Investigation of 6 Vulnerable Strains of MRSA, High Areas of Occurrence for MRSA, and Protein Structure and Function Variants for BlaZ gene.

Hasika Oggi

Table of Contents:

Abstract:

Methicillin-resistant *Staphylococcus aureus* is a bacterium that is quickly developing resistance to many antibiotics including methicillin, penicillin, beta-lactam antibiotics and more. It causes many infections and can be fatal in some cases if not diagnosed and treated properly. While MRSA incidence has declined in some regions, it still is a clinical threat due to its high level of resistance to modern antibiotics. Successfully eradicating MRSA will take time, but this review aims to look at the gene diversity between various strains of MRSA. This ultimately can be useful for doctors when deciding the most effective treatment regimen for a patient. NCBI, PathogenWatch, and BLAST were used to search for MRSA strains and discover what they were resistant to. NCBI was used to download various genome assemblies using the search terms *Staphylococcus aureus* and to look at papers that provided further information on the onset and problems that MRSA causes. PathogenWatch was used to keep track of all the assemblies and to create a tree that would clearly showcase resistance genes present in the genomes. Finally, BLAST was used to check for gene diversity and see what genes would MRSA be most resistant to. I blasted 51 MRSA strains that are included in this paper and concluded that gene diversity is present in all but 6 strains which are extremely vulnerable to any sort of treatment. Most commonly, the resistance genes were *mecA*, *tetM*, and *ermA*, the most popular types of antibiotics used to treat MRSA. In addition, several strains in my collection had high gene diversity, having the sequence of almost every gene I blasted against them. When I looked at my metadata, I noticed that most of these strains were from South America, in areas such as Brazil and Argentina, or they were from East Asia, near Taiwan, which is further supported by the current literature. In conclusion, methicillin-resistant *S. aureus* is a health risk to most people, especially the elderly, athletes, soldiers, and those with weaker immune systems. It is imperative that more successful treatments be created for MRSA, but until then, we may have to resort to using multiple antibiotics to treat this disease. Possible treatments for MRSA may include multi-drug therapy or alternative therapeutic treatments.

Introduction:

Methicillin-resistant *Staphylococcus aureus* (MRSA) is caused by a type of *S. aureus* that is resistant to not only methicillin, but also other antibiotics used to treat staph infections (Mayo Clinic, n.d.). It mainly causes skin infections, but if left untreated, it can result in pneumonia, or even sepsis (CDC, 2019). MRSA is sometimes fatal, depending on the severity of the infection. Infection with MRSA may occur when healthy individuals touch objects that have been contaminated by infected people or are carrying the bacteria (Yuen *et al.,* 2015). This includes contact with an infected person as well. Those who are at higher risk for contracting MRSA are athletes, the elderly, daycare and school students, and military personnel in barracks because the risk of contracting MRSA increases in areas or activities that involve crowding, skin-to-skin contact, and shared equipment or supplies (Weber, 2009). Every 2 in 100 people carry the MRSA strain and MRSA is highly prevalent in hospitals throughout the world (Harvard Health, 2016). In addition, it’s common in regions in East Asia where there is an excessive amount of antibiotics used to treat staph infections as seen in Figure 1 (One Health Trust, 2010). MRSA presents a large threat to society, especially to those who are in the hospital or in nursing homes and are at higher risk of contracting this infection as shown in Figure 2 (Lee *et al.,* 2013). The resistance rate for strains of *S. aureus* globally to penicillin was 85.8%, for erythromycin 87.2%, and 90.8% resistance to ciprofloxacin (Rağbetli *et al.,* 2016). The mortality rate for those who have contracted hospital-acquired MRSA is 29% while those who have contracted community-acquired MRSA is 18%. This amounts to a rate of 6.3 deaths per 100,000 people in the United States (Clevens *et al.,* 2007). In addition, not only are there physical impacts for those with MRSA, but there are also psychological impacts on patients due to fear, discrimination, and isolation (Muenks *et al.,* 2018).

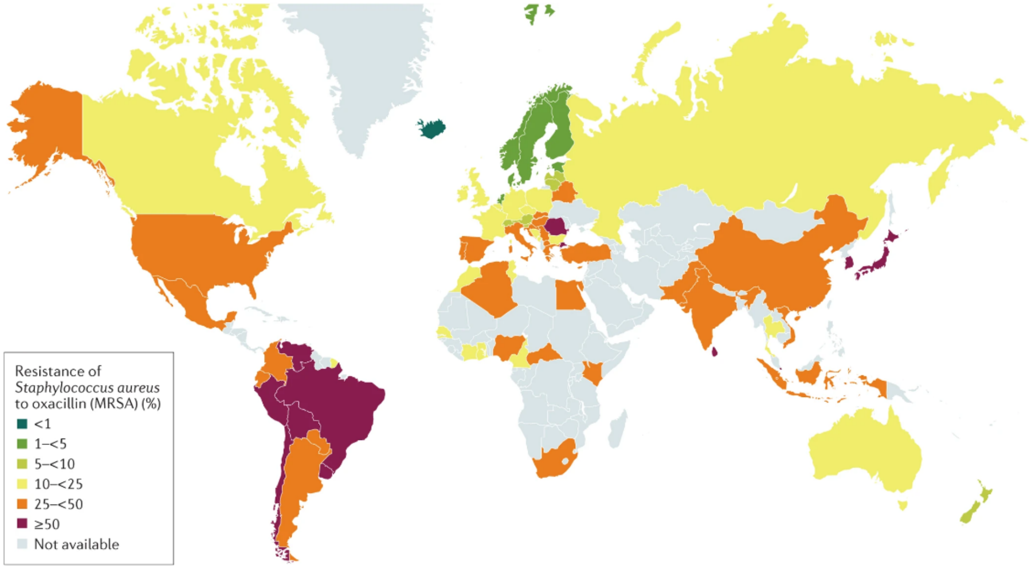


Figure 1: Resistance of *S. aureus* to oxacillin (MRSA strains) in 2018 globally. Taken from Nature (Lee *et al*., 2018). This image shows the resistance of *S. aureus* to oxacillin, which is in the same class of drugs as methicillin. While this study was done in 2018, it shows the prevalence of MRSA strains throughout the world, especially in regions such as South America, East Asia, and the US.

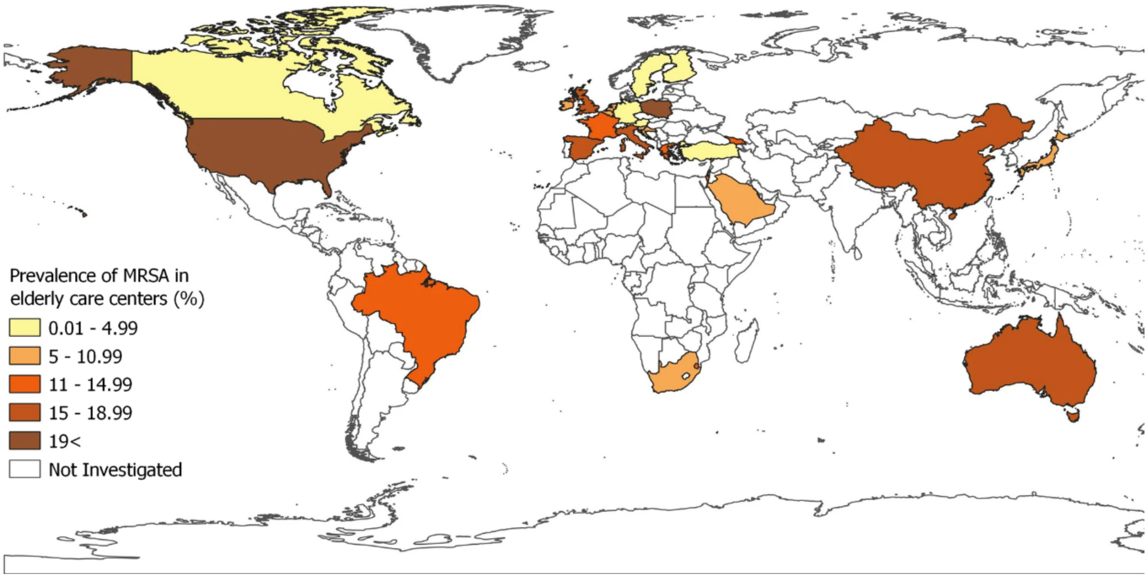


Figure 2: Prevalence of MRSA in elderly care centers in 2023 globally. Taken from Biomed Central (Hasanpour *et a*l, 2023). The elderly are among those who are at a high risk for contracting MRSA. In comparison to Figure 1, this image shows that there was a slight decrease in cases of MRSA; however, this was probably in part due to the pandemic. The lockdown isolated those with the infection, preventing much contact, thus leading to lower transmission rates. However, this does not negate the need for treatment, as people still suffer from this disease. In addition, now that the lockdown is no longer in place, cases may likely start to rise again.

