CHAPTER THREE

RESULTS

OVERVIEW

This chapter is sub-divided into two sections. The first section describes the results of quality control analyses, phylogenetic investigation as well as molecular characterization of Kenya isolates. The second section highlight the phylogenetic relationship of Kenya isolates in combination with 126 public genomes and correlation of certain genes with phylogenetic lineages.

SECTION A

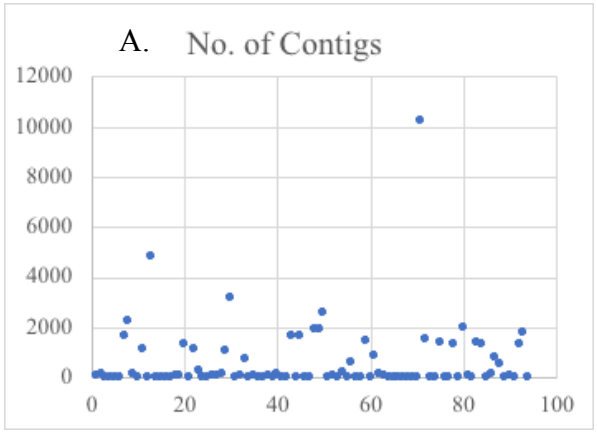
3.1.1 Study Population

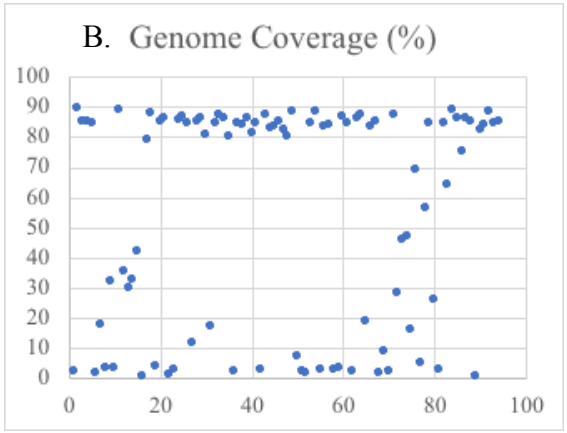
Ninety-four strains that were successfully sequenced, were randomly selected from the collections of samples from a study undertaken in Kiambu county, Kenya between October 2015 and August 2016 (unpublished data). These 94 sample isolates were obtained from nasal swabs of 46 humans and 48 swine in 29 different homesteads. Only up to a maximum of 3 farmers and 3 pigs from an average of 11 herd size (range 2-125 pigs per farm) were selected for sampling in the same homestead. All the pigs sampled were reared under zero grazing on a mean average of 2 acres piece of land with exceptions of 8 swine isolates that were grazing freely.

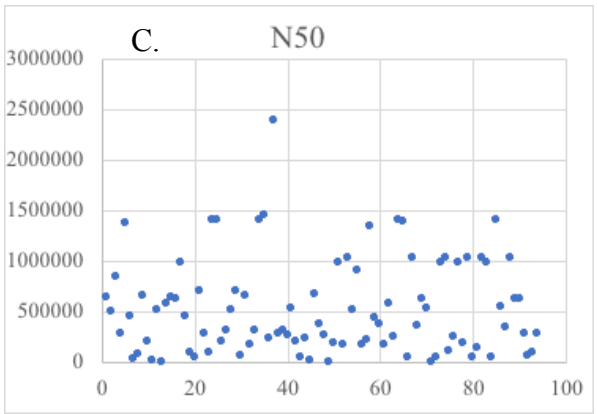
3.1.2 Genomic Metrics

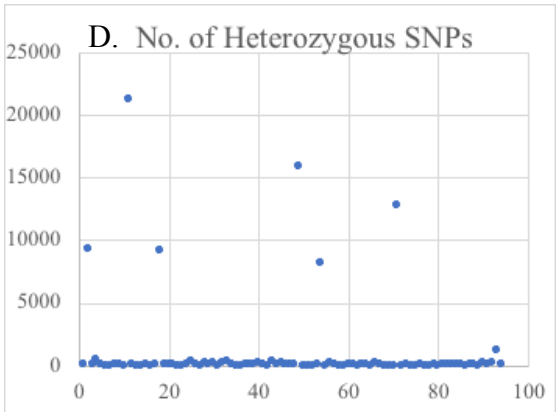
To ensure that genomes used for downstream analyses were of good quality, 94 sequence data were initially evaluated by plotting comparison graphs (as shown in fig 3) using different indicators of contamination such as number of contigs, N50, assembly length, genome coverage and number of heterozygous Single Nucleotides Polymorphisms (SNPs). Then rationale parameters (see table 7) were set for stepwise filtering of the suspected contaminant sequence strains. First, 60 isolates were eliminated because of low genome coverage on mapping to TW20 reference genome FN433596 and were either mixed with other bacterial species or different species as shown in Fig. 4. Next step was exclusions of

four isolates which had assembly length out of this range (2.6 to 3.6 Mbs). These genomic size thresholds were set based on estimation reported in the genome assembly and annotation of *S. aureus* in the NCBI (National Center for Biotechnology Information) (https://www.ncbi.nlm.nih.gov/genome/genomes/154?). Among 30 remaining isolates, 4 were suspected to be contaminated because they had larger number of contigs (544-857) than set maximum (400). The 26 genomes were further screened based on the number of heterozygous SNPs (maximum n=1000) in which 3 genomes were excluded because this could suggest the genomes might have been mixed with other *S. aureus* strains. To rule out that these SNPs were arising possibly from prophages, we checked manually their distribution by visualizing their BAM files with ARTEMIS. As a consequence of applying the above thresholds, genomes with low depth coverage, large number of unmapped reads, high number of coding genes were also eliminated resulting in 23 genomes for downstream analyses and 71 isolates being excluded for further analyses.









