Capturing the Evolutionary Divergence of Antibiotic Resistant Genes in Pathogenic vs. Non-pathogenic Staphylococcus Species

Yuanyuan Wu, Ashley Ahmed, and Sam Greenspun

PLBIO4000

5/15/22

Introduction

Staphylococcus is a widely-known bacteria genus that contains many species known to cause severe physiological distress in humans to other animals (pathogenic). In fact, many common lethal conditions that are associated with staphylococcus infection include Pneumonia (*S.aureus*), Toxic Shock Syndrome (*S.aureus*), and urinary tract infections (*S.saprophyticus*), with many of those who are prone to develop such infections are most often immuno-comprimised (e.g. neonates, patients in critical condition). Although the majority of staphyloccus species are not harmful to other organisms (non-pathogenic), the adverse effects of pathogenic species has resulted in the rise of antibiotics to manage and kill bacterial infections.

Antibiotics are medications that control and fight pathogenic bacterial infections in humans and in other organisms. In staphylococcus infections, the most commonly used antibiotics are Penicillin and Methicillin. However, since the initial release of Penicillin(~1940) and Methicillin(~1980), in both a hospital and commercial setting, staphylococcus infections have been more difficult to treat, leading to the hypothesis that bacterial species have evolved mechanisms that protect them from the effects of antibiotics (Fig 1); this is known as antibiotic resistance. Antibiotic resistance has been observed in other bacterial species and fungi and has been hypothesized to be acquired by horizontal/vertical transfer of genes and selective pressure between bacterial species (Fig 2).

In the staphylococcus genus, antibiotic resistance genes have been shown to be shared between specific pathogenic and non-pathogenic Staphylococcus species. This has led many to believe that rapid antibiotic resistance may be attributed to frequent horizontal and vertical transfer of resistance genes, driven by selective pressure, between specific pathogenic and non-pathogenic species. To better characterize the relationships between antibiotic resistance and pathogenicity in staphylococcus species, we aim to employ machine learning and phylogenetic analyses to assess 1) the most common resistance genes across pathogenic vs non-pathogenic species, 2) to determine if there is a resistance gene that can be used to classify the pathogenic and non-pathogenic strains, and 3) to characterize possible specific species transfers of antibiotic resistance genes between pathogenic and non-pathogenic strains.

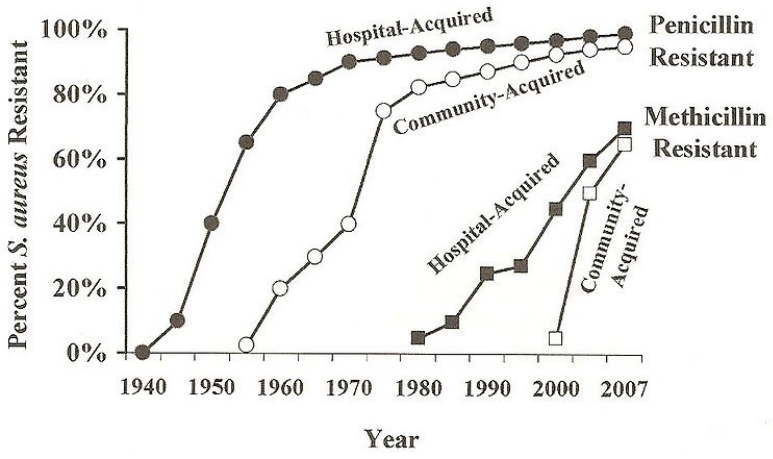

Fig 1. The discovery of Penicillin was revolutionary in treating S. aureus infections in the 1940s; figure is provided by [1]. However, after a few years after its introduction, penicillin resistance was encountered in S. aureus in commercial and in clinical practice. As a result, the prevalence of antibiotic-resistant strains of S. aureus has increased and challenged the healthcare community to continuously update antibiotic medications. Antibiotic resistance has also been observed in Methicillin, another antibiotic used to treat S. aureus strains infection [1].

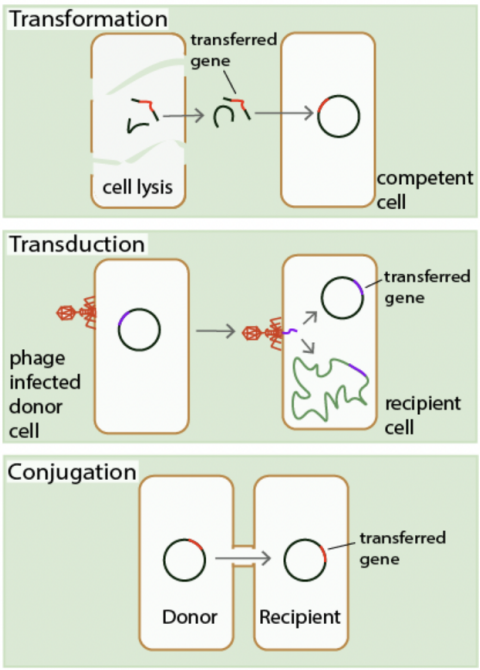

Fig 2. There are three horizontal gene transfer mechanisms that occur between different

bacterial species: Transformation, Transduction, and Conjugation; figure is provided by [2]. Transformation is the uptake of DNA from other bacterial cells and transduction is the transfer of DNA between bacterial species by a bacteriophage. Conjugation is the transfer of DNA through cell to cell contact between bacterial species. Antibiotic resistance has been attributed to horizontal gene transfer across bacterial species, specifically in the staphylococcus genus [2] .

Materials and Methods

Data Availability

Whole genome and 16S ribosomal RNA (rRNA) fasta files containing DNA sequences for 12 different *Staphylococcus* species were retrieved from National Center for Biotechnology Information (NCBI): 6 pathogenic *Staphylococcus* species and 6 non-pathogenic *Staphylococcus* species. Whole genome files were used to identify antibiotic-resistant genes of interest. 16S rRNA files were used to construct a species phylogenetic tree.