There are seven common antibiotics used against MRSA, which are: vancomycin, daptomycin, linezolid, Sulfamethoxazole and trimethoprim (TMP-SMZ), quinupristin-dalfopristin, clindamycin and tigecycline (Okwu *et al.,* 2019). Treatment of MRSA at home usually includes a 7-10 day course of an antibiotic (by mouth) such as trimethoprim-sulfamethoxazole, clindamycin, minocycline, linezolid, or doxycycline (UpToDate, 2022). Right now, the most effective antibiotic to treat MRSA is vancomycin or daptomycin. However, MRSA is quickly developing resistance even to these antibiotics, so some healthcare providers have turned to therapeutic treatments to combat MRSA infections. These include quorum sensing inhibition, lectin inhibition, phage therapy, and more (Guo *et al.,* 2020). Researchers have also turned to using combination therapy, either with vancomycin or daptomycin with beta-lactam in order to see if there is successful clearance of persistent bacteremia caused by MRSA strains (Choo *et al.,* 2016).

There are individuals in the world who are suffering from this disease and MRSA is slowly becoming more widespread throughout the world, so it is imperative that some sort of antibiotic treatment be developed that can provide individuals with relief and also decrease the resistance of MRSA (Ventola, 2015). Treating MRSA costs about $10 billion per year, which averages about $60,000 per patient (Shinkman*,* 2016). Many people who are affected by MRSA in third-world countries may not have access to this kind of money (Zhen *et al.,* 2020). Therefore, the purpose of this paper is to evaluate gene diversity between different genes that strains of MRSA will develop resistance to if acquired. This information may be useful for doctors because based on the gene diversity and resistance in a strain, they can decide which treatment regimen to use to treat the patient.

Methods:

Data collection:

To begin collecting data, I downloaded both genomes and genes that I would be able to test using the software blast. Using the NCBI website on September 2, 2023, I searched for *Staphylococcus aureus* and downloaded the first 51 assemblies. When I searched for the *S. aureus* genome, the NCBI database automatically applied 2 filters: Latest and Exclude Anomalous. I clicked on the GenBank number and downloaded the complete record for each genome. Finally, I placed all of these genomes into a file titled “MRSA Genomes”. The reason that I selected these genomes is because I was looking for MRSA strains and I wanted to have some diversity between all of the strains, so that not all of them would be resistant to the same things. If they all had the same resistance genes, the aim of my project would’ve needed to be changed. By randomly selecting the first 51 genomes, I was ensuring that there would be some sort of genetic diversity among all of my strains.

I then uploaded each of these genomes into PathogenWatch, using version 21.2.0. I uploaded single genome FASTAs because that was how I had downloaded and compiled the 51 genomes. Once all the genomes were uploaded into PathogenWatch, I selected all the genomes that I had downloaded and created a personal collection in PathogenWatch. Once I had done that, I viewed the collection to ensure that all of the genomes I had downloaded were in PathogenWatch. A tree was generated that included all of my genomes and to view the resistance genes, I clicked on “typing” and selected “genes”. I looked through all of the genes that were present in my strains and selected the 6 most common: *mecA, blaZ, ermA, tetM, aaca-aphD,* and *aphA-3*. These were selected because when I clicked on each gene, the number of strains with the resistance gene would light up on the tree. The 6 genes mentioned above were those with the most amount of strains conferring resistance to their respective antibiotics. Finally, I recorded those genes in a spreadsheet on Google Sheets.

After downloading these genomes and viewing them in PathogenWatch, I returned to NCBI on September 4, 2023 to collect the metadata of each of the genomes I had retrieved. I collected their metadata by clicking on the BioSample ID and noting the information presented in Table 1. The metadata I recorded was the accession number, strain, host, collection date, isolation source, and geographical location. I added these to the spreadsheet that had originally contained only my genes that I had taken from PathogenWatch.

This collection is important because it can be used to compare and analyze the different MRSA genomes, which can help us understand the evolving history of resistance and genetic diversity of the various strains. These data can also be used to develop new treatments, such as vaccines, for diseases caused by the organisms.

Next, I took all the genes I had recorded from PathogenWatch and found their accession numbers (listed in Table 1) in NCBI. I did this by searching in the nucleotide database for my gene, clicking on its CDS, and downloading its FASTA file. I then compared each gene with its different files from GenBank, EMBL, and DDJB in order to see if there were any differences between the databases. I used version 257.0 for GenBank and version 130.0 for DDBJ. I looked at both the general database information, as well as the information given in the FASTA files and recorded any differences and similarities that I saw, as well as readability. The files in GenBank were easily accessible if I just searched for the accession number of the gene I was looking for. Specifically for DDBJ, as their website was less user-friendly, I had to go to their website, click on search, and then click on ARSA. Once I had found the gene, I clicked on both the flat-file and FASTA file to be able to compare both to the formats found in GenBank and EMBL.

Gene presence and absence analysis:

Over the period of the next two weeks, beginning September 4, 2023 - September 18, 2023, I took the sequence of each gene - *mecA, ermA, aacA-aphD, aphA-3, blaZ, and tetM* - and I blasted my set of genomes with each of these genes. Table 1 includes the accession numbers of all of the genomes that I blasted, and the accession numbers of the genes that I blasted are MW682923 for *mecA*, CP003194 for *aphA-3*, CP010526 for *aacA-aphD,* MT536162 for *blaZ,* CP002120 for *ermA,* and M21136 for *tetM*. I used the program BLASTn found in the NCBI database, version 2.14.1. I blasted 3 genomes at once as the BLASTn program would crash if I added a 4th genome. This process allowed me to compare the sequence of each gene across all genomes, which can help to identify genes that are highly conserved between different organisms and those that are highly variable. This can help to identify genes that may be important for pathogenicity or antibiotic resistance and allows me to determine patterns between the homologous genes in the genomes. After the analysis was complete, I recorded my results in a spreadsheet - both the hits and the description of the BLAST, looking for patterns in my results.

Nucleotide and AA diversity:

My final step was to look at various protein structures of the genes that I have used in my analysis of gene diversity in MRSA strains. I checked to see if there were any Single Nucleotide Polymorphisms (SNPs) in the genes or any variants in the genes present in the genomes. While I was looking at the alignment of the BLAST results, only *ermA, blaZ,* and *aacA-aphD* showed some differences.

Functional changes within genes:

With this information in mind, I decided to delve a little further to understand whether these SNPs would cause any structural variants and thus functional changes. To gather more information, I went to secondary databases such as PDB, UniProt and Phobius to further examine the two variations of the gene there. I searched up the name of the gene that had shown some differences and looked at their structure on PDB. If PDB showed differences, I went to UniProt to confirm if there was a difference in the function of the protein. If there was, I would go to Phobius to look at the transmembrane topography and further confirm a difference in function. To do so, I would upload the protein sequence into Phobius and it would generate a graph with information about the protein. I then recorded whether or not the differences resulted in changed protein structure and location. I used the version UniProt release 2023_04, PDB V4.0, and Phobius 1.01 to conduct these analyses.

Results:

Metadata and Database Comparison Shows Genbank is the Best Database for Information on MRSA:

To begin, I analyzed each of my genomes and recorded their metadata using the process mentioned in my methods portion. Supplementary Tables 1 and 2 show the information that I recorded for each strain and Figures 3 and 4 show some of the comparisons of the metadata from the various strains.