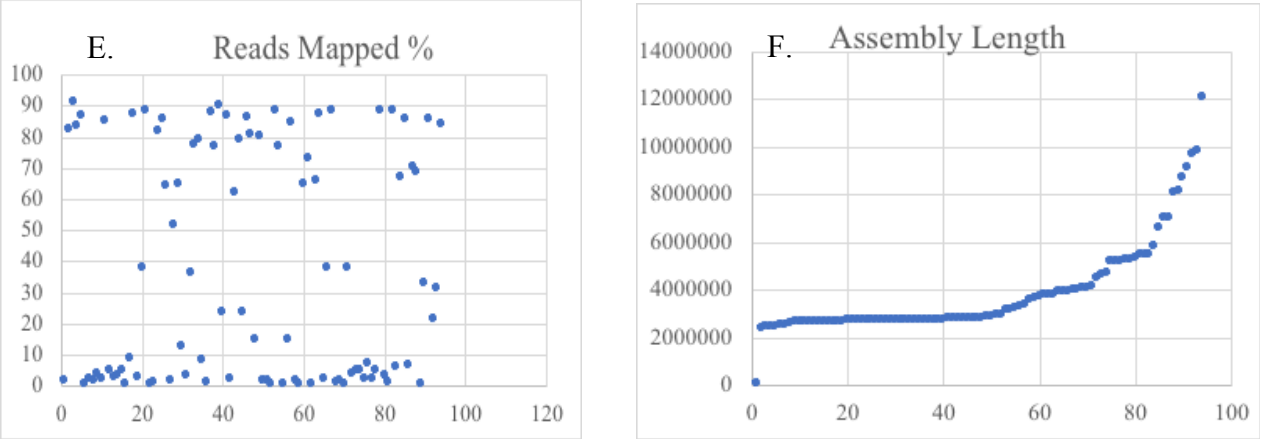


Fig. 3: Graphs showing distribution of 94 genomes based on number of contigs (A), genome coverage (B), N50 (C), number of heterozygous SNPs (D), Reads mapped % (E) and genome size length (F) before the QC filtering parameters were applied.

Table 7: Quality control filtering parameters

Filtering Parameter	**Threshold**	
	Minimum	**Maximum**
Genome size length	2.4 Mbs	3.6 Mbs
Number of contigs	N/A	400
Genome coverage compared to reference (%)	70	100
Number of Heterozygous SNPs	N/A	1000

The 23 *S. aureus* isolates have the following genomic metric features

Table 8: Genomic metrics of 23 high-quality genomes

Parameter	**Minimum**	**Maximum**	**Median**
Genome size length (bp)	2674035	2831812	2739599
Number of contigs	13	106	17
% Genome coverage to reference	85.32	96.05	93.94
Number of Heterozygous SNPs	39	401	115
Number of CDS genes	2420	2602	2518
Number of Genes	2549	2738	2648

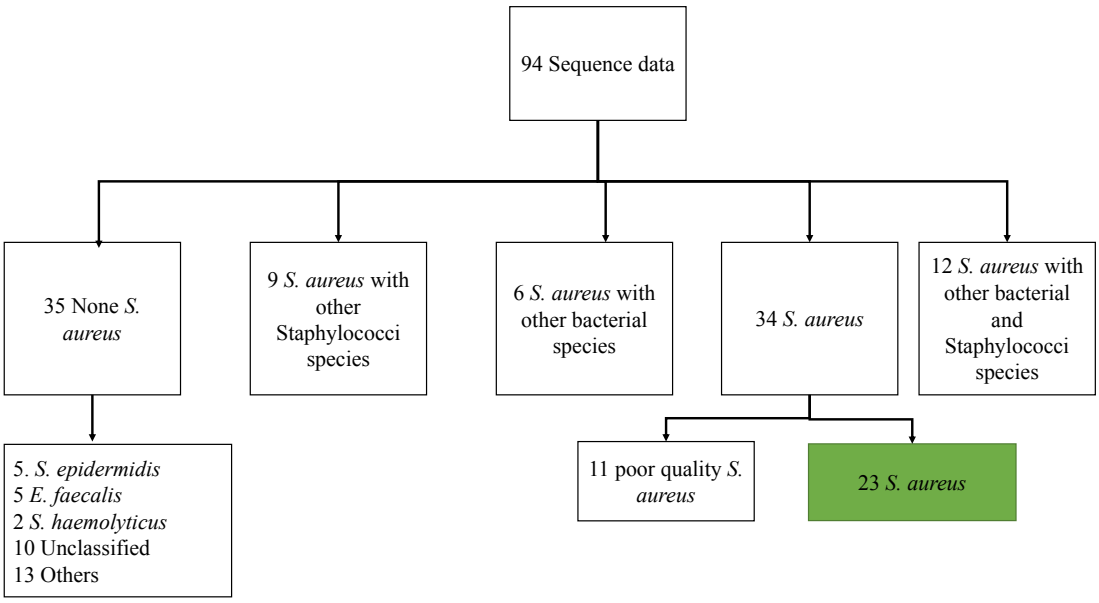


Fig. 4: A schematic diagram of a summary of general flow description of species identification and contamination of sequence data