Resistive Gene Identification and Retrieval

All bacterial genomes were entered into the resistance gene identifier (RGI) to identify genes that are associated with antibiotic resistance. Identified resistance genes are then separated in an output file containing the unaligned sequence by RGI. RGI uses Hidden markov model and 'Blast' to extract all possible resistance genes and output the %identity of the extracted sequence to the reference. A script was developed to retrieve resistance gene sequences from the output file and identify common genes present in the 12 staphylococcus species in an excel sheet. The RGI output was also used to generate strict hit genes (identity>95%) heatmap.

Supervised machine learning to identify possible genes that identify pathogenic and non-pathogenic strains

The RGI output with %identity to the reference antibiotic genes were used as the features, and bacterial pathogenicity were used as the labels. The genes not shared in specific strains were filled with 0. Pycaret was used to generate and determine the best model.

Evolutionary Analysis

Molecular Evolutionary Genetics Analysis (MEGA) v6 [6] software was used to perform multiple sequence alignment on collected datasets and construct maximum likelihood phylogenetic trees for species and gene sequences. The best model for the substitution pattern was predicted using MEGA software and selected based on the lowest Bayesian Information Criterion (BIC).

Notung 2.9 platform [8] was used to root the un-rooted species tree obtained from MEGA. Additionally, Notung was used to reconcile the gene and species tree and infer duplication, transfer, and loss events based on the Newick phylogenetic tree obtained from MEGA.

Data Monkey is an adaptive evolution server that was used to detect possible recombination events caused by horizontal and vertical gene transfers in resistive genes. Detection of recombination events was used to develop phylogenetic trees based on sequence segments between staphylococcus species.The tool that was used to perform this analysis was the Genetic Algorithm for Recombination Detection tool (GARD). GARD algorithm screens sequence alignments for possible recombination breaking points between species to infer unique phylogenetic trees for each detected recombination event region. Recombination breaking points are detected based on the algorithm described in [11]. The algorithm infers a neighbor joining end tree for all alignment sequences to obtain an Alkaline Information Criterion Score(A0) score that uses maximum likelihood to estimate rate parameters and branch lengths. Alignment sites are partitioned into 2 continuous blocks to screen for signficant variability; possible breaking point will coincide with a variable site. If the break point is placed at site , an NJ tree is created individually for each block and computes the AICc score (Ai) of the model that fits branch lengths to each partition independently.

Results and Discussion

Extract shared Antibiotic-Resistance Genes Between Pathogenic and Non-pathogenic Staphylococcus Species from RGI outputs for phylogenetic analysis

With the RGI output json files, we reimported the file into RGI in terminal to generate a heatmap for the shared genes, and cluster both strains and the genes. Yellow indicates a perfect hit to the reference gene sequence in RGI, Green indicates strict hit, and purple indicates no strict or perfect hits. According to this, we picked 4 most shared genes: sepA, sdrM, norC and quaJ.

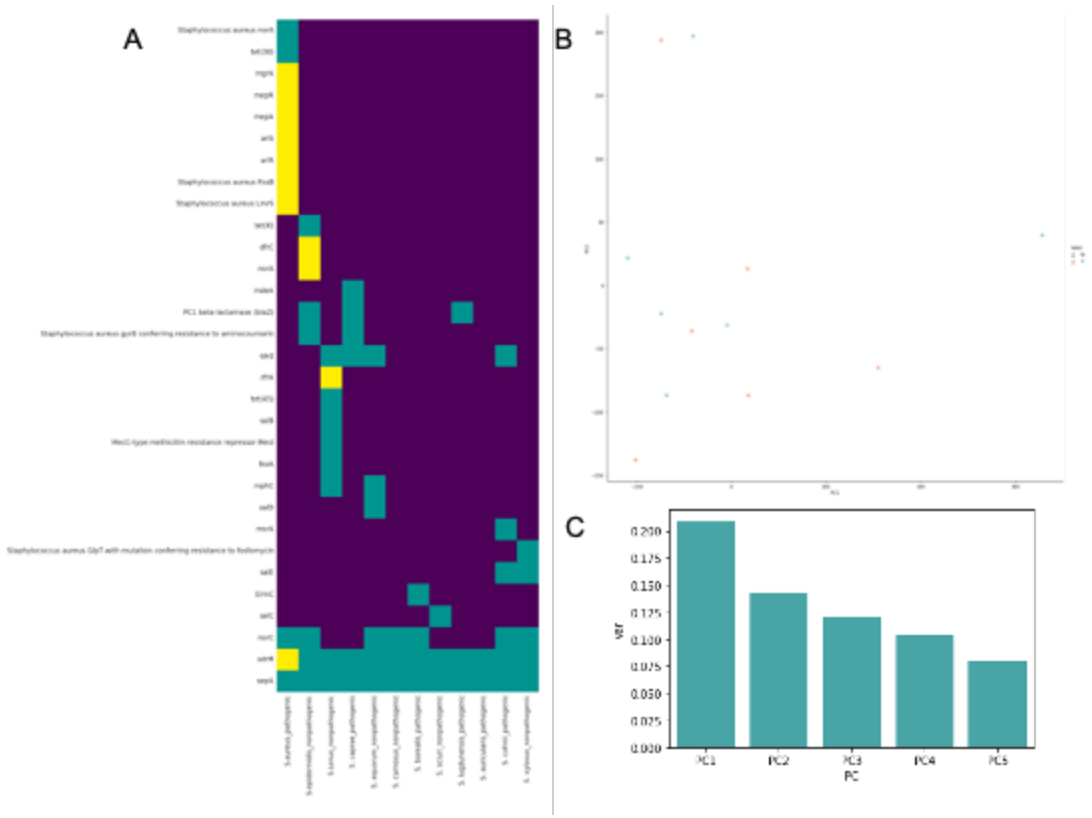

Fig 3: RGI outputs summary. **(A)** Heatmap for shared strict hit RGI genes. **(B)** PCA for all identified genes. **(C)** The top5 explained variance ratio for PCA.