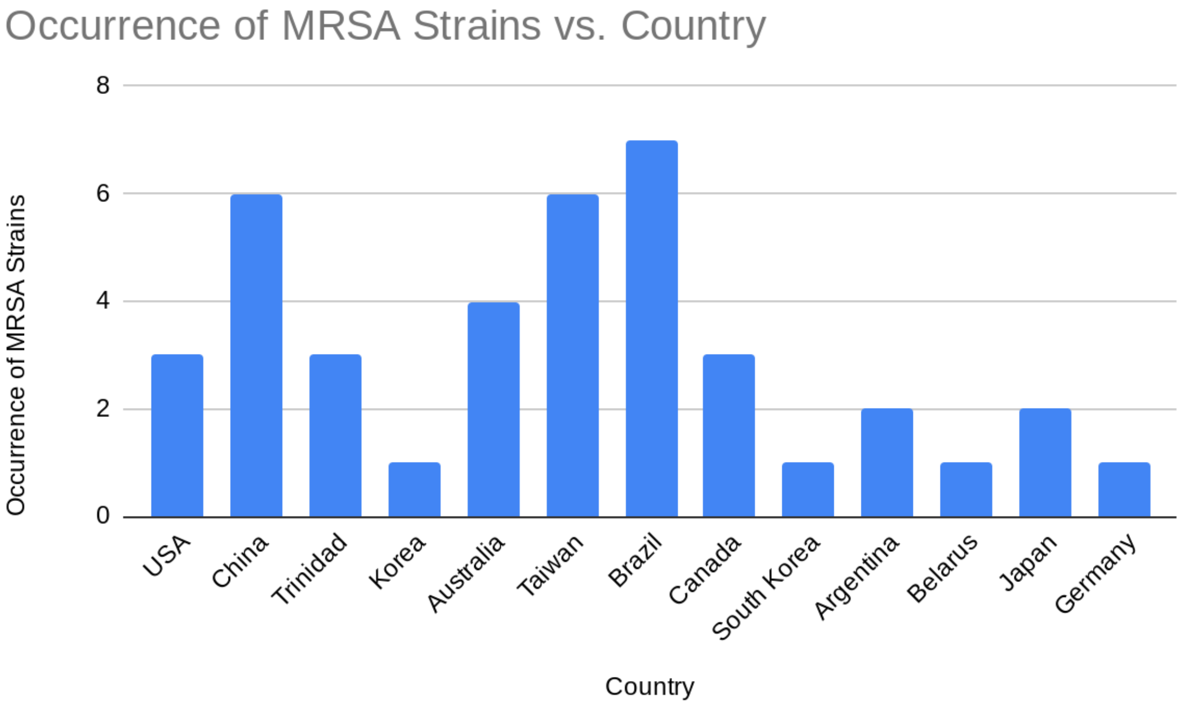


Figure 3: The above figure is a depiction of the metadata gathered from each genome. In this chart, the country where the MRSA strains were isolated was recorded. As shown in the chart, the genomes were taken from all over the world - primarily Taiwan and Brazil, both of which are humid areas which allow bacteria to grow faster.

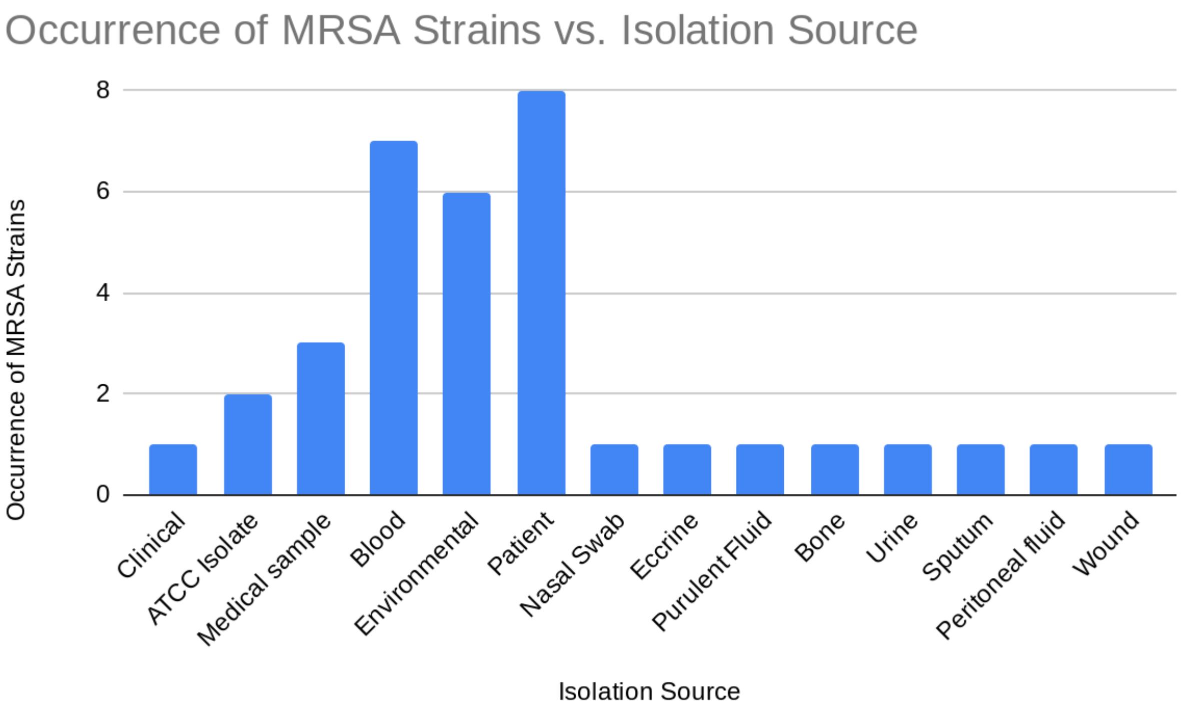
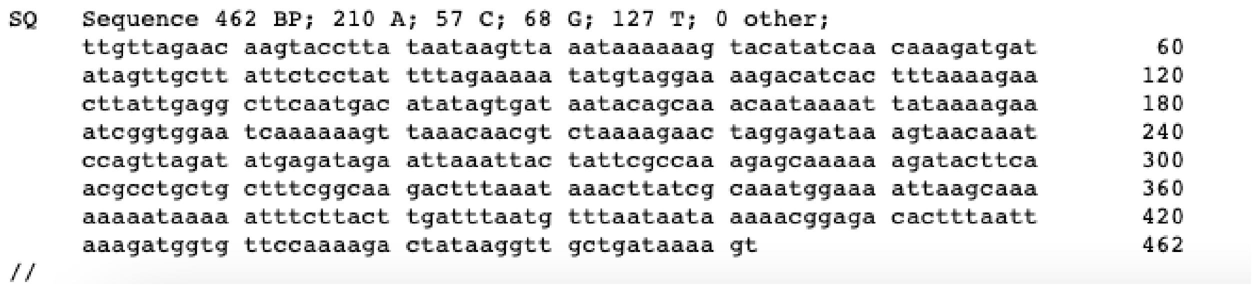


Figure 4: The above figure is another depiction of the metadata gathered from each of the 51 downloaded genomes. Many of the isolation sources were from blood or the environment. As shown in the graph, the samples were taken mainly from patients in order to be more accurate in terms of effects on the human body.

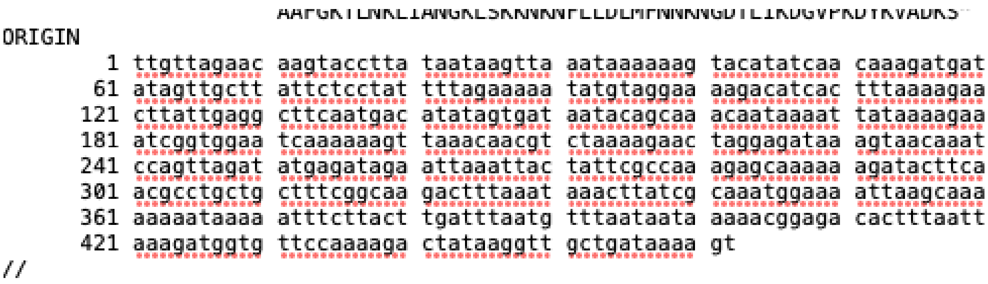
Once I had recorded all of my metadata, I turned my attention to the genes that I would eventually be blasting with. After searching through the various primary databases for each of my genes, comparing both the formats and the FASTA files, I noticed a few slight discrepancies but overall, the information given in each database was the same, as is recorded in Tables 2 and 2.1 below.