Among 23 *S. aureus* high-quality genomes, 11 were sampled from swine and 12 from humans. Of these, 12 strains were paired in the homesteads (table 9). These paired isolates from the same homestead were sampled at the same time. The remaining 11 isolates (6 from swine and 5 from human) are un-mated and are distributed in 11 different farms. The average age of the pigs were 9 months while that of farmers were 25 years (range, 12 to 43 years). The range of herd size of pigs per farm were between 6 and 125 and were carried out under zero grazing (except in two homesteads that were grazing freely) in an approximately 2 acres piece of land. Only 3 farmers and none of the pigs had received antibiotics within the last 3 months prior to their sampling date (supplementary file 2, provide information of the farmers and pigs).

Table 9: showing pairing of strains in the homesteads (No. indicate number of strains while id. means serial identity of the homestead)

Host pairing	No. of strains	Homestead id.
Swine – human	6	24, 6, 33
Human – human	4	5, 13
Swine – swine	2	21

3.1.3 *In Silico* Prediction of Multi-locus Sequence Types and *spa* typing

To determine Multi-locus Sequence Types (MLST) of the strains, MLSTcheck script was used to compare sequence assemblies of these genomes against alleles of seven housekeeping genes in PubMLST reference database. The analyses of sequence types (ST) revealed a significant degree of genetic diversity in terms of clonal lineages in Kiambu county, Kenya. Among the 9 STs (Fig. 5A) identified in the 23 isolates, the dominant clone was ST188 26% (n=6) the predominant clone of livestock and hospitals in Pan-Asia continent (Wang et al., 2018). This was followed in frequency by ST789, ST580 and ST25 with 3 isolates (13%) in each ST. The ST188 is a double locus variant (DLV) of ST1, ST789 is a single locus variant of ST7 while ST580 is a DLV of livestock-associated ST398, a dominant clone in Europe and North America. These dominant population clones were shared among human and swine isolates (Fig. 5B).

In order to find out the distribution of Protein A (*spa*) types among the hosts, an in-house script was used to extract *spa* nucleotides from the annotated assemblies for determination of repeat units and types using Ridom and Kreiswith nomenclatures. However, the prediction of the *spa* type based on assemblies may not be accurate since the assigning of *spa* type number depend on tandem sequence repeats of polymorphic X region which could have been affected by assembly and sequencing errors. Nine *spa* types were predicted among 23 strains and were distributed among swine and human (fig 6) as defined in MLST results. The four 60% most common *spa* type of isolates were t189 (n=6), t7817 (n=3), t1176 (n=3) and t091(n=3). Overall, *spa* type and MLST results could be indicating higher diversity of clonal lineages circulating in Kiambu and their presence in both human and swine suggested possibility of these clonal lineages colonizing different types of hosts.

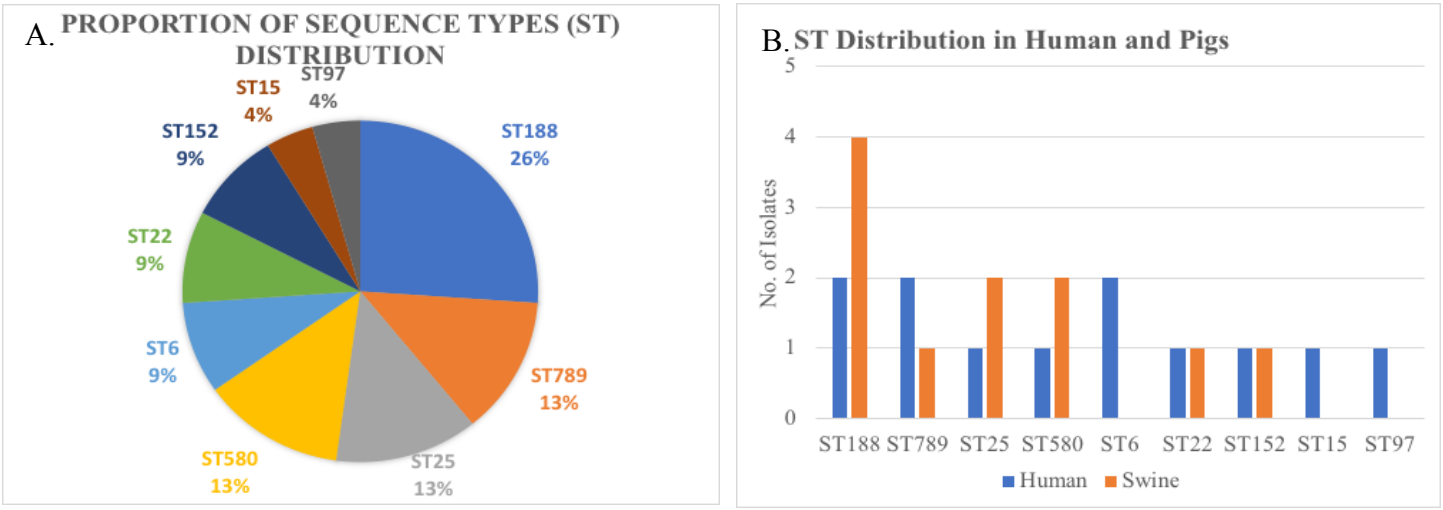


Fig. 5: Prevalence of STs (A) and distribution of STs per host (B)

