T-distributed Stochastic Neighbour Embedding

Authors' contributions
NZ performed the main analyses and wrote the draft manuscript. MSD and ES participated in the design of the study and supervised and directed the research. MSD and ES and VDMS revised the manuscript. All authors contributed to the writing of the final version of the manuscript. The authors read and approved the final manuscript.

Availability of data and materials
The authors declare that all data supporting the findings of this study are available within the article and its supplementary information.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Stippeneng 4, 6708WE Wageningen, Netherlands. [2]LifeGlimmer GmBH, Markelstraße 38, 12163, Berlin, Germany.

References
1. Graham MR, Smoot LM, Migliaccio CAL, Virtaneva K, Sturdevant DE, Porcella SF, et al. Virulence control in group a streptococcus by a two-component gene regulatory system: global expression profiling and in vivo infection modeling. Proc Natl Acad Sci U S A. 2002;99(21):13855–60 Available from: http://www.pnas.org/cgi/content/long/99/21/13855.

2. Krismer B, Weidenmaier C, Zipperer A, Peschel A. The commensal lifestyle of Staphylococcus aureus and its interactions with the nasal microbiota. Nat Rev Microbiol. 2017;15(11):675–87. https://doi.org/10.1038/nrmicro.2017.104.

3. Henriques-Normark B, Normark S. Commensal pathogens, with a focus on Streptococcus pneumoniae, and interactions with the human host. Exp Cell Res. 2010;316(8):1408–14. https://doi.org/10.1016/j.yexcr.2010.03.003.

4. Causey WA. Staphylococcal and streptococcal infections of the skin. Prim Care. 1979;6(1):127–39 Available from: http://www.ncbi.nlm.nih.gov/pubmed/8902344.

5. Patenge N, Pappesch R, Khani A, Kreikemeyer B. Genome-wide analyses of small non-coding RNAs in streptococci. Front Genet. 2015;6(MAY):1–13.

6. Suzuki H, Lefébure T, Bitar PP, Stanhope MJ. Comparative genomic analysis of the genus Staphylococcus including Staphylococcus aureus and its newly described sister species Staphylococcus simiae. BMC Genomics. 2012;13(1):38 Available from: http://www.biomedcentral.com/1471-2164/13/38.

7. Saccenti E, Nieuwenhuijse D, Koehorst JJ, Dos Santos VAPM, Schaap PJ. Assessing the metabolic diversity of streptococcus from a protein domain point of view. PLoS One. 2015;10(9):1–20.

8. Bosi E, Monk JM, Aziz RK, Fondi M, Nizet V, Palsson BØ. Comparative genome-scale modelling of Staphylococcus aureus strains identifies strain-specific metabolic capabilities linked to pathogenicity. Proc Natl Acad Sci. 2016;113(26):E3801–9. https://doi.org/10.1073/pnas.1523199113.

9. Koehorst JJ, Van Dam JCJ, Van Heck RGA, Saccenti E, Dos Santos VAPM, Suarez-Diez M, et al. Comparison of 432 pseudomonas strains through integration of genomic, functional, metabolic and expression data. Sci Rep. 2016;6(December):1–13. https://doi.org/10.1038/srep38699.

10. Koehorst JJ, Saccenti E, Schaap PJ, dos Santos VAP M, Suarez-Diez M. Protein domain architectures provide a fast, efficient and scalable alternative to sequence-based methods for comparative functional genomics. F1000Research. 2017;5(0):1987 Available from: https://f1000research.com/articles/5-1987/v3.

11. Rouli L, Merhej V, Fournier PE, Raoult D. The bacterial pangenome as a new tool for analysing pathogenic bacteria. New Microbes New Infect. 2015;7:72–85. https://doi.org/10.1016/j.nmni.2015.06.005.

12. Fuchs S, Mehlan H, Bernhardt J, Hennig A, Michalik S, Surmann K, et al. AureoWiki-the repository of the Staphylococcus aureus research and annotation community. Int J Med Microbiol. 2018;308(6):558–68. https://doi.org/10.1016/j.ijmm.2017.11.011.

13. Gao XY, Zhi XY, Li HW, Klenk HP, Li WJ. Comparative genomics of the bacterial genus streptococcus illuminates evolutionary implications of species groups. PLoS One. 2014;9(6):e101229.