When comparing the *mecA* gene, some differences that I noticed were that the EMBL database doesn't have any links. While the way the information is formatted is different, the content is still the same. With the DDBJ database, it was virtually the exact same as the GenBank database - the only difference was that both EMBL and DDBJ included a base count in the information about the gene. The same results applied when I was comparing both *tetM* and *blaZ* across the different databases. However, when I looked at the nucleotide sequences, I found some differences between EMBL and GenBank (see Figures 3 and 4) as they seemed to record their sequences differently. I soon found out that EMBL recorded the number of the last base and GenBank recorded the number of the first base in a sequence line.



```
SQ   Sequence 462 BP; 210 A; 57 C; 68 G; 127 T; 0 other;
     ttgttagaac aagtacctta taataagtta aataaaaaag tacatatcaa caaagatgat        60
     atagttgctt attctcctat tttagaaaaa tatgtaggaa aagacatcac tttaaaagaa       120
     cttattgagg cttcaatgac atatagtgat aatacagcaa acaataaaat tataaaagaa       180
     atcggtggaa tcaaaaaagt taaacaacgt ctaaaagaac taggagataa agtaacaaat       240
     ccagttagat atgagataga attaaattac tattcgccaa agagcaaaaa agatacttca       300
     acgcctgctg ctttcggcaa gactttaaat aaacttatcg caaatggaaa attaagcaaa       360
     aaaaataaaa atttcttact tgatttaatg tttaataata aaaacggaga cactttaatt       420
     aaagatggtg ttccaaaaga ctataaggtt gctgataaaa gt                          462
//
```

Figure 5: Picture of EMBL database and the nucleotide sequence of MRSA. Notice the numbers on the right side.



```
                     AAFGKTLNKLIANGKLSKNNKNFLLDLMFNNNNGDTLIKDGVPKDYKVADKS
ORIGIN
        1 ttgttagaac aagtacctta taataagtta aataaaaaag tacatatcaa caaagatgat
       61 atagttgctt attctcctat tttagaaaaa tatgtaggaa aagacatcac tttaaaagaa
      121 cttattgagg cttcaatgac atatagtgat aatacagcaa acaataaaat tataaaagaa
      181 atcggtggaa tcaaaaaagt taaacaacgt ctaaaagaac taggagataa agtaacaaat
      241 ccagttagat atgagataga attaaattac tattcgccaa agagcaaaaa agatacttca
      301 acgcctgctg ctttcggcaa gactttaaat aaacttatcg caaatggaaa attaagcaaa
      361 aaaaataaaa atttcttact tgatttaatg tttaataata aaaacggaga cactttaatt
      421 aaagatggtg ttccaaaaga ctataaggtt gctgataaaa gt
//
```

Figure 6: Picture of the GenBank database and the nucleotide sequence of MRSA. This is the same gene used in the above photo of the EMBL database. Notice the numbers on the left side and the differences between the two pictures.

However, when comparing the FASTA files between the databases for *mecA*, I noticed that the EMBL and DDBJ databases were a little more similar in their presentation of information. They also didn't include a FASTA file for the protein sequence, unlike GenBank.

Regarding *tetM* and *blaZ*, I retrieved the same results from my *mecA* analysis. In their FASTA files, EMBL and DDBJ both mention if the gene is the complete CDS or not, while GenBank does not. My results from comparing the primary databases started to vary when I began looking at genes such as *ermA*, *aphA-3,* and *aacA-aphD.* When I took a look at the information about the *aacA-aphD* gene in the various databases, I noticed that GenBank looked like it had with the previous genes. However, EMBL had a huge section with the initials DR (Database Cross-Reference). This means that it cross-references other databases which contain information related to the entry in which the DR line appears. GenBank doesn't have this and the arrangement of cds is different, but still contains the same information. In addition, DDBJ gave the whole base count and the whole sequence of the genome; the other databases didn't do that. I received the same results when I compared the *ermA* and *aphA-3* genes. When comparing the FASTA files for these genes, I noticed that while EMBL had sufficient information about the gene on the database page itself, it had significantly less information for the FASTA files regarding the nucleotide sequence. DDBJ included the whole sequence but didn't divide it at all. Out of these formats, GenBank was a bit "nicer" as it broke up the separate genes into different sections, making it easier to locate a specific gene and the protein it creates. A database and FASTA file comparison is given below in Tables 2 and 2.1.

Database comparison	***mecA***	***tetM***	***blaZ***	***ermA***	***aphA-3***	***aacA-aphD***

GenBank	Links included in information, no base count, fully written out what the topic is.	Links included in information, no base count, fully written out what the topic is.	Links included in information, no base count, fully written out what the topic is.	Links included in information, more information given than in EMBL, fully written out what the topic is, no base count.	Links included in information, more information given than in EMBL, fully written out what the topic is, no base count.	Links included in information, more information given than in EMBL, fully written out what the topic is, no base count.
EMBL	No links in EMBL, but includes base count, date of creation and date of update. Abbreviations for certain information is hard to understand.	No links in EMBL, but includes base count, date of creation and date of update. Abbreviations for certain information is hard to understand.	No links in EMBL, but includes base count, date of creation and date of update. Abbreviations for certain information is hard to understand.	EMBL had a huge section with initials DR (Database Cross-Reference). It cross-references other databases which contain information related to the entry in which the DR line appears but GenBank and DDBJ don't have this. Arrangement of CDS is different but contains the same information.	EMBL had a huge section with initials DR (Database Cross-Reference). It cross-references other databases which contain information related to the entry in which the DR line appears but GenBank and DDBJ don't have this. Arrangement of CDS is different but contains the same information.	EMBL had a huge section with initials DR (Database Cross-Reference). It cross-references other databases which contain information related to the entry in which the DR line appears but GenBank and DDBJ don't have this. Arrangement of CDS is different but contains the same information.
DDBJ	DDBJ includes links just like Genbank. The way that it organizes and calls things is much more similar to GenBank than it is to EMBL. DDBJ includes a base count.	DDBJ includes links just like Genbank. The way that it organizes and calls things is much more similar to GenBank than it is to EMBL. DDBJ includes a base count.	DDBJ includes links just like Genbank. The way that it organizes and calls things is much more similar to GenBank than it is to EMBL. DDBJ includes a base count.	Similar to the previous databases; however, DDBJ also gave the whole base count and the whole sequence; the other databases didn't do that.	Similar to the previous databases; however, DDBJ also gave the whole base count and the whole sequence; the other databases didn't do that.	Similar to the previous databases; however, DDBJ also gave the whole base count and the whole sequence; the other databases didn't do that.

Table 2: Compares the various genes across the different primary databases: GenBank, EMBL, and DDBJ. Looked at the information in the database to compare readability, determine if there were any discrepancies, and figure out the best database to use when searching for information.

Fasta Files Comparison	***mecA***	***tetM***	***ermA***	***blaZ***	***aphA-3***	***aacA-aphD***
GenBank	Only GenBank includes a FASTA file with the protein sequence and includes protein ID in FASTA nucleotide file.	GenBank doesn't mention if the gene is complete CDS, otherwise everything is the same between the files, similar to *mecA*.	GenBank is a little nicer because there are various gene, so they break up each separate gene into different sections so it's easier to locate a specific gene and the protein that it is.	No difference in the way GenBank looks compared to the portrayal of *mecA* in FASTA files	GenBank is a little nicer because there are various gene, so they break up each separate gene into different sections so it's easier to locate a specific gene and the protein that it is.	GenBank is a little nicer because there are various gene, so they break up each separate gene into different sections so it's easier to locate a specific gene and the protein that it is.
EMBL	No FASTA file for the protein itself, otherwise GenBank and EMBL are the same.	EMBL mentions if the gene is complete CDS	EMBL has significantly less info for the FASTA files regarding nucleotides.	No FASTA file for the protein itself, otherwise GenBank and EMBL are the same.	EMBL has significantly less info for the FASTA files regarding nucleotides.	EMBL has significantly less info for the FASTA files regarding nucleotides.
DDBJ	DDBJ is very similar to the other 2 databases.	DDBJ mentions if the gene is complete CDS	DDBJ includes the whole genome, so it's quite long, but no other differences between the databases.	DDBJ is very similar to the other 2 databases.	DDBJ includes the whole genome, so it's quite long, but no other differences between the databases.	DDBJ includes the whole genome, so it's quite long, but no other differences between the databases.