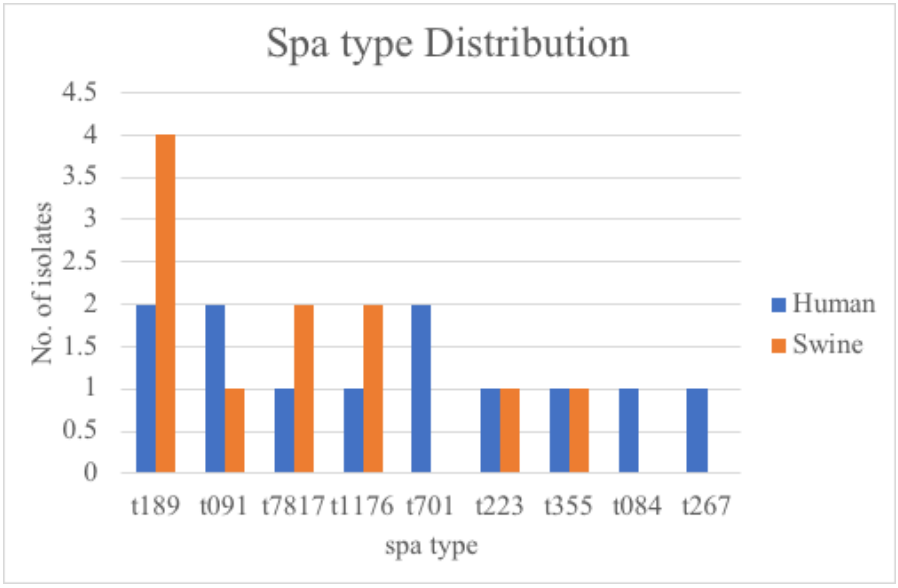


Fig. 6: *spa* type distribution between humans and swine.

3.1.4 Phylogenetic Analyses

In order to reconstruct phylogenetic tree for genetic diversity investigation between swine and humans in this study, an in-house pipeline (developed by Pathogens Informatics) was first used to retrieve a pseudogenome alignment of 2.82 Mb in length (without reference) that was generated by mapping raw reads to MSSA476 reference (ST1 by MLST) and subsequently calling SNPs. The 20 regions (185 kb size) of mobile genetic elements of the reference such as indels, prophages, transposons and other repetitive elements were removed from the alignment. This resulted in a concatenated core genome alignment of 2.63 Mb size that was used for SNPs extractions. The comparison of this concatenated core alignment of all 23 *S. aureus* strains identified 57010 variants that were exclusively ATCG, further supporting higher genetic diversity between clonal lineages of the 23 isolates. The phylogeny formed distinct clades according to ST and *spa* type designation with humans and swine isolates within and across homesteads intermingling together in the tree.

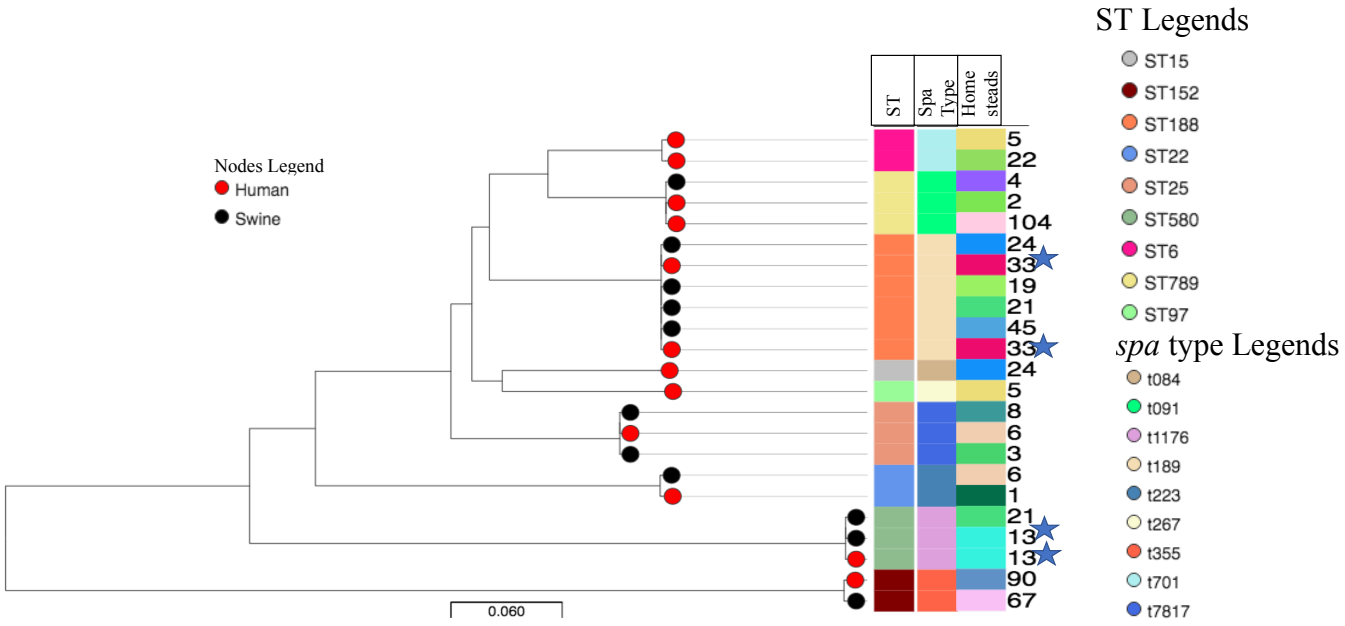


Fig. 7: Midpoint rooted maximum likelihood phylogenetic tree based on 57010 core genome SNPS. The source (human and swine) in the phylogeny nodes are linked to ST and spa type and homesteads number