We clustered all the strains in PCA with their RGI hits and found that no clear clustering for pathogenic and non-pathogenic strains were found. Fig 3C also indicates that little variance can be explained by the topThis is consistent with Fig 3A where pathogenic and non-pathogenic strains were mixed.

mexS Gene Identified as a Grouping Variable Between Pathogenic and Non-Pathogenic Bacterial Species

The best model to identify the pathogenicity of the bacterial strains were AdaBoostClassifier(algorithm='SAMME.R', base_estimator=None, learning_rate=1.0, n_estimators=50, random_state=8606). With this, AUC = 0.83 was achieved. We identify mexS as the only gene that can possibly identify bacterial pathogenicity.

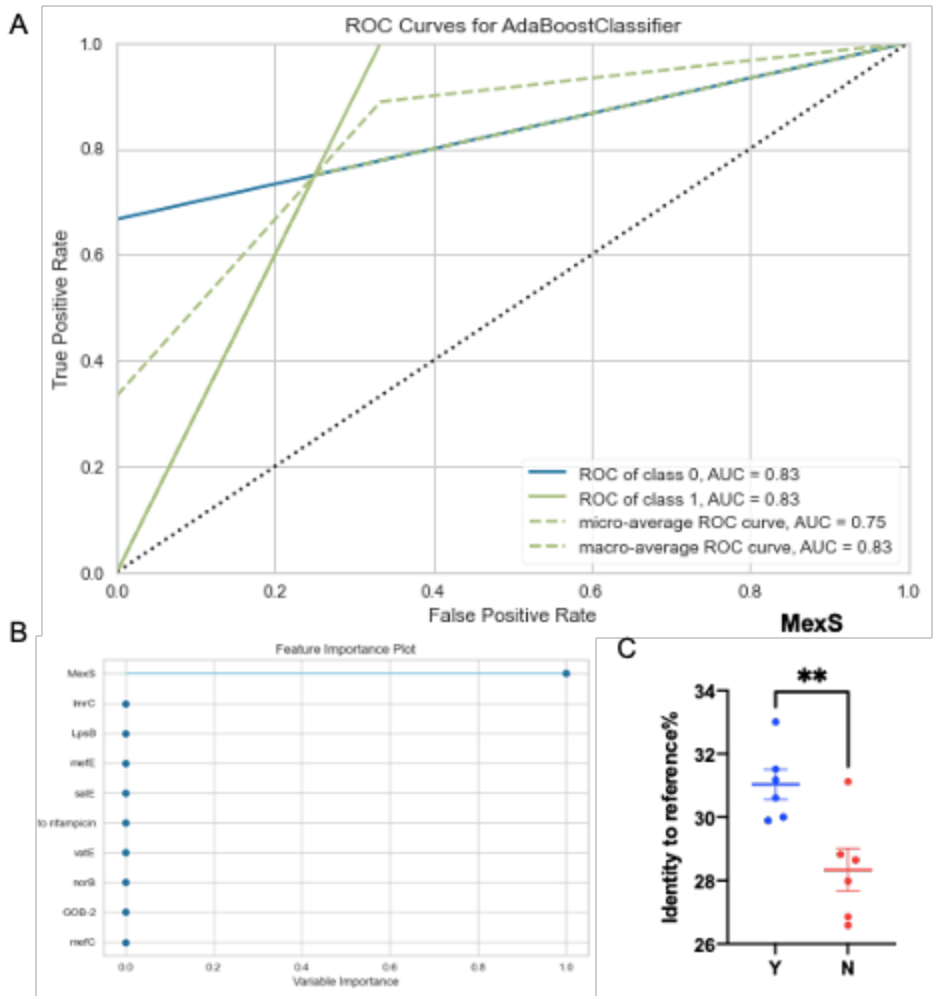

Fig 4. MexS is identified as a gene for pathogenicity prediction. **A)** ROC curves for the selected model performance. **B)** Identified features. **C)** MexS identity to reference distribution in pathogenic(Y) and non-pathogenic(N) strains.

A closer look at the MexS gene showed that MexS is a secondary antibiotic gene expression regulator. It is responsible for multiple antibiotic resistances. The functional effects of MexS can be further explored.

Staphylococcus **species tree**

Fig 5 demonstrates that no observable pattern in evolutionary relatedness between pathogenic and non-pathogenic *Staphylococcus* species.

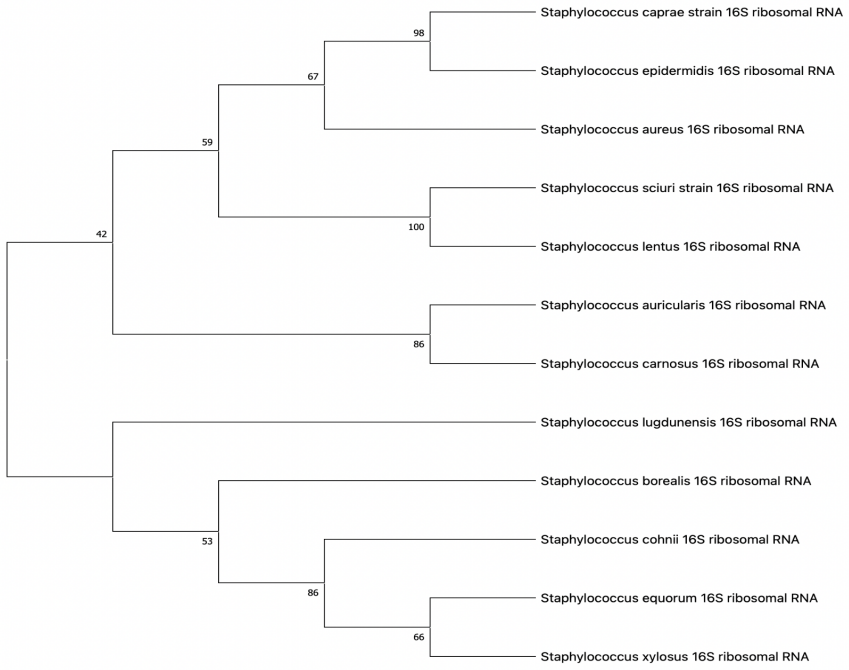

Fig 5. Phylogenetic tree constructed in MEGA using 16S rRNA sequences from pathogenic and non-pathogenic *Staphylococcus* species with bootstrap values near nodes.