Table 2.1: Compares the various genes across the different primary databases: GenBank, EMBL, and DDBJ. Looked at the FASTA files to determine if there were any discrepancies, as well as the best database to use when searching for information.

In terms of the primary databases, I discovered that when simply looking for gene information, either GenBank or DDBJ would be the most user-friendly places to go. This is because EMBL does not write out exactly what they are describing in the gene, whether it is locus, accession number, etc. This makes it a lot harder to compare any similarities or differences between the 3 primary databases I looked at. However, when looking at the FASTA files, either GenBank or EMBL would be the most ideal to look at to retrieve information. GenBank (especially for genomes) organizes information in a readable format while EMBL provides a shorter, condensed version with the same amount of information. DDBJ, on the other hand, provides a lot of information with no way to parse it down.

Discovery of 6 Vulnerable Strains and High Areas of Occurrence Through BLAST:

Once I had gone through the various databases to look at trends regarding file formats, I began to blast my genomes with the genes I had taken from PathogenWatch. When I first tried to blast, I didn't see an area for me to enter in the genomes that I wanted to blast. After some time, I realized that I had to click the button that said "align two or more sequences." Once I had done that, I assumed the BLAST would work. However, when I tried to blast 10 sequences at once, the software kept returning an error function that said, "Length limit exceeded. Please limit your query/subject sequence length to 10,000,000 characters or less. I was extremely confused because my query sequences for both genes and genomes were obviously short - they were just accession numbers. After fiddling around with the software for a few hours, I realized that it wasn't my genes or genomes that were the problem, but rather the size of my genomes. 10 genomes were too much for the BLAST to handle, so I shrunk it down to 3. Just to check, I tried to blast with 4 genomes, but the system crashed. After figuring out that I could only blast 3 genomes at a time, I began to blast all of the genomes with the mecA strain, which took me about an hour. Each time I received a result, I downloaded both the Hit Table (as a CSV) and the Description Table (as a CSV). The purpose of downloading it as a CSV was so that the files would easily be uploaded into a spreadsheet and would be readable. Once I had blasted all of my genomes against the mecA gene, I placed them all into a folder called MecA_CSV. I then ran a terminal command to concatenate all the CSVs into one combined CSV using the command: cat *.csv > concatenated.csv. I repeated this process for all of the other genes that I needed to blast, which took about 4 hours. After combining all the CSVs for each gene, I uploaded them into separate spreadsheets as shown in Supplementary Tables 3 - 8.

Supplementary Tables 3 - 8 all depict the results from the BLAST of each gene against the 50 genomes I downloaded. They include the hits that the BLAST generated as well as the description table. The description table for each gene includes the max score, total score, query cover, E value, percentage identity, accession length, and accession number. Across all of the hits, something that was always consistent was the E value for all of the hits was always 0 and the percentage identity ranged from 96% to as high as 100%.

While comparing my results, a common trend across all the strains that I blasted was that there were 6 consistent strains that didn't have a hit with any of the genes I blasted them with. These 6 were (by accession number): CP000253, CP104478, CP011526, CP035101, CP040998, CP064365. The names of the strains are NCTC 8325, DSM 20231, DSM 20231, ATCC 12600, FDAARGOS_773, and PartF-Saureus-RM8376 respectively. When I looked on PathogenWatch to compare my results from the BLASTn, I noticed that all of these strains did not have resistance to Amikacin, Gentamicin, Tobramycin, Kanamycin, Methicillin, Penicillin, Clindamycin, Erythromycin, and Tetracycline. Thus, I concluded that these 6 strains, if found in any organism, could be treated using any of these drugs, as they would be extremely vulnerable and currently, the gene isn't present in the genome.

In addition, when looking at my results, I noticed *mecA*, *tetM*, and *ermA* were present in most of the genomes while the other sequences weren't as common. *MecA* appeared 42 times, *tetM* 41, and *ermA* 39 times throughout the genomes showing that these are genes that these strains of MRSA would be far more resistant to. In addition, there were several strains in my collection that had high gene diversity, having the sequence of almost every gene I blasted against them. When I looked at my metadata, I noticed that most of these strains were from South America, in areas such as Brazil and Argentina, or they were from East Asia, near Taiwan. If you refer back to Figure 1, you will notice that these are the exact areas that often report high cases of resistant strains of *S. aureus*. Thus, the results produced from my BLAST do confirm the facts about those areas having higher cases of resistance.

Discovery of Protein Function Variation in *BlaZ* gene:

The final part of my results was to see if the same gene had variants and if those variants affected the function of the gene. While mecA, tetM, and aphA-3 did not have any variants that I found, blaZ, ermA, and aacA-aphD all had structural variants. While I was blasting my genes and looking at the alignment for them, I noticed that blaz, ermA, and aacA-aphD, had some variation in the nucleotide sequence, and so I decided to take a closer look into those 3 genes. I wanted to figure out whether or not those slight variations would end up causing a change in protein function or structure.

The first thing I did was access the Protein Data Bank (PDB) database. I used this database because it would show me the 3D structure of the protein I was looking for which would help in distinguishing any structural differences between the genomes. While looking through PDB, I noticed that the *blaZ* gene found in genomes throughout my dataset has 2 different structures attached to it. One is classified as a signaling protein and the other is classified as a hydrolase. Due to this different classification, their structures differ greatly from each. Figures 7A and 7B show that *blaZ* protein that functions as a hydrolase is less symmetrical than the signaling protein's structure. This structural difference does lead to differences in their functions (which I discovered in Phobius), because, in the case of most proteins, structure determines function. After discovering this, I went to UniProt to see if I could learn more about the function of these variants, as UniProt would show the annotation of protein sequences. In UniProt, I found that the function of the hydrolase is a "beta-lactam + H2O = a substituted beta-amino acid" (Uniprot) and its biological process is antibiotic resistance. The signaling protein has this definition: "An integral membrane protein involved in sensing of the presence of beta-lactam antibiotics and transduction of the information to the cytoplasm" (Uniprot). The definition goes on to mention that "mechanistically, activation of the signal transducer involves acylation of a serine in the C-terminal sensor domain upon binding of the beta-lactam antibiotic. In turn, a conformational change occurs and the signal is transmitted from the cell surface to the cytoplasm. There, the zinc protease domain is activated and initiates autoproteolysis as well as cleavage of the transcriptional repressor BlaI leading to derepression of antibiotic resistance genes" (UniProt).