14. Snipen L, Almøy T, Ussery DW. Microbial comparative pan-genomics using binomial mixture models. BMC Genomics. 2009;10:1–8.

15. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol. 2008;11(5):472–7.

16. Protein F, Jeng A, Sakota V, Li Z, Datta V, Beall B, et al. Molecular genetic analysis of a group a streptococcus operon encoding serum opacity factor and a novel. J Bacteriol. 2003;185(4):1208–17 Available from: http://jb.asm.org/content/185/4/1208.full.pdf.

17. Kuroda M, Ohta T, Uchiyama I, Baba T, Yuzawa H, Kobayashi I, et al. Whole genome sequencing of meticillin-resistant Staphylococcus aureus. The Lancet 2001;357(9264):1225--40. https://doi.org/10.1016/S0140-6736(00)04403-2.

18. Matyi SA, Dupre JM, Johnson WL, Hoyt PR, White DG, Brody T, et al. Isolation and characterization of Staphylococcus aureus strains from a Paso del Norte dairy. J Diary Sci HHS Public Access. 2017;96(6):3535–42.

19. Staphylococcus TF. Chapter 12 in microbiology. In: S B, editor. Microbiology 4th edition. 4th ed. Galveston: University of Texas Medical Branch at Galveston; 1996. Chapter 12. Available from: https://www.ncbi.nlm.nih.gov/books/NBK8448/996.

20. Li Y, Cao B, Zhang Y, Zhou J, Yang B, Wang L. Complete genome sequence of Staphylococcus aureus T0131, an ST239-MRSA-SCCmec type III clone isolated in China. J Bacteriol. 2011;193(13):3411–2.

21. Sass P, Berscheid A, Jansen A, Oedenkoven M, Szekat C, Strittmatter A, et al. Genome sequence of Staphylococcus aureus VC40, a vancomycin- and daptomycin-resistant strain, to study the genetics of development of resistance to currently applied last-resort antibiotics. J Bacteriol. 2012;194(8):2107–8.

22. Baba T, Bae T, Schneewind O, Takeuchi F, Hiramatsu K. Genome sequence of Staphylococcus aureus strain Newman and comparative analysis of staphylococcal genomes: polymorphism and evolution of two major pathogenicity islands. J Bacteriol. 2008;190(1):300–10.

23. Zautner AE, Krause M, Stropahl G, Holtfreter S, Frickmann H, Maletzki C, et al. Intracellular persisting *Staphylococcus aureus* is the major pathogen in recurrent tonsillitis. PLoS One. 2010;5(3).

24. Surmann K, Depke M, Dhople VM, Pané-Farré J, Hildebrandt P, Gumz J, et al. Analysis of Staphylococcus aureus proteins secreted inside infected human epithelial cells. Int J Med Microbiol. 2018;308(6):664-74. https://doi.org/10.1016/j.ijmm.2018.06.002.

25. Sendi P, Proctor RA. Staphylococcus aureus as an intracellular pathogen: the role of small colony variants. Trends Microbiol. 2009;17(2):54–8.

26. Kreikemeyer B, McIver KS, Podbielski A. Virulence factor regulation and regulatory networks in streptococcus pyogenes and their impact on pathogen-host interactions. Trends Microbiol. 2003;11(5):224–32.

27. Loof TG, Deicke C, Medina E. The role of coagulation/fibrinolysis during streptococcus pyogenes infection. Front Cell Infect Microbiol. 2014;4(September):1–8. https://doi.org/10.3389/fcimb.2014.00128/abstract.

28. Askarian F, Ajayi C, Hanssen A, Van Sorge NM. The interaction between *Staphylococcus aureus* SdrD and desmoglein 1 is important for adhesion to host cells. Nat Publ Gr. 2016;(October 2015):1–11. https://doi.org/10.1038/srep22134.

29. Eriksen NHR, Espersen F, Rosdahl VT, Jensen K, Godwin H, Wyllie DH, et al. Prevalence of Staphylococcus aureus protein a (spa) mutants in the community and hospitals in Oxfordshire. Epidemiol Infect. 1995;115(01):51–60 Available from: http://www.journals.cambridge.org/abstract_S095026 8800058118.