Pathogenic and Non-Pathogenic Staphylococcus Species Shows Evolutionary Divergence At Different Genetic Regions of Shared Antibiotic Resistance Genes

We identified four common antibiotic resistance genes in pathogenic and non-pathogenic staphylococcus species: qacJ, norC, sepA, and sdr. For each gene, we compared phylogenetic trees between shared antibiotic resistance genes using MEGA and Notung and identified high areas of recombination in specific genetic regions using GARD.

The quaJ gene is associated with resistance to ammonium based disinfectants. According to RGI analysis, quaJ is present only in the *S. caprae*, *S. cohnii*, *S. equorum*, and *S. lentus* species. Fig 6A describes the phylogenetic tree constructed using the T92 + G substitution model. Based on the tree in Fig 6A, *S. caprae* and *S. cohnii*, both pathogenic species, demonstrate similar relatedness at the molecular level relative to *S. equorum* and *S. lentus*, both non-pathogenic species. This evidence supports the development of antibiotic resistance in pathogenic *Staphylococcus* species. The quaJ phylogenetic tree was rooted using Notung's scoring algorithm where the highest scored branch was selected to root the tree. Based on Fig 6B, Notung predicted a duplication event at the ancestral node for *S. caprae*, *S. cohnii*, *S. equorum*, and *S. lentus* species which could suggest possible mutations that resulted in the rise of pathogenic *S. caprae*

and *S. cohnii*. Additionally, Notung predicted 4 loss events in all species, possibly suggesting the loss of genomic material from the prior duplication event. GARD analysis showed 2 possible recombination breaking points in the 40-50 bp and 185-200 bp region of the qacJ gene between *S. caprae*, *S. cohnii*, *S. equorum*, and *S. lentus* species (Fig 7A). Phylogenetic trees based on the two different breaking point regions showed a divergence between pathogenic and non-pathogenic bacteria in the 52-324 bp region (Fig 7B). In the 1-51 bp region, pathogenicity was not a factor in the tree phylogeny. This indicates that a mutation may be in the 52-324 bp region in pathogenic species that is not present in non-pathogenic bacteria and this region may indicate a transfer of DNA between the two pathogenic species.

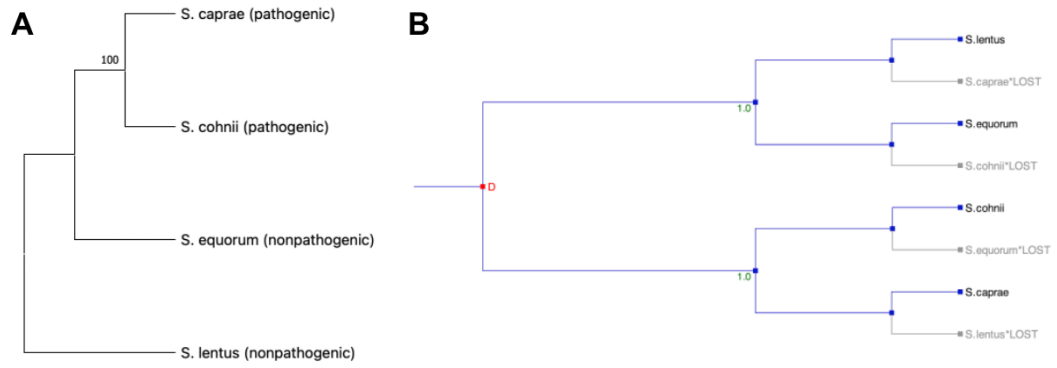

Fig 6. (A) MEGA phylogenetic tree construction of aligned quaJ resistant genes in *S. caprae*, *S. cohnii*, *S. equorum*, and *S. lentus* with bootstrap values. **(B)** Notung-inferred duplication (red D) and loss events (in gray) for quaJ.

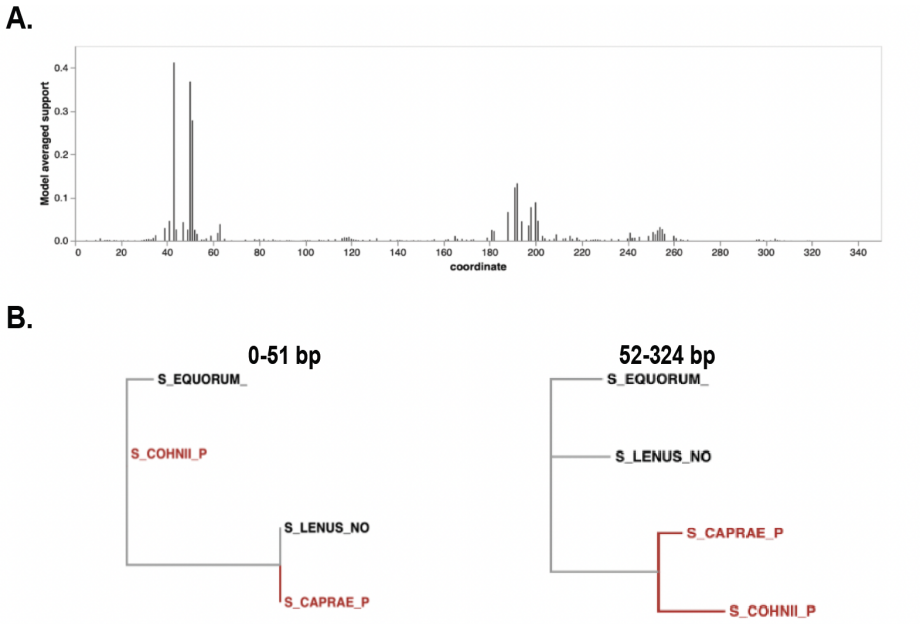

Fig 7. (A) Coordinate prediction of recombination breaking point region in the qacJ gene between S.*caprae*, *S. cohnii*, *S. equorum*, and *S. lentus.* Regions include the 0-51 bp region and the 52-324 bp region of qacJ. **(B)** Phylogenetic trees for each breaking point region show that 1) S.*caprae(P)* and *S. lentus(NP)* and 2) *S. cohnii(P)* and *S. equorum(NP)* are closely related at the 0-51 bp region. At the 52-324 bp region, pathogenic and non-pathogenic species show a genetic divergence; red is used to indicate pathogenic species.