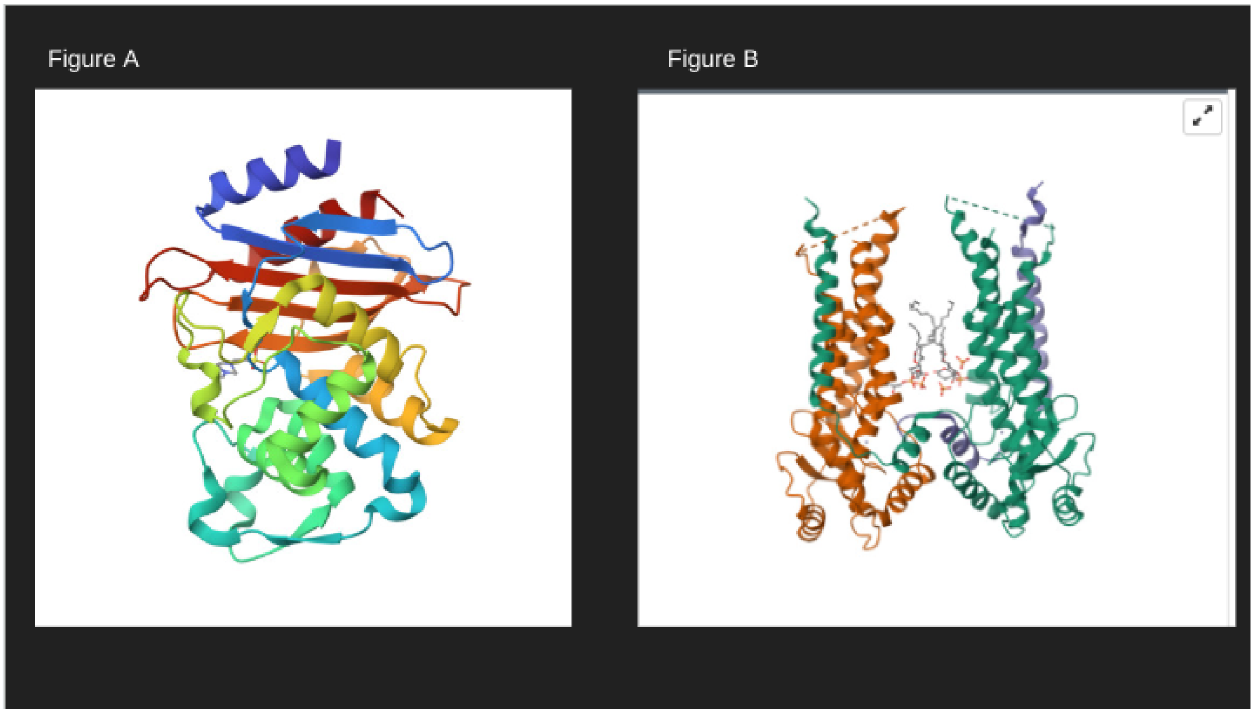
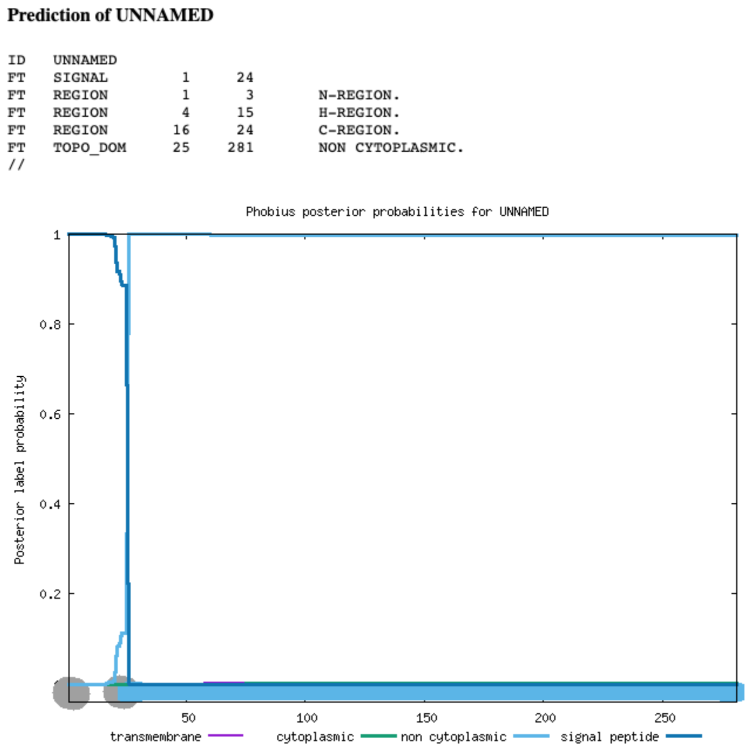


Figure 7A: This is a diagram of the hydrolase protein version of *blaZ*. Its structure is much more denatured/loosely coiled than that of the signaling peptide. **Figure 7B:** Above is a picture of the signaling protein that varies in structure compared to the hydrolase of *blaZ* in Figure 7A.

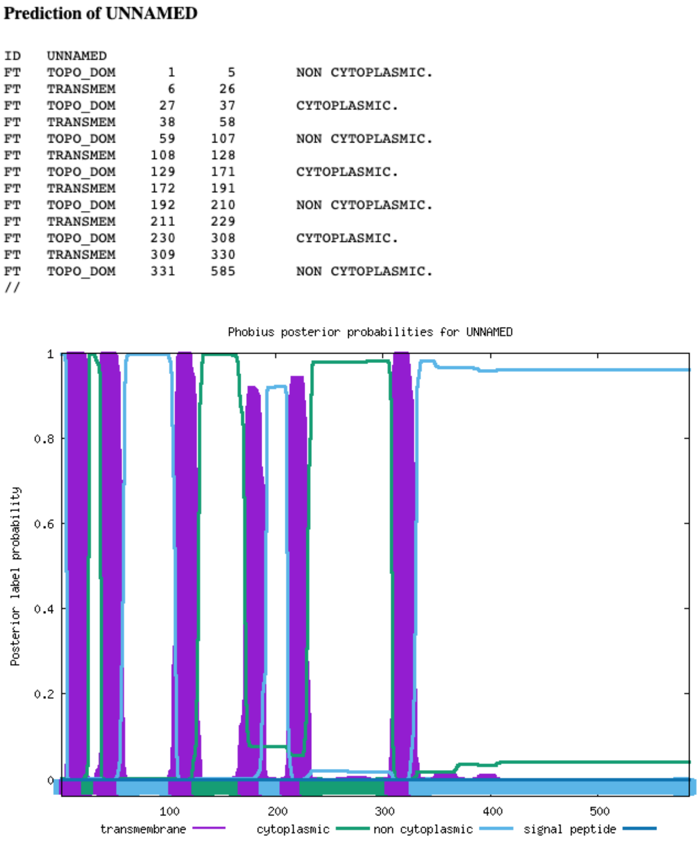
To further solidify whether the functions of these two proteins would be different, I gathered their protein sequences from NCBI by searching the accession number of the gene, finding its FASTA file, and put that sequence into Phobius, another secondary database that determines where certain proteins will remain and how they act. Phobius further proved that the difference in structures caused a change in functions because the hydrolase acted as a signaling peptide for a little bit before becoming non-cytoplasmic. The signaling peptide was often in the transmembrane, cytoplasm, acted as a signaling peptide occasionally, and more. Figures 8 and 9 depict the graphs created by Phobius.



Prediction of UNNAMED

```
ID   UNNAMED
FT   SIGNAL          1     24
FT   REGION          1      3         N-REGION.
FT   REGION          4     15         H-REGION.
FT   REGION         16     24         C-REGION.
FT   TOPO_DOM       25    281         NON CYTOPLASMIC.
//
```

Figure 8: This is a graph of the location and function of the hydrolase protein version of *blaZ*. Its function is quite straightforward - a signaling peptide for the first 24 amino acids and then non-cytoplasmic for the rest of the sequence.