30. Hermans SJ, Baker HM, Sequeira RP, Langley RJ, Baker EN, Fraser JD. Structural and functional properties of staphylococcal superantigen-like protein 4. Infect Immun. 2012;80(11):4004–13.

31. Jenkins A, Diep A, Mai TT, Vo NH, Warrener P, Suzich J, et al. Differential expression and roles of Staphylococcus aureus virulence determinants during colonization and disease. MBio. 2015;6(1):1–10.

32. Virulence factors of *Streptococcus pyogenes* and their roles [Internet]. [cited 2018 Dec 17]. Available from: https://microbeonline.com/virulence-factors-streptococcus-pyogenes-roles/.

33. Alborzi SZ, Devignes M, Ritchie D, Alborzi SZ, Devignes M, Ritchie D, et al. Associating gene ontology terms with Pfam protein domains to cite this version : HAL id : HAL-01531204 associating gene ontology terms with Pfam protein domains; 2017.

34. Stemberk V, Jones RPO, Moroz O, Atkin KE, Edwards AM, Turkenburg JP, et al. Evidence for steric regulation of fibrinogen binding to *Staphylococcus aureus* Fibronectin-binding. Protein A. 2014;289(18):12842–51.

35. Fries BC, Varshney AK. Bacterial toxins — staphylococcal enterotoxin B description of agent; 2013. p. 1–12.

36. Zheng X, Bi C, Brooks M, DSH. Collagen-like proteins of pathogenic streptococci. Anal Chem. 2015;25(4):368–79.

37. Rohde M, habi rer nat, Cleary PP. Adhesion and invasion of *Streptococcus pyogenes* into host cells and clinical relevance of intracellular streptococci. Univ Oklahoma Heal Sci Cent. 2016:1–30.

38. Gottlieb M, Long B, Koyfman A. The evaluation and management of toxic shock syndrome in the emergency department: a review of the literature. J Emerg Med. 2018;(December 2017):1–8. https://doi.org/10.1016/j.jemermed.2017.12.048.

39. Otto M. Staphylococcus aureus toxins Michael. Curr Opin Microbiol. 2015:32–7.

40. Kolar SL, Antonio Ibarra J, Rivera FE, Mootz JM, Davenport JE, Stevens SM, et al. Extracellular proteases are key mediators of Staphylococcus aureus virulence via the global modulation of virulence-determinant stability. Microbiologyopen. 2013;2(1):18–34.

41. Chen WH, Lu G, Chen X, Zhao XM, Bork P. OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. Nucleic Acids Res. 2017;45(D1):D940–4.

42. Mäder U, Nicolas P, Depke M, Pané-Farré J, Debarbouille M, van der Kooi-Pol MM, et al. Staphylococcus aureus Transcriptome architecture: from laboratory to infection-mimicking conditions. PLoS Genet. 2016;12(4):1–32.

43. Chaudhuri RR, Allen AG, Owen PJ, Shalom G, Stone K, Harrison M, et al. Comprehensive identification of essential Staphylococcus aureus genes using transposon-mediated differential hybridisation (TMDH). BMC Genomics. 2009;10:1–18.

44. Forsyth RA, Haselbeck RJ, Ohlsen KL, Yamamoto RT, Xu H, Trawick JD, et al. A genome-wide strategy for the identification of essential genes in Staphylococcus aureus. Mol Microbiol. 2002;43(6):1387–400.

45. Henderson B, Nair S, Pallas J, Williams MA. Fibronectin: a multidomain host adhesin targeted by bacterial fibronectin-binding proteins. FEMS Microbiol Rev. 2011;35(1):147–200.

46. Edwards AM, Potts JR, Josefsson E, Massey RC. *Staphylococcus aureus* host cell invasion and virulence in sepsis is facilitated by the multiple repeats within FnBPA. PLoS Pathog. 2010;6(6):e1000964. https://doi.org/10.1371/journal.ppat.1000964.

47. Timmer AM, Kristian SA, Datta V, Jeng A, Gillen CM, Walker MJ, et al. Serum opacity factor promotes group A streptococcal epithelial cell invasion and virulence. Mol Microbiol. 2006;62(1):15–25 Available from: http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L44386547%5Cn, https://doi.org/10.1111/j.1365-2958.2006.05337.x, http://mun-primo.hosted.exlibrisgroup.com/openurl/01MUN/01MUN_SERVICES?sid=EMBASE&issn=0950382X&id=doi:10.1111%2Fj.