Four pairs of strains (fig 7) sampled from the same homestead on the same date (i.e., homestead number 5 with human-human, 6 with swine-human, 21 with swine-swine and 24 with swine-human), belonged to different STs and had large number of SNPs differences between 8796 and 21793. On the other hand, two pairs of strains swine/human and human/human collected in homesteads 33 and 13 (blue star in fig 7), respectively, belonged to the same ST and had near identical genotypes with 2 and 6 SNPs differences, respectively. The genomes of humans and pigs within clonal lineages ST789, ST25 and ST188 were highly similar, with less than 5 SNPs differences, even though isolates were sampled from different homesteads (exact GPS locations of homesteads and date of sampling not provided).

The typing methods such as *spa* typing, are routinely used to infer transmission of *S. aureus* strains between humans and animals for epidemiological surveillance studies (Harris et al., 2010b). However, strains within each cluster of *spa* type t701, t223 and t355 had SNPs difference of 222, 86 and 68 respectively. Using less than 50 SNPs difference as a cut-off threshold for defining suspected recent sharing of common ancestor as determined by Coll et al. in longitudinal surveillance studies of MRSA in UK (Coll et al., 2017) indicate these isolates were distantly related.

3.1.5 *In Vitro* Antibiotic Susceptibility Test Results

The *in vitro* antibiotic susceptibility distribution profiles of the 23 strains were examined using Kirby – Bauer Disk Diffusion and their zone diameter breakpoints were interpreted with Clinical & Laboratory Institute Reference guidelines (supplementary file 3). Both isolates of humans and swine showed high prevalence of ampicillin (n=21, 91%), followed by erythromycin 40% n=9. Linezolid and chloramphenicol resistance were only detected in humans while one strain of swine exhibited non-susceptibility to ciprofloxacin. Only 3 strains were phenotypically resistance to cefoxitin (MRSA) (fig 8).

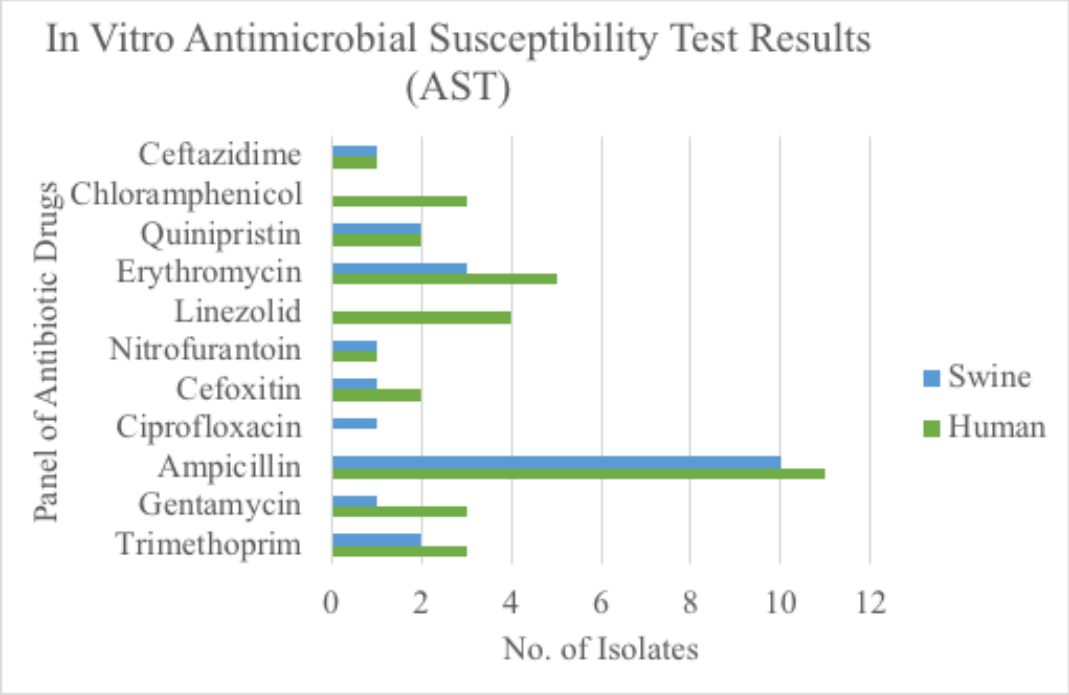


Fig. 8: Histogram showing phenotypic distribution of antibiotic drugs between pigs and humans

3.1.6 *In Silico* Prediction of Antimicrobial Resistance Genes

To investigate antimicrobial resistance (AMR) genes among the 23 WGS raw reads, reference databases of public resource Resfinder (Zankari et al., 2012) and Argannot (Gupta et al., 2014) and custom databases of known variants regions associated with resistance to fusidic acid, fluoroquinolones, vancomycin, rifampin, daptomycin and muciprocin were determined with ARIBA.