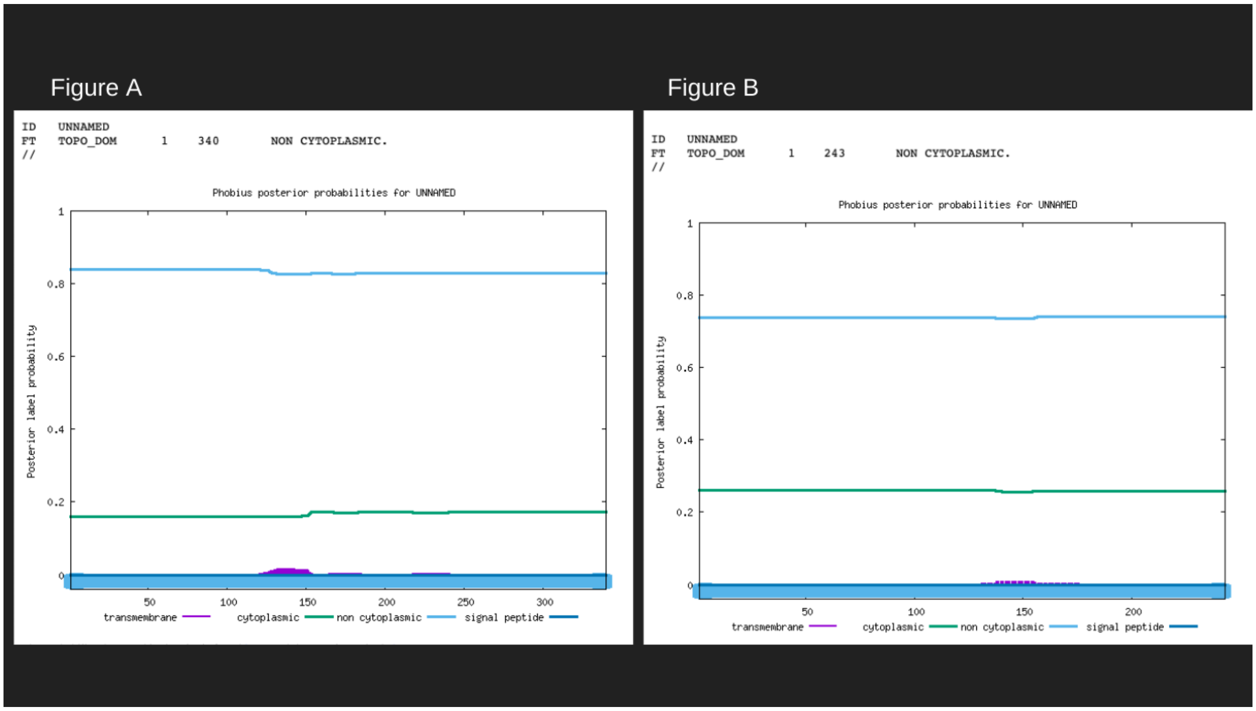


Prediction of UNNAMED

```
ID   UNNAMED
FT   TOPO_DOM      1     5       NON CYTOPLASMIC.
FT   TRANSMEM      6    26
FT   TOPO_DOM     27    37       CYTOPLASMIC.
FT   TRANSMEM     38    58
FT   TOPO_DOM     59   107       NON CYTOPLASMIC.
FT   TRANSMEM    108   128
FT   TOPO_DOM    129   171       CYTOPLASMIC.
FT   TRANSMEM    172   191
FT   TOPO_DOM    192   210       NON CYTOPLASMIC.
FT   TRANSMEM    211   229
FT   TOPO_DOM    230   308       CYTOPLASMIC.
FT   TRANSMEM    309   330
FT   TOPO_DOM    331   585       NON CYTOPLASMIC.
//
```

Figure 9: This is a graph of the signaling protein version of *blaZ*. Its function and location is a lot more varied than that of the hydrolase, with this protein being present nearly everywhere in the cell at different points in the sequence.

After looking at *blaZ*, I moved on to *ermA* and found variants of this gene as well. It also had 2 different structures - they looked a little similar in PDB, so I went to UniProt to see if they had different functions. While both variants were involved in erythromycin resistance and were classified as transferase, the function of the protein *ermA* in *S. aureus* is more detailed: “This protein produces a dimethylation of the adenine residue at position 2085 in 23S rRNA, resulting in reduced affinity between ribosomes and macrolide-lincosamide-streptogramin B antibiotics.” (UniProt). I went to Phobius once more to see if the function was different between the two variants, but they both were non-cytoplasmic. However, as you can see in Figures 10A and 10B, areas where the protein may be in the transmembrane do vary.



```
ID   UNNAMED
FT   TOPO_DOM      1    340       NON CYTOPLASMIC.
//
```

```
ID   UNNAMED
FT   TOPO_DOM      1    243       NON CYTOPLASMIC.
//
```

Figure 10A: This is a diagram of the protein *ermA* whose definition is to aid in erythromycin resistance. Its location seems to be mainly non-cytoplasmic, indicating activity outside of the cell membrane. **Figure 10B:** This is a diagram of the protein *ermA* which produces a dimethylation of the adenine residue. Its location is relatively similar to that of the previous protein, with both of them being non-cytoplasmic proteins.

Finally, I looked at the protein for *aacA-aphD*. While it also had 2 different structures, they were both classified as transferase proteins, and just like the *ermA* protein, they were both non-cytoplasmic with no major differences in function despite their structural difference.

When looking at the secondary databases for the proteins of *blaZ, ermA*, and *aacA-aphD*, I noticed that despite differing structures in the latter two, there wasn't really a big difference in their function. This could be due to the fact that scientists are still not sure what these proteins look like, so slight variations might just be different depictions of the same protein. However, the *blaZ* protein has a confirmed structure - a helix with beta strands - and therefore, with a significant difference in shape and structure, its function also differs significantly. The Phobius diagram depicts the two proteins as having different functions and locations, depending on what they do. It is therefore possible that the two proteins have different functions, even though they are both part of the same family. This suggests that they have different mechanisms of action, which could explain the different effects of the two proteins. Ultimately, only *blaZ* had a change in function due to a change in protein structure unlike the other genes, which can lead to the conclusion that not all structural changes in proteins result in functional differences. This suggests that other mechanisms might be at play to cause functional changes in proteins. It highlights the importance of studying the interactions between proteins and their ligands in order to gain a better understanding of their functions.

Discussion:

While looking at the databases, I noticed that for this particular project regarding MRSA, GenBank and/or DDBJ (especially GenBank) were the easiest to navigate and presented the information in such a way that it was easy for me to understand it (Benson *et al.,* 2010). However, this specific project required a lot of metadata which is why those databases were the easiest for me to use. For people who are doing projects that require them to find information from other databases, EMBL might be the most helpful for them. While looking at my genes, I noticed that EMBL had a large section with initials DR (Database Cross-Reference). This meant that it cross-references other databases which contain information related to the entry in which the DR line appears (Baker *et al.,* 2000). GenBank and DDBJ don't possess this characteristic which therefore means that it would be much harder to extract that kind of information from one of those databases. Ultimately, it is useful to know the differences between various databases because knowing what each database offers can help a researcher recognize which would be the most useful database for their project (Grewal *et al.,* 2016). In addition, using multiple databases and knowing their similarities and differences allows researchers to know what to look for, especially if they are trying to spot any discrepancies or additional information that may be necessary for their project.

While I was searching through my collection of genomes, I noticed that the *mecA* gene didn't have any SNPs in it. I perused the current literature and found that this was a little unusual. Some researchers have found that the *mecA* gene does have some SNPs that are involved with resistance to other antibiotics as well (Salehi *et al.*, 2020). Therefore, what I discovered in my collection was relatively uncommon as sometimes SNPs in *mecA* have even led to mediated beta-lactam resistance in some *S. aureus* (Rolo *et al.,* 2017). This likely may have happened due to the fact that 51 genomes is a small sample size, so a large sample may have yielded some SNPs in the *mecA* gene. This would affect the treatment of patients with MRSA because if these strains had SNPs, researchers could compare the genetic make-up of an individual and an antibiotic in order to provide them with treatment that is effective and safer (Okwu *et al.,* 2019). With SNPs, it would be easier to determine an individual's risk of contracting various illnesses and well as predict their responses to drugs (Bin Alwi, 2005).

In addition, throughout the blasting process, I noticed that there were 6 genomes that had absolutely no resistance genes. To confirm this, I checked on PathogenWatch and those 6 strains did not show resistance to any type of gene. This meant that they were extremely vulnerable to any antibiotic and would die the moment antibodies touched them. However, if they are so vulnerable, why do the strains exist in the first place and how can they survive? There are several possible explanations to this question; however, I will just name a few. A possible explanation to this question is that genomes that carry more resistance genes tend to be less fit than those that carry less or none (Vogwill *et al.,* 2017). It could also potentially mean that the genomes have a few plasmids floating around them (Ternent *et al.,* 2015). When an antibiotic enters the body, the genomes might "pick up" those plasmids to be resistant for a short time and survive (Landecker, 2015). Once the antibiotic has been purged from the body, the genomes might drop the plasmids once more and become vulnerable but also more fit (Inoue, 2007). Another possible explanation is that the gene may be dormant, meaning that it only activates under certain circumstances (Courtot *et al.,* 2018). Another possibility is that these genes are not actually resistant to antibodies, but rather have some other form of defense to protect them when the antibodies do enter the body to kill the bacteria. Further research is needed to understand these genes and how they survive.

This study, however, does have potential limitations. Some limitations throughout this whole research project were that we only tested 51 different strains of MRSA. Perhaps if we tested more, our results would be more accurate, but 51 genomes is a bit on the smaller sample size side to apply results to a whole population (Nayak, 2010). There were also time constraints which may have made it harder to dive deeper into the literature and see what is already present. This would've affected our ability to present deeper thoughts on this matter. Finally, the whole results analysis may have to leave room for some error as computer programs may have slightly erred while retrieving results for the BLAST, protein structure, etc (Yu *et al.*, 2006). This could lead to some inaccurate results, even though the program was carefully tested. Therefore, it is important to note that the results presented may not be 100 percent accurate. While the protein structure databases are being updated constantly with new information, the BLAST program may sometimes vary in the homology that it presents between the gene and genome (Schaffer *et al.*, 2001).