48. Spaulding AR, Salgado-Pabón W, Kohler PL, Horswill AR, Leung DYM, Schlievert PM. Staphylococcal and streptococcal superantigen exotoxins. Clin Microbiol Rev. 2013;26(3):422–47.

49. Klingelhutz AJ, Salgado-pabón W. The Superantigen toxic shock syndrome toxin 1 alters human aortic endothelial cell function. Infect Immun. 2018;86(3):1–16. https://doi.org/10.1128/IAI.00848.17.

50. Stoll H, Ost M, Singh A, Mehling R, Neri D, Schäfer I, et al. Staphylococcal enterotoxins dose-dependently modulate the generation of myeloid-derived suppressor cells. Front Cell Infect Microbiol. 2018;8(September):1–15.

51. Bae T, Banger AK, Wallace A, Glass EM, Aslund F, Schneewind O, et al. Staphylococcus aureus virulence genes identified by bursa aurealis mutagenesis and nematode killing. Proc Natl Acad Sci. 2004;101(33):12312–7.

52. Kamminga T, Koehorst JJ, Vermeij P, Slagman S-J, Martins dos Santos VAP, Bijlsma JJE, et al. Persistence of functional protein domains in mycoplasma species and their role in host specificity and synthetic minimal life. Front Cell Infect Microbiol. 2017;7:31 [cited 2017 Feb 7]. https://doi.org/10.3389/fcimb.2017.00031/full.

53. Cooper VS, Honsa E, Rowe H, Deitrick C, Iverson AR, Whittall JJ, et al. Experimental evolution in vivo to identify selective pressures during pneumococcal colonization. bioRxiv. 2020;5(3):1–17.

54. Pain M, Hjerde E, Klingenberg C, Cavanagh JP. Comparative genomic analysis of Staphylococcus haemolyticus reveals key to hospital adaptation and pathogenicity. Front Microbiol. 2019;10(September):1–13.

55. Anisimova M, Bielawski J, Dunn K, Yang Z. Phylogenomic analysis of natural selection pressure in streptococcus genomes. BMC Evol Biol. 2007;7:1–13.

56. Jamrozy D, Coll F, Mather AE, Harris SR, Harrison EM, MacGowan A, et al. Evolution of mobile genetic element composition in an epidemic methicillin-resistant Staphylococcus aureus: temporal changes correlated with frequent loss and gain events. BMC Genomics. 2017;18(1):1–12.

57. Mccarthy AJ, Lindsay JA, Loeffler A. Are all meticillin-resistant *Staphylococcus aureus* (MRSA) equal in all hosts? Epidemiological and genetic comparison between animal and human MRSA. Vet Dermatol. 2012;23(4):267–75.

58. Osaki M, Takamatsu D, Shimoji Y, Sekizaki T. Characterization of *Streptococcus suis* genes encoding proteins homologous to sortase of gram-positive bacteria. J Bacteriol. 2002;184(4):971–82 Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=134807&tool=pmcentrez&rendertype=abstract.

59. Kao C, Chen P, Huang F, Chen C, Chi C, Lin Y, et al. Clinical and genetic analysis of invasive and non-invasive group A streptococcal infections in Central Taiwan; 2005. p. 105–11.

60. McNeilly CL, McMillan DJ. Horizontal gene transfer and recombination in *Streptococcus dysgalactiae subsp. equisimilis*. Front Microbiol. 2014;5(DEC):1–6.

61. Bork P, Doolittle RF. Proposed acquisition of an animal protein domain by bacteria. Proc Natl Acad Sci. 1992;89(19):8990–4. https://doi.org/10.1073/pnas.89.19.8990.

62. Goyette-Desjardins G, Auger JP, Xu J, Segura M, Gottschalk M. Streptococcus suis, an important pig pathogen and emerging zoonotic agent-an update on the worldwide distribution based on serotyping and sequence typing. Emerg Microbes Infect. 2014;3(March):e45.