The resistance genes that were identified in 23 strains encode for aminoglycosides (*aadC*, *aac6_le_APH*), tetracyclines (*tetM*, *tetK* and *tetL*), β-lactams (*mecA* and *blaZ*), fosfomycin (*fosA* and *fosB*), trimethoprim-sulfamethoxazole (*dfrG*, *dfrC*, and *dfrK* variant) and macrolide-lincosamides-streptogramins (*mphD*, *ermC*, *lnuA*, *str*) (as shown in fig 9).

Generally, there was low level prevalence of AMR. For instance, known variants associated with resistant to ciprofloxacin, muciprocin, fusidic acid, daptomycin, vancomycin and rifampin were not detected in these isolates. However, novel variants and frameshift changes as a result of deletion were detected in *gyrA* at T818 in ST580 and ST789 strains (Supplementary File 4).

The resistance gene *blaZ* which encodes for penicillin resistance was the most prevalent (22/23, 95%). Genes that were significantly identified in higher frequency are those which encode fosfomycin, trimethoprim and macrolide each with 4 isolates (17%) represented by *fosB*, *dfrG* and *ermC* genes respectively. Only one strain carried *mecA* gene that confer resistance to methicillin. Two ST789 isolates (one swine and human) were found to be multidrug resistant and carried *blaZ*, *tetL*, *str* and *lsaa* genes, and one human ST25 isolate that had blaZ, dfrG, fosB, ermC and tetK genes. There was no significant difference in the distribution of classes of antimicrobial resistance genes among the swine and human isolates (Fishers Chi^2 exact test ($p=0.1845$) p-value < 0.010).

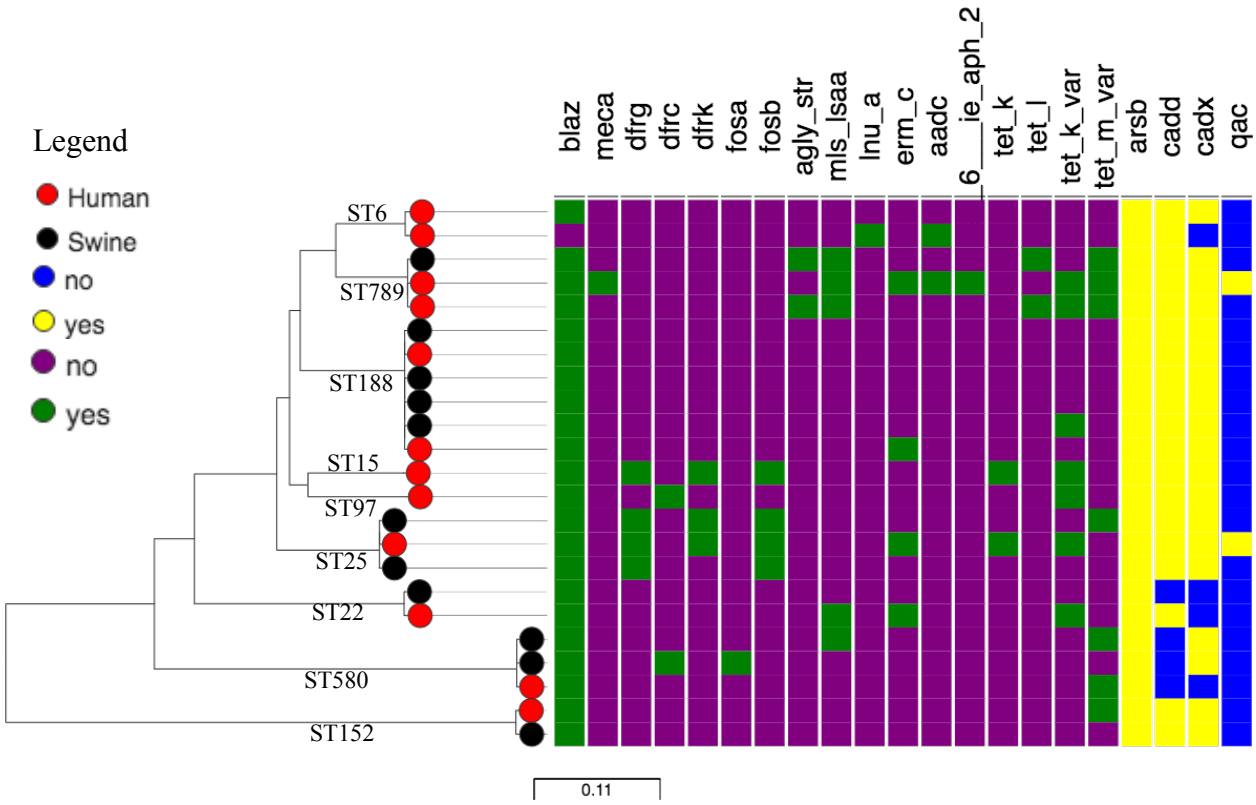


Fig. 9: A midpoint rooted maximum likelihood phylogenetic tree with heatmap distribution of AMR genes and heavy metals. The names at the top of heatmap indicate AMR and heavy metal gene names

The comparison of *in silico* prediction and phenotypic resistance available in supplementary file 5 based on the commonly used antimicrobial agents in the treatment of *S. aureus* infections as suggested by (Aanensen et al., 2016b, Gordon et al., 2014) in their studies for determination for the usefulness of whole genome sequencing in prediction of AMR. The results may not be accurate since the single colony used for antibiotic susceptibility test could not have been used for DNA extraction for subsequent sequencing.