Conclusion:

In conclusion, MRSA is a contagious disease that can lead to serious infections if not treated earlier. The people who are at risk for contracting MRSA are the elderly, athletes, and those who are in areas or activities that involve crowding, skin-to-skin contact, and shared equipment or supplies. The countries that report the most cases of MRSA tend to be those where hospitals overuse the amount of antibiotics needed to treat the disease. This leads to the bacteria developing antibiotic resistance against various types of medication such as penicillin and methicillin (which is currently where MRSA gets its name from). To conduct my research regarding gene diversity in strains of MRSA, I took multiple genomes of MRSA and blasted different genes such as *mecA* and *ermA* to see the gene diversity that various strains of MRSA had. From this process, I concluded that most strains of MRSA contain resistance genes to *mecA, tetM,* and *ermA*, but the strains do not carry only one resistance gene - rather, they carry many, inferring the possibility of multidrug resistance. In the future, the author hopes that this information will be useful when treating various strains of MRSA - doctors, if they are able to identify the strain of MRSA in a patient, will be able to treat them accordingly with the proper antibiotics that will have a positive effect on eradicating the disease from their body.

References

Alwi, Z. B. (2005). The Use of SNPs in Pharmacogenomics Studies. *The Malaysian Journal of Medical Sciences : MJMS*, *12*(2), 4–12. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3349395/

Baker, W. (2000). The EMBL Nucleotide Sequence Database. *Nucleic Acids Research*, *28*(1), 19–23. https://doi.org/10.1093/nar/28.1.19

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2009). GenBank. *Nucleic Acids Research*, *38*(suppl_1), D46–D51. https://doi.org/10.1093/nar/gkp1024

Choo, E. J., & Chambers, H. F. (2016). Treatment of Methicillin-Resistant Staphylococcus aureus Bacteremia. *Infection & Chemotherapy*, *48*(4), 267. https://doi.org/10.3947/ic.2016.48.4.267

Doherty, N., Trzcinski, K., Pickerill, P., Zawadzki, P., & Dowson, C. G. (2000). Genetic Diversity of the *tet* (M) Gene in Tetracycline-Resistant Clonal Lineages of *Streptococcus pneumoniae*. *Antimicrobial Agents and Chemotherapy*, *44*(11), 2979–2984. https://doi.org/10.1128/aac.44.11.2979-2984.2000

Grewal, A., Kataria, H., & Dhawan, I. (2016). Literature search for research planning and identification of research problem. *Indian Journal of Anaesthesia*, *60*(9), 635–639. NCBI. https://doi.org/10.4103/0019-5049.190618

Guo, Y., Song, G., Sun, M., Wang, J., & Wang, Y. (2020). Prevalence and Therapies of Antibiotic-Resistance in Staphylococcus aureus. *Frontiers in Cellular and Infection Microbiology*, *10*(107). https://doi.org/10.3389/fcimb.2020.00107

Inoue, Y. (1997). Spontaneous loss of antibiotic-resistant plasmids transferred to Escherichia coli in experimental chronic bladder infection. *International Journal of Urology: Official Journal of the Japanese Urological Association*, *4*(3), 285–288. https://doi.org/10.1111/j.1442-2042.1997.tb00191.x

Klevens, R. M. (2007). Invasive Methicillin-Resistant Staphylococcus aureus Infections in the United States. *JAMA*, *298*(15), 1763. https://doi.org/10.1001/jama.298.15.1763

Landecker, H. (2016). Antibiotic Resistance and the Biology of History. *Body & Society*, *22*(4), 19–52. https://doi.org/10.1177/1357034x14561341

Lee, B. Y., Bartsch, S. M., Wong, K. F., Singh, A., Avery, T. R., Kim, D. S., Brown, S. T., Murphy, C. R., Yilmaz, S. L., Potter, M. A., & Huang, S. S. (2013). The Importance of Nursing Homes in the Spread of Methicillin-resistant Staphylococcus aureus (MRSA) Among Hospitals. *Medical Care*, *51*(3), 205–215. https://doi.org/10.1097/mlr.0b013e3182836dc2

Lilas Courtot, Hoffmann, J.-S., & Valérie Bergoglio. (2018). The Protective Role of Dormant Origins in Response to Replicative Stress. *International Journal of Molecular Sciences*, *19*(11), 3569–3569. https://doi.org/10.3390/ijms19113569

Muenks, C. E., Sewell, W. C., Hogan, P. G., Thompson, R. M., Ross, D. G., Wang, J. W., Morelli, J. J., Gehlert, S. J., & Fritz, S. A. (2018). Methicillin-Resistant Staphylococcus aureus : The Effects Are More Than Skin Deep. *The Journal of Pediatrics*, *199*, 158–165. https://doi.org/10.1016/j.jpeds.2018.04.002

Nayak, B. K. (2010). Understanding the relevance of sample size calculation. *Indian Journal of Ophthalmology*, *58*(6), 469. https://doi.org/10.4103/0301-4738.71673

Okwu, M. U., Olley, M., Akpoka, A. O., & Izevbuwa, O. E. (2019). Methicillin-resistant *Staphylococcus aureus* (MRSA) and anti-MRSA activities of extracts of some medicinal plants: A brief review. *AIMS Microbiology*, *5*(2), 117–137. https://doi.org/10.3934/microbiol.2019.2.117

Rağbetli, C., Parlak, M., Bayram, Y., Guducuoglu, H., & Ceylan, N. (2016). Evaluation of Antimicrobial Resistance inStaphylococcus aureusIsolates by Years. *Interdisciplinary Perspectives on Infectious Diseases*, *2016*, 1–4. https://doi.org/10.1155/2016/9171395

Rolo, J., Worning, P., Boye Nielsen, J., Sobral, R., Bowden, R., Bouchami, O., Damborg, P., Guardabassi, L., Perreten, V., Westh, H., Tomasz, A., de Lencastre, H., & Miragaia, M. (2017). Evidence for the evolutionary steps leading to mecA-mediated β-lactam resistance in staphylococci. *PLOS Genetics*, *13*(4), e1006674. https://doi.org/10.1371/journal.pgen.1006674

Salehi, M., S Abdolhamid Angaji, Nader Mosavari, & Mahsa Ahrabi. (2020). SNP Scanning in mecA Gene for Methicillin-Resistant Staphylococcus aureus. *PubMed*, *18*(3), e2242–e2242. https://doi.org/10.30498/ijb.2020.2242

Schaffer, A. A. (2001). Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research*, *29*(14), 2994–3005. https://doi.org/10.1093/nar/29.14.2994

Ternent, L., Dyson, R. J., Krachler, A.-M., & Jabbari, S. (2015). Bacterial fitness shapes the population dynamics of antibiotic-resistant and -susceptible bacteria in a model of combined antibiotic and anti-virulence treatment. *Journal of Theoretical Biology*, *372*, 1–11. https://doi.org/10.1016/j.jtbi.2015.02.011

Ventola, C. L. (2015). The Antibiotic Resistance crisis: Part 1: Causes and Threats. *P & T : A Peer-Reviewed Journal for Formulary Management*, *40*(4), 277–283. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4378521/

Vogwill, T., & MacLean, R. C. (2014). The genetic basis of the fitness costs of antimicrobial resistance: a meta-analysis approach. *Evolutionary Applications*, *8*(3), 284–295. https://doi.org/10.1111/eva.12202

Weber, K. (2009). Community-Associated Methicillin-Resistant Staphylococcus aureus Infections in the Athlete. *Sports Health: A Multidisciplinary Approach*, *1*(5), 405–410. https://doi.org/10.1177/1941738109343653

Yu, Y.-K., Gertz, E. M., Agarwala, R., Schäffer, A. A., & Altschul, S. F. (2006). Retrieval accuracy, statistical significance and compositional similarity in protein sequence database searches. *Nucleic Acids Research*, *34*(20), 5966–5973. https://doi.org/10.1093/nar/gkl731

Yuen, J., Chung, T., & Loke, A. (2015). Methicillin-Resistant Staphylococcus aureus (MRSA) Contamination in Bedside Surfaces of a Hospital Ward and the Potential Effectiveness of Enhanced Disinfection with an Antimicrobial Polymer Surfactant. *International Journal of Environmental Research and Public Health*, *12*(3), 3026–3041. https://doi.org/10.3390/ijerph120303026

Zhen, X., Lundborg, C. S., Zhang, M., Sun, X., Li, Y., Hu, X., Gu, S., Gu, Y., Wei, J., & Dong, H. (2020). Clinical and economic impact of methicillin-resistant Staphylococcus aureus : a multicentre study in China. *Scientific Reports*, *10*(1), 1–8. https://doi.org/10.1038/s41598-020-60825-6