63. Kerdsin A, Puangpatra P, Tanimura S, Gottschalk M, Polwichai P, Dejsirilert S, et al. Genotypic profile of Streptococcus suis serotype 2 and clinical features of infection in humans, Thailand. Emerg Infect Dis. 2011;17(5):835–42.

64. Nghia HDT, Ngo TH, Le DL, Campbell J, To SD, Chau NVV, et al. Human case of Streptococcus suis serotype 16 infection. Emerg Infect Dis. 2008;14(1):155–7.

Zondervan *et al. BMC Genomics*   (2021) 22:102

Page 20 of 21

65. Wertheim HFL, Nghia HDT, Taylor W, Schultsz C. Streptococcus suis: an emerging human pathogen. Clin Infect Dis. 2009;48(5):617–25.

66. Hasegawa N, Sekizuka T, Sugi Y, Kawakami N, Ogasawara Y, Kato K, et al. Characterization of the pathogenicity of Streptococcus intermedius TYG1620 isolated from a human brain abscess based on the complete genome sequence with Transcriptome analysis and transposon mutagenesis in a murine subcutaneous abscess model. Infect Immun. 2017;85(2):1–15.

67. Allen AG, Bolitho S, Lindsay H, Khan S, Bryant C, Norton P, et al. Generation and characterization of a defined mutant of streptococcus suis lacking suilysin. Infect Immun. 2001;69(4):2732–5.

68. He Z, Pian Y, Ren Z, Bi L, Yuan Y, Zheng Y, et al. Increased production of suilysin contributes to invasive infection of the Streptococcus suis strain 05ZYH33. Mol Med Rep. 2014;10(6):2819–26.

69. Botelho ACN, Ferreira AFM, Fracalanzza SEL, Teixeira LM, Pinto TCA. A perspective on the potential zoonotic role of Streptococcus agalactiae: searching for a missing link in alternative transmission routes. Front Microbiol. 2018;9(March):1–5.

70. Remmington A, Turner CE. The DNases of pathogenic lancefield streptococci. Microbiology. 2018;164(3):242–50.

71. Sharma P, Lata H, Arya DK, Kashyap AK, Kumar H, Dua M, et al. Role of pilus proteins in adherence and invasion of streptococcus agalactiae to the lung and cervical epithelial cells. J Biol Chem. 2013; 288(6):4023–34.

72. Bryan JD, Shelver DW. Streptococcus agalactiae CspA is a serine protease that inactivates chemokines. J Bacteriol. 2009;191(6):1847–54.

73. European Bioinformatics Institute. EnaBroswerTools [Internet]. [cited 2019 Nov 7]. Available from: https://github.com/enasequence/enaBrowserTools.

74. Koehorst JJ, van Dam JCJ, Saccenti E, dos Santos VAP M, Suarez-Diez M, Schaap PJ. SAPP: functional genome annotation and analysis through a semantic framework using FAIR principles. Bioinformatics. 2017;(March):1–3. https://doi.org/10.1093/bioinformatics/btx767/4653704.

75. van Dam JCJ, Koehorst JJ, Vik JO, VAP MDS, Schaap PJ, Suarez-Diez M. The Empusa code generator and its application to GBOL, an extendable ontology for genome annotation. Sci Data. 2019;6(1):254. https://doi.org/10.1038/s41597-019-0263-7.

76. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.

77. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, et al. InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res. 2017;45(D1):D190–9.

78. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44(D1):D279–85.

79. Tange O. Gnu parallel-the command-line power tool.; login: The USENIX Magazine, 2011;36(1):42–47.

80. SPARQL endpoint interface to python [Internet]. [cited 2018 Aug 8]. Available from: https://rdflib.github.io/sparqlwrapper/.

81. Duncan Temple Lang and the CRAN team. CRAN - Package RCurl [Internet]. [cited 2018 Nov 1]. Available from: https://cran.r-project.org/web/packages/RCurl/index.html

82. Fang G, Rocha E, Danchin A. How essential are nonessential genes ? 2004.

83. Christiansen MT, Kaas RS, Chaudhuri RR, Holmes MA, Hasman H, Aarestrup FM. Genome-wide high-throughput screening to investigate essential genes involved in methicillin-resistant *Staphylococcus aureus* sequence type 398 survival. PLoS One. 2014;9(2):e89018.