The norC gene is associated with resistance of the Fluoroquinolone antibiotics and common disinfectants and antiseptics including Clorox, Lysol and Sanisol.() Based on Fig 8A, no observable relationship is inferred from the phylogenetic tree. Fig 8A is a phylogenetic tree constructed using the T92 + G substitution model. Pathogenic species *S. caprae*, *S. borealis*, and *S. auricularis* appear to have diverged from non-pathogenic *S. carnosus*. The rest of the tree does not demonstrate any pattern in norC gene evolution between pathogenic and non-pathogenic bacteria. Additionally, there does not seem to be much similarity in the norC sequence across the various pathogenic and non-pathogenic *Staphylococcus* species which could be an indicator that this gene is not heavily used in the pathogenic species to resist antibiotics or it is involved in low-level resistance where other genes are more involved. Fig 8B depicts multiple common ancestral nodes that Notung predicts as possible transfer events of norC. There is a general pattern in the reconciled tree of norC being transferred to pathogenic species such as *S. borealis*, *S. lugdunensis*, and *S. aureus*. Additionally, Notung predicted a loss of the norC gene from *S. caprae*, *S. aureus*, *S. borealis*, *S. equorum*, and *S. lugdunensis*

which are all pathogenic species. The predicted transfer and eventual loss is contradicting, thus, this could potentially signify limited use of this specific gene. However, GARD analysis showed 3 possible recombination breaking points in the 0-1500 bp,1501-2760 bp, and 2761-3000 bp region of the norC gene between pathogenic species and non-pathogenic species (Fig 9A). Phylogenetic trees showed a divergence between pathogenic and non-pathogenic bacteria in the 1501-2760 bp region (Fig 9B); other norC regions showed no relationship with pathogenicity. The 1501-2760bp could be a region that contains a mutation specific for pathogenicity and high levels of horizontal/vertical gene transfer.

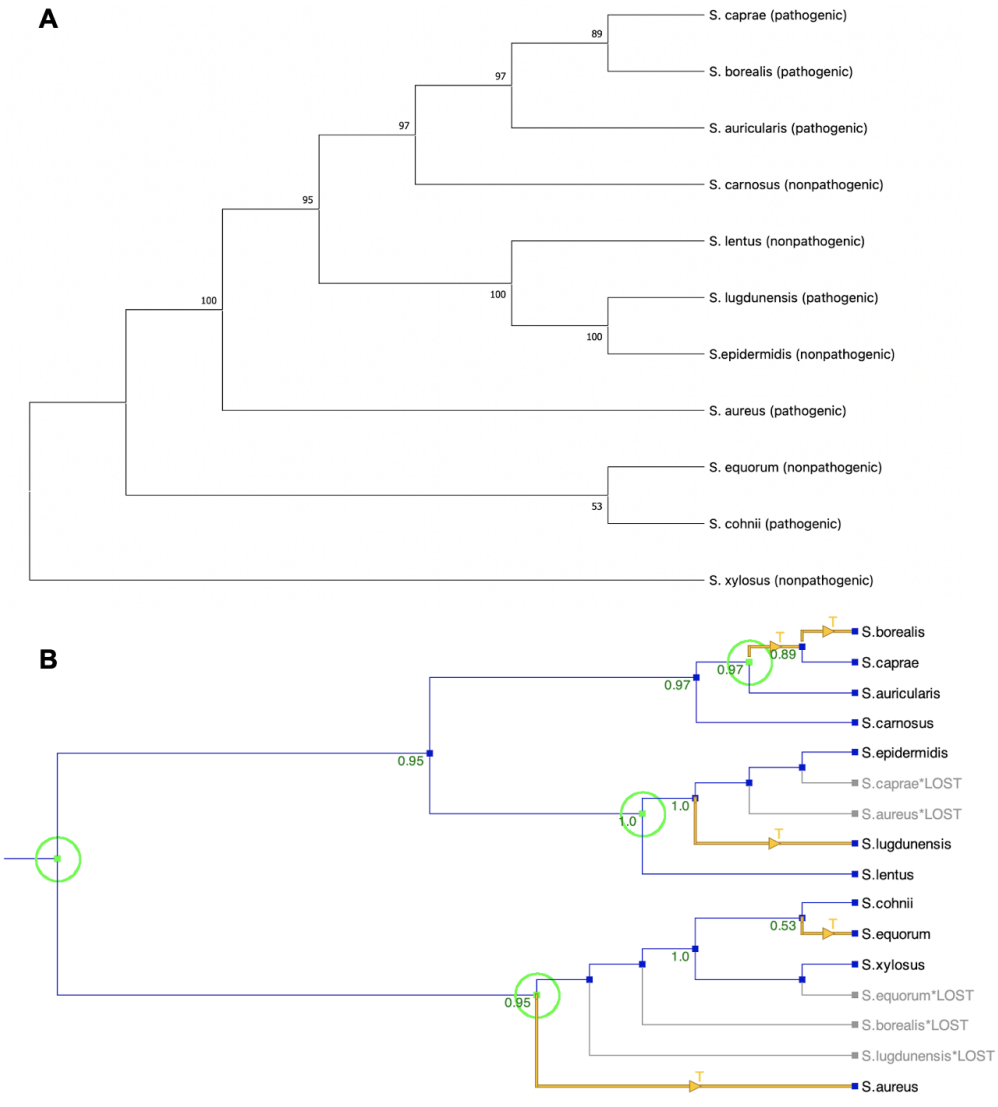

Fig 8. (A) MEGA phylogenetic tree construction of aligned norC resistant genes in 12 *Staphylococcus* species with bootstrap values **(B)** Notung-inferred duplication, loss, and transfer events (in yellow and denoted using 'T') for norC sequences.