3.1.7 *In Silico* Prediction of Virulence Genes

The presence or absence of virulence genes among 23 WGS data were predicted with ARIBA using a custom database of 102 set of virulence genes as the reference with a minimum threshold of 95% nucleotide identity and 99% coverage of query contig length to predict presence of a gene. A variety of virulence genes were identified in at least one of the strains encode for enterotoxins, leukocidin, capsular 8 polysaccharide, immune evasion genes clusters, haemolysins, biofilm formations, iron proteins, toxic shock syndrome, arginine catabolic metabolite enzymes (ACME),adhesive and serine proteases genes (fig 10).

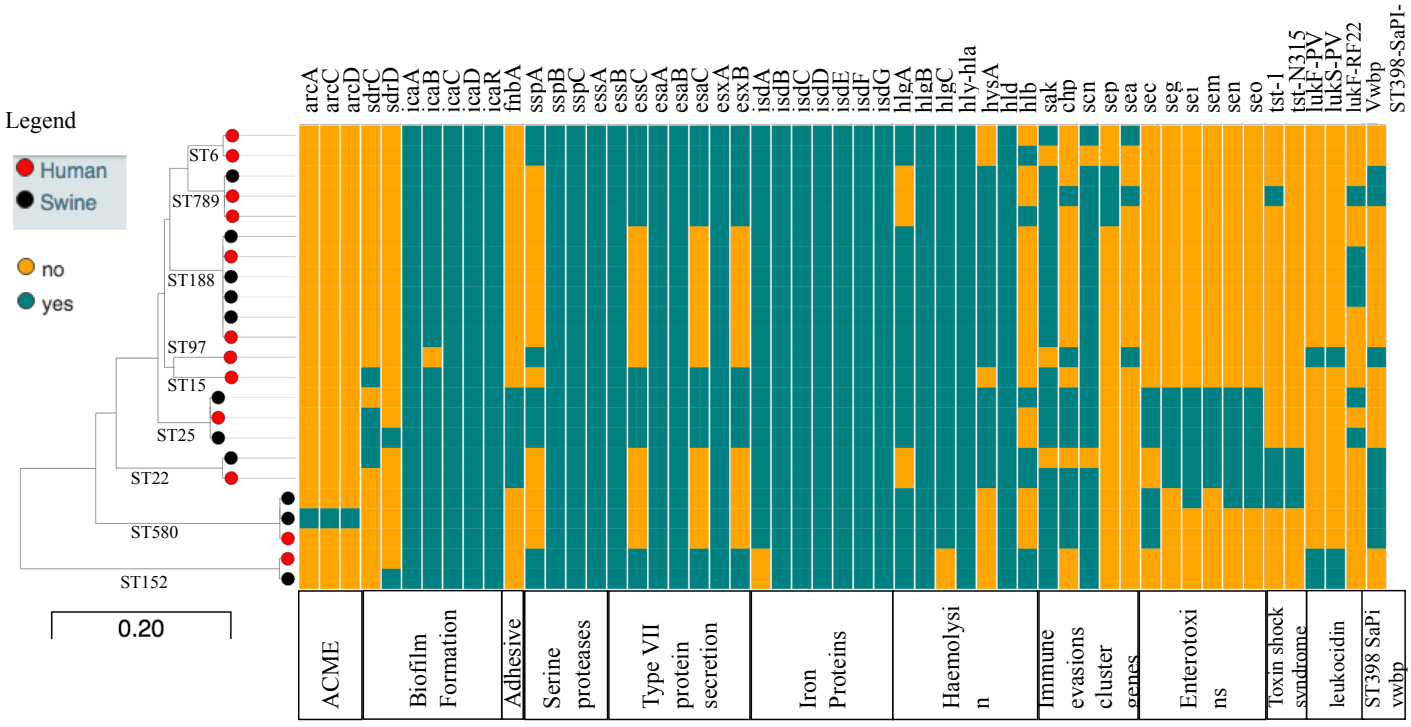


Fig. 10: A Midpoint rooted phylogenetic tree with a heatmap of virulence genes (not included capsular 8 polysaccharide) that were present in at least of one strain of the 23 isolates. Nodes indicate source of the isolates and genes names indicated at the top and the broad classification of the genes at the bottom

The Panton- Valentine leucocidin (PVL), encoded by two sub-unit proteins, lukF-PV and lukS-PV which is associated with skin and soft tissues infections such as necrotizing dermatitis (Lina et al., 1999) was present in 13% (n=3) strains including one isolate from swine origin of ST152.

Staphylococcal food poisoning characterized with emetic activity and nausea is likely to occur due to consumption of food contaminated with *S. aureus* which had been improperly cooked or stored allowing for their growth and expression of enterotoxins (Hennekinne et al., 2012). The classical enterotoxins (SEA-SEE) have been commonly implicated in food poisoning. About 15% (n=3) of human and none of the swine isolates carried the *sea* genes, and nearly a quarter (4 from swine and 2 from human) of the strains carried *sec* genes.