84. Fey PD, Endres JL, Yajjala VK, Fey PD, Endres JL, Yajjala K, et al. A genetic resource for rapid and comprehensive phenotype. MBio. 2013;4(1):1–8.

85. Le Breton Y, Belew AT, Valdes KM, Islam E, Curry P, Tettelin H, et al. Essential genes in the core genome of the human pathogen streptococcus pyogenes. Sci Rep. 2015;5:9838.

86. Levering J, Fiedler T, Sieg A, van Grinsven KWA, Hering S, Veith N, et al. Genome-scale reconstruction of the streptococcus pyogenes M49 metabolic network reveals growth requirements and indicates potential drug targets. J Biotechnol. 2016;232:25–37. https://doi.org/10.1016/j.jbiotec.2016.01.035.

87. Gurobi Optimization LLC. Gurobi optimizer reference manual [internet]; 2018. Available from: http://www.gurobi.com.

88. Ebrahim A, JALBO P, Hyduke DR. COBRApy: constraints-based reconstruction and analysis for python. BMC Syst Biol. 2013;7:74. https://doi.org/10.1186/1752-0509-7-74.

89. Gostev M, Faulconbridge A, Brandizi M, Fernandez-Banet J, Sarkans U, Brazma A, et al. The BioSample database (BioSD) at the European bioinformatics institute. Nucleic Acids Res. 2012;40(D1):64–70.

90. Zhang BC, Zhang J, Sun L. Streptococcus iniae SF1: complete genome sequence, proteomic profile, and immunoprotective antigens. PLoS One. 2014;9(3):e91324.

91. Pridgeon JW, Zhang D, Zhang L. Complete genome sequence of the attenuated Novobiocin-resistant streptococcus iniae vaccine strain ISNO. Genome Announc. 2014;2(3):2007–8.

92. Sun JR, Yan JC, Yen CY, Lee SY, Lu JJ. Invasive infection with streptococcus iniae in Taiwan. J Med Microbiol. 2007;56(9):1246–9.

93. Rajoo S, Jeon W, Park K, Yoo S, Yoon I, Lee H, et al. Complete genome sequence of streptococcus iniae YSFST01-82, isolated from olive flounder in Jeju, South Korea. Genome Announc. 2015;3(2):10–1.

94. Holden MTG, Heather Z, Paillot R, Steward KF, Webb K, Ainslie F, et al. Genomic evidence for the evolution of Streptococcus equi host restriction, increased virulence, and genetic exchange with human pathogens. PLoS Pathog. 2009;5(3):e1000346.

95. Pelkonen S, Lindahl SB, Suomala P, Karhukorpi J, Vuorinen S, Koivula I, et al. Transmission of streptococcus equi subspecies zooepidemicus infection from horses to humans. Emerg Infect Dis. 2013;19(7):1041–8.

96. Ma Z, Geng J, Zhang H, Yu H, Yi L, Lei M, et al. Complete genome sequence of Streptococcus equi subsp. Zooepidemicus strain ATCC 35246. J Bacteriol. 2011;193(19):5583–4.

97. Zhang Y, Ding D, Liu M, Yang X, Zong B, Wang X, et al. Effect of the glycosyltransferases on the capsular polysaccharide synthesis of Streptococcus suis serotype 2. Microbiol Res. 2016;185:45–54. https://doi.org/10.1016/j.micres.2016.02.002.

98. Zhang Y, Lu C, Dong W, Pan Z, Zhong X, Lu P, et al. SssP1, a Streptococcus suis fimbria-like Protein transported by the SecY2/A2 system, contributes to bacterial virulence. Appl Environ Microbiol. 2018;84(18):1–17.

99. Chen C, Tang J, Dong W, Wang C, Feng Y, Wang J, et al. A glimpse of streptococcal toxic shock syndrome from comparative genomics of S. suis 2 Chinese isolates. PLoS One. 2007;2(3):e315.

100. Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I, et al. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. Nucleic Acids Res. 2012;40(D1):57–63.

101. Wang K, Chen J, Yao H, Lu C. Whole-genome sequence of Streptococcus suis serotype 4 reference strain 6407. Genome Announc. 2014;2(4):9–10.