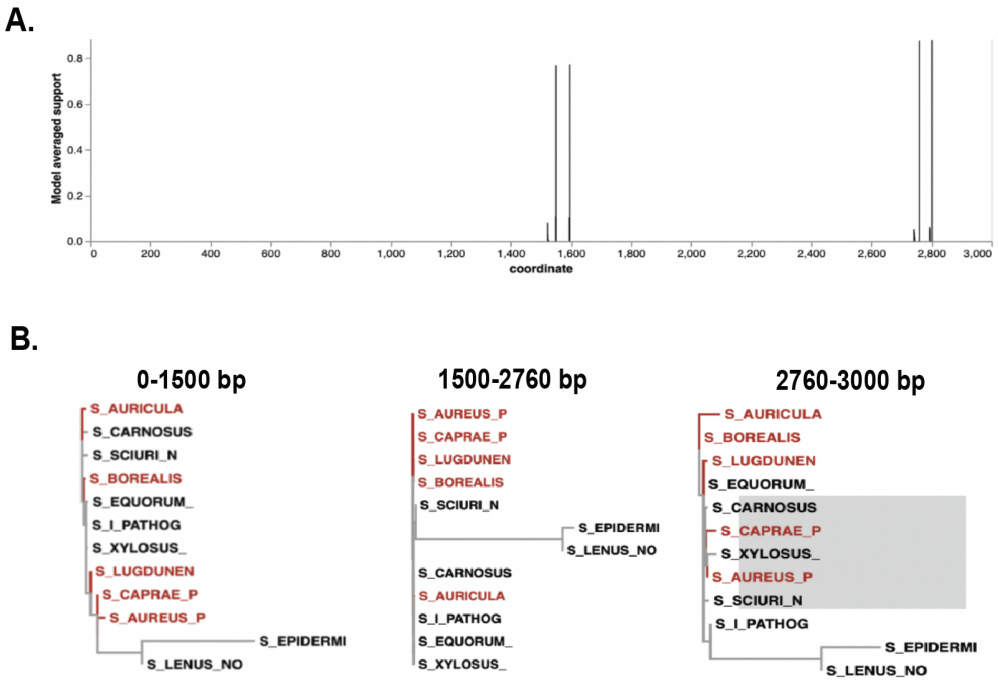

Fig 9. (A) Coordinate prediction of recombination breaking point region in the norC gene between pathogenic and non-pathogenic staphylococcus species. Regions include the 0-1500 bp region, the 1501-2760 bp region, and the 2760-3000 bp region of norC. **(B)** Phylogenetic trees for each breaking point region show that pathogenic and non-pathogenic species show a genetic divergence between the 1500-2760 region.

The sepA gene is associated with resistance of common disinfectants and antiseptics and was found in all pathogenic and non-pathogenic staphylococcus species. T92 + G substitution model was used to construct the phylogenetic tree in Fig 10A. Pathogenic and non-pathogenic Staphylococcus species share a common ancestral node. Non-pathogenic species S. sciuri and S. lentus appear to diverge relatively early. Pathogenic species such as *S. aureus*, *S. borealis*, and *S. lugdunensis* that likely express sepA more due to human-induced selection pressures largely appear to evolve separately from other non-pathogenic species by sharing molecular similarity. This evidence further supports the claim that pathogenic species are evolving antibiotic resistance separately from non-pathogenic species, and that pathogenic species share higher molecular similarity which is likely attributed to similar functional use of the gene (Fig 10A). Notung predicted a single duplication event at the ancestral node of bifurcated phylogenetic tree. Notung predicted the loss of sepA function in *S. carnosus* and *S. auricularis*. Based on the Notung reconciled tree, the majority of the pathogenic species including *S. caprae*, *S. aureus*, *S. lugdunensis*, and *S. borealis* diverged from the non-pathogenic species *S. equorum*, *S. xylosus*, *S. lentus*, and *S. sciuri* after the duplication event. Furthermore, Notung predicted

that the common ancestor for S. cohnii and S. xylosus had transferred the sepA gene only to S. cohnii (Fig 10B). This could potentially serve as evidence that antibiotic resistance contributes to pathogenicity in some *Staphylococcus* species. This is supported by GARD analysis which showed 3 possible recombination breaking points in 0-151 bp region, the 152-434 bp region, and the 435-500 bp region of norC (Fig 11A). Phylogenetic trees showed a genetic divergence between pathogenic and non-pathogenic bacteria in the 152-434 bp region between pathogenic and non-pathogenic bacteria (Fig 11B). This region may be where the duplication may have occurred that resulted in the divergence of pathogenic and non pathogenic bacteria.

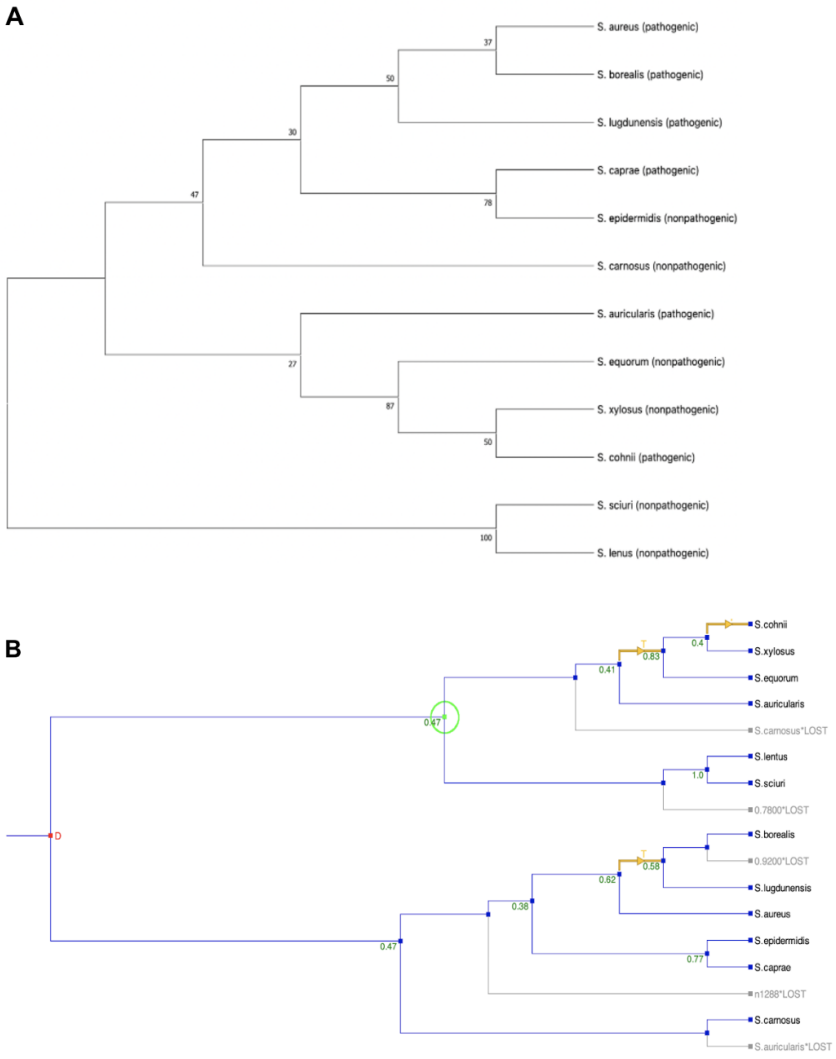

Fig 10. (A) MEGA phylogenetic tree construction of aligned sepA resistant genes in 12 *Staphylococcus* species with bootstrap values **(B)** Notung-inferred duplication, loss, and transfer events for sepA sequences.

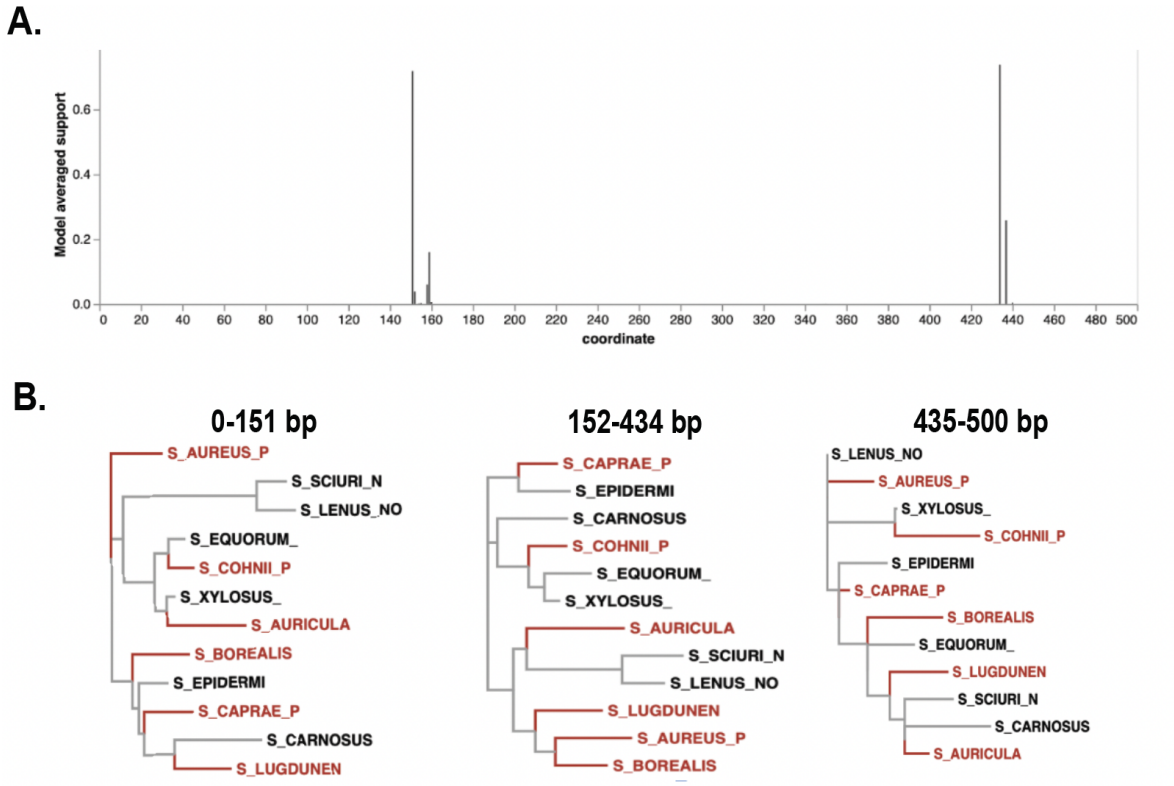

Fig 11. **(A)** Coordinate prediction of recombination breaking point region in the sepA gene between pathogenic and non-pathogenic staphylococcus species. Regions include the 0-151 bp region, the 152-434 bp region, and the 435-500 bp region of norC. **(B)** Phylogenetic trees for each breaking point region show that pathogenic and non-pathogenic species show a genetic divergence between the 152-434 bp region.

The srdM gene is associated with resistance of common disinfectants and antiseptics and fluoroquinolone. GTR + G was the substitution model used to construct the phylogenetic tree in Fig 12A. Non-pathogenic *Staphylococcus* species have similar sdrM sequences as many of them are found within the same clade. Additionally, the non-pathogenic species generally evolve from a pathogenic *Staphylococcus* species. This could indicate limited use of the sdrM gene and low-level antibiotic resistance. Fig 12B predicts multiple transfer events for sdrM. Generally, sdrM appears to transfer to clades containing majority non-pathogenic *Staphylococcus*. This further reinforces that sdrM may not be heavily involved in antibiotic resistance as its transfer shows no observable relationship with pathogenicity.