102. Pan Z, Ma J, Dong W, Song W, Wang K, Lu C, et al. Novel variant serotype of Streptococcus suis isolated from piglets with meningitis. Appl Environ Microbiol. 2015;81(3):976–85.

103. Xiao J, Hu P, Chen H, Wu J, Hua Y, Zhang A, et al. Comparative genomic analysis of Streptococcus suis reveals significant genomic diversity among different serotypes. BMC Genomics. 2011;12(1):523 Available from: http://www.biomedcentral.com/1471-2164/12/523.

104. Chatellier S, Harel J, Zhang Y, Higgins R, Brousseau R, Gottschalk M, et al. Phylogenetic diversity of Streptococcus suis strains of various serotypes as revealed by 16S rRNA gene sequence comparison. Int J Syst Bacteriol. 2009; 48(2):581–9.

105. Pan X, Tang J, Zhao Y, Li M, Yao X, Hu D, et al. Isolation and characterization of a native avirulent strain of Streptococcus suis serotype 2: a perspective for vaccine development. Sci Rep. 2015;5(1):1–8.

106. Boyle B, Vaillancourt K, Bonifait L, Charette SJ, Gottschalk M, Grenier D. Genome sequence of the swine pathogen streptococcus suis serotype 2 strain S735. J Bacteriol. 2012;194(22):6343–4.

107. Wang K, Yao H, Chengping Lu JC. Complete genome sequence of Streptococcus suis serotype 16 strain TL13. J Bacteriol. 2011;193(9): 2375–6.

108. Zheng H, Du P, Qiu X, Kerdsin A, Roy D, Bai X, et al. Genomic comparisons of Streptococcus suis serotype 9 strains recovered from diseased pigs in Spain and Canada. Vet Res. 2018;49(1):1–13. https://doi.org/10.1186/s13567-017-0498-2.

109. Szafrański SP, Deng ZL, Tomasch J, Jarek M, Bhuju S, Rohde M, et al. Quorum sensing of Streptococcus mutans is activated by Aggregatibacter actinomycetemcomitans and by the periodontal microbiome. BMC Genomics. 2017;18(1):1–15.

110. Cook LC, LaSarre B, Federle MJ. Interspecies communication among commensal and pathogenic streptococci. MBio. 2013;4(4):1–11.

111. Brouwer S, Cork AJ, Ong C-LY, Barnett TC, West NP, McIver KS, et al. The endopeptidase PepO regulates the SpeB cysteine protease and is essential for the virulence of invasive M1T1 *Streptococcus pyogenes*. J Bacteriol. 2018; (January):JB.00654–17 10.1128/JB.00654-17.

112. Cheung GYC, Otto M. Understanding the significance of Staphylococcus epidermidis bacteremia in babies and children. Curr Opin Infect Dis. 2010; 23(3):208–16.

113. Herman-Bausier P, Labate C, Towell AM, Derclaye S, Geoghegan JA, Dufrêne YF. Staphylococcus aureus clumping factor a is a force-sensitive molecular switch that activates bacterial adhesion. Proc Natl Acad Sci. 2018;115(21): 5564–9.

114. GO CM. db: a set of annotation maps describing the entire gene ontology; 2017.

115. Galili T. Dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics. 2015;31(22):3718–20.

116. Galili T, O'Callaghan A, Sidi J, Sievert C. Heatmaply: an R package for creating interactive cluster heatmaps for online publishing. Bioinformatics. 2018;34(9):1600–2.

117. R Core Team. R: A language and environment for statistical computing [internet]. Vienna: R Core Team; 2017. Available from: https://www.r-project.org/

118. Van Der ML. Accelerating t-SNE using tree-based algorithms. J Mach Learn Res. 2014;15:3221–45 Available from: http://jmlr.org/papers/v15/vandermaaten14a.html%0A, http://jmlr.org/papers/v15/vandermaaten14a.html%257B%2525%257D5Cnfiles/1017/JMLR-van%257B_%257Dder%257B_%257DMaaten-2014-Accelerating%257B_%257Dt-SNE%257B_%257Dusing%257B_%257DTree-Based%257B_%257DAlgorith.

119. Wickham H. ggplot2: elegant graphics for data analysis [internet]. New York: Springer-Verlag; 2016. Available from: https://ggplot2.tidyverse.org

Publisher's Note