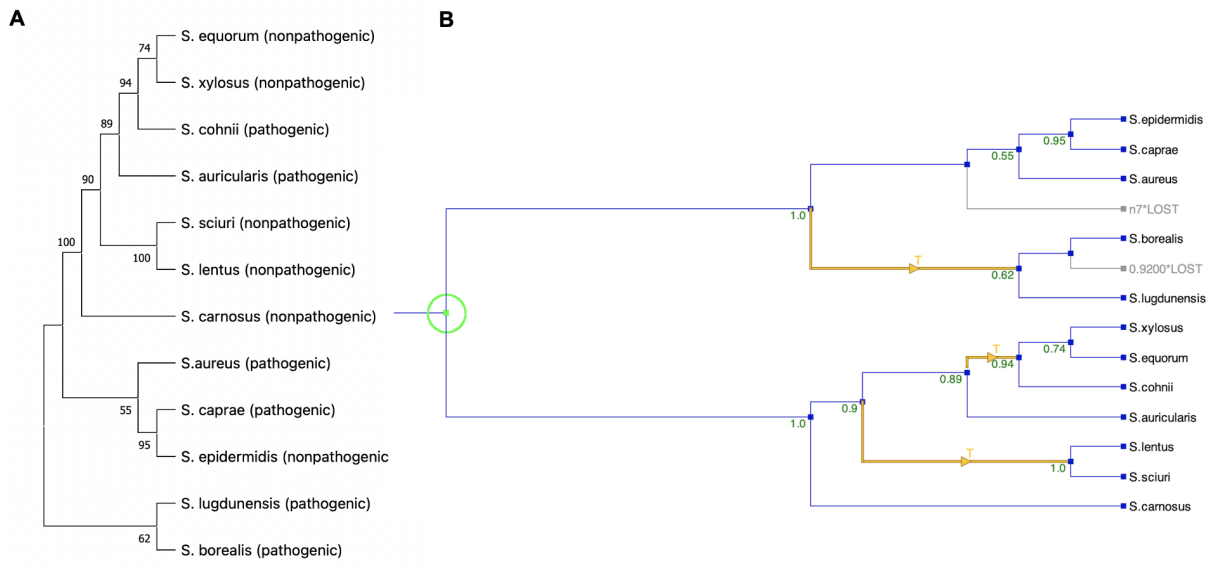

Fig 12. (A) MEGA phylogenetic tree construction of aligned sdrM resistant genes in 12 *Staphylococcus* species with bootstrap values **(B)** Notung-inferred transfer events for sdrM sequences.

Conclusion:

To start with, we found that non-pathogenic strains contain as many resistance genes and the RGI difference between the pathogenic and non-pathogenic strains is not distinguishable. We exhausted all the genes we got and only found MexS as a possible identifier. It possibly indicates MexS's function above resistance gene regulation. And indeed we found an alternative MexS function in the literature to regulate type III secretion activity [10]. We also identified that the major antibiotic resistance mechanism shared among all strains was antibiotic efflux, by using antibiotic efflux pumps. And the most prevalent antibiotic resistance were targeting fluoroquinolone, disinfecting agents and antiseptics.

Based on phylogenetic tree analysis, our findings show that the quaJ and sepA genes demonstrate evolutionary relatedness at the genetic level in pathogenic species which appear to have evolved from non-pathogenic *Staphylococcus*. We provide potential evolutionary hypotheses such as duplication, transfer, and loss events using Notung which may explain this phenomenon.

By mapping out the evolution of antibiotic resistance genes, we have been able to observe possible non-pathogenic to pathogenic species relationships in trans. Our data suggests that common antibiotic resistance genes share a genetic divergence possibly based on pathogenicity in specific sequence regions, indicating that this may be a specific target area for therapeutic treatment. However, further analysis is needed to

explore the evolution of MexS in pathogenicity and confirm specific genetic regions to pathogenic and non-pathogenic species.

Disclosure:

It should be noted that we used the reference genome of the bacteria for analysis and the reference genome is usually the pangenome of all the bacteria of the same strains. It does not reflect the distribution of the antibiotic genes in the individual bacteria.

References:

1. Chambers HF, Deleo FR. Waves of resistance: Staphylococcus aureus in the antibiotic era. *Nat Rev Microbiol*. 2009;7(9):629-641. doi:10.1038/nrmicro2200
2. Pigłowski M. Pathogenic and Non-Pathogenic Microorganisms in the Rapid Alert System for Food and Feed. *Int J Environ Res Public Health*. 2019;16(3):477. Published 2019 Feb 6. doi:10.3390/ijerph16030477
3. D'Costa, V. M., McGrann, K. M., Hughes, D. W., and Wright, G. D. (2006). Sampling the antibiotic resistome. *Science* 311, 374–377. doi: 10.1126/science.1120800
4. McArthur *et al*. 2013. The Comprehensive Antibiotic Resistance Database. *Antimicrobial Agents and Chemotherapy*, 57, 3348-57.
5. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database
6. Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Molecular Biology and Evolution, 30, 2725–2729. https://doi.org/10.1093/molbev/mst197.
7. Bello-López, J.M., et al., *Horizontal Gene Transfer and Its Association with Antibiotic Resistance in the Genus Aeromonas spp.* Microorganisms, 2019. 7(9): p. 363.
8. Charlotte A Darby, Maureen Stolzer, Patrick J Ropp, Daniel Barker, Dannie Durand, Xenolog classification, Bioinformatics, Volume 33, Issue 5, 1 March 2017, Pages 640–649, https://doi.org/10.1093/bioinformatics/btw686
9. Kosakovsky Pond, SL et al. "Automated Phylogenetic Detection of Recombination Using a Genetic Algorithm." Mol. Biol. Evol. 23, 1891–1901 (2006).
10. Jin, Yongxin, et al. "MexT regulates the type III secretion system through MexS and PtrC in Pseudomonas aeruginosa." Journal of bacteriology 193.2 (2011): 399-410.
11. Kosakovsky Pond, SL et al. "Automated Phylogenetic Detection of Recombination Using a Genetic Algorithm." Mol. Biol. Evol. 23, 1891–1901

(2006).