The enterotoxin gene cluster (*egc*) that comprises of *seg, sei, sem, sen* and *seo* genes were the most prevalent 6 (26% of the isolates) among the novel enterotoxins and were found mostly in strains of ST25 and ST22 lineages. To determine their genetic arrangement in each lineage, an in-house Blastn script was used to identify presence and location of these genes in the ordered assemblies and visualized their annotated GFF format file in Artemis (as shown in the figure 12). These lineages have similar genetic arrangement of these *egc* except presence of enterotoxin sec-like 1 (*entC1*) in ST25 and sec-like 3 (*entC3*) in ST22.

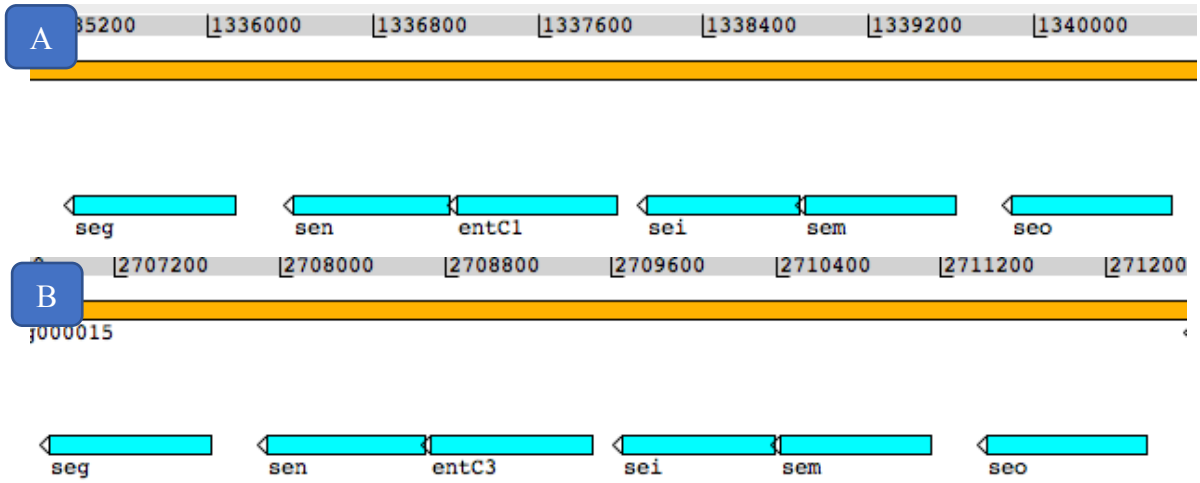


Fig. 11: Artemis visualization of enterotoxins gene clusters in ST25 (A) and ST22 (B)

Genes *tst-1* that mediate for toxic shock syndrome in humans were identified in 4 (17%, 2 from humans and 2 from swine) isolates, and 3 were co-detected with *egc* genes.

The distribution of human immune evasion gene clusters (IECs) among human and swine isolates was further explored. These genes are found in prophages of integrase group 3 (*phi*3) that are integrated into the ß-hemolysin genes (*hlb*), and are highly specific for human neutrophils and therefore, their presence in the host's isolate could indicate human origin.

Approximately 90% (n=21) of the 23 isolates tested for immune evasion gene clusters (IECs) were positive for two or three of the IECs genes (table 10). This suggests that about 90% (n=10) of the swine isolates in this study were colonized by isolates that were likely to have originated from humans. Only two isolates (ST22 from swine and ST6 from human) belong to phage type H that are associated with livestock adapted lineages.

Table 10: Different types of IECs (van Wamel et al., 2006b) with their prevalence among 23 strains

IECs type	Genes	Overall Prevalence	Human (n=)	Pigs (n=)
Phage type E	*sak* and *scn*	39%	4	5
Phage type B	*sak*, *chp* and *scn*	30%	3	4
Phage type D	*sea*, *sak* and *scn*	4%	1	0
Phage type H	lack all the IECs genes	8%	1	1
Not-yet typed	*sea*, *sep*, *sak*, *chp* and *scn*	4%	1	0
Phage type G	*sak*, *scn* and *sep*	8%	1	1
Not-yet typed	*sea*, *chp* and *scn*	4%	1	0

SECTION B

3.2.1 Relatedness of Kenya Isolates with Publicly Available Genomes

To contextualize humans and swine *S. aureus* genomes of this study in global genetic perspective, their genetic relatedness was compared with 126 publicly available genomes drawn from diverse countries as guided by MLST results but with a major focus on African genomes (supplementary file 6). The ST distribution of these public genomes is as follows: ST152 (n=25), ST188 (n=24), ST6 (n=14), ST22 (n=12), ST25 (n=19), ST398 (n=30) and ST789 (n=2). The public genomes of two STs (ST97 and ST15) that had only single strain in each were not included and the search for literature for public genomes in PubMed for ST580 were missing.

First, the genomes of 105 randomly selected genomes and 23 Kenyan isolates were mapped to TW20 complete genome to identify core genome SNPs (completed by Pathogen Informatics pipeline). After excluding 367 kb size of associated regions of MGEs of the reference in the alignment, 42851 core genome SNPs were identified among 128 isolates. This core genomes SNPs was used for maximum likelihood phylogenetic reconstruction with RAxML.

Generally, the SNPs diversity of isolates of each ST were increasing over time in terms of SNPs pairwise differences on mapping to TW20 reference (Fig. 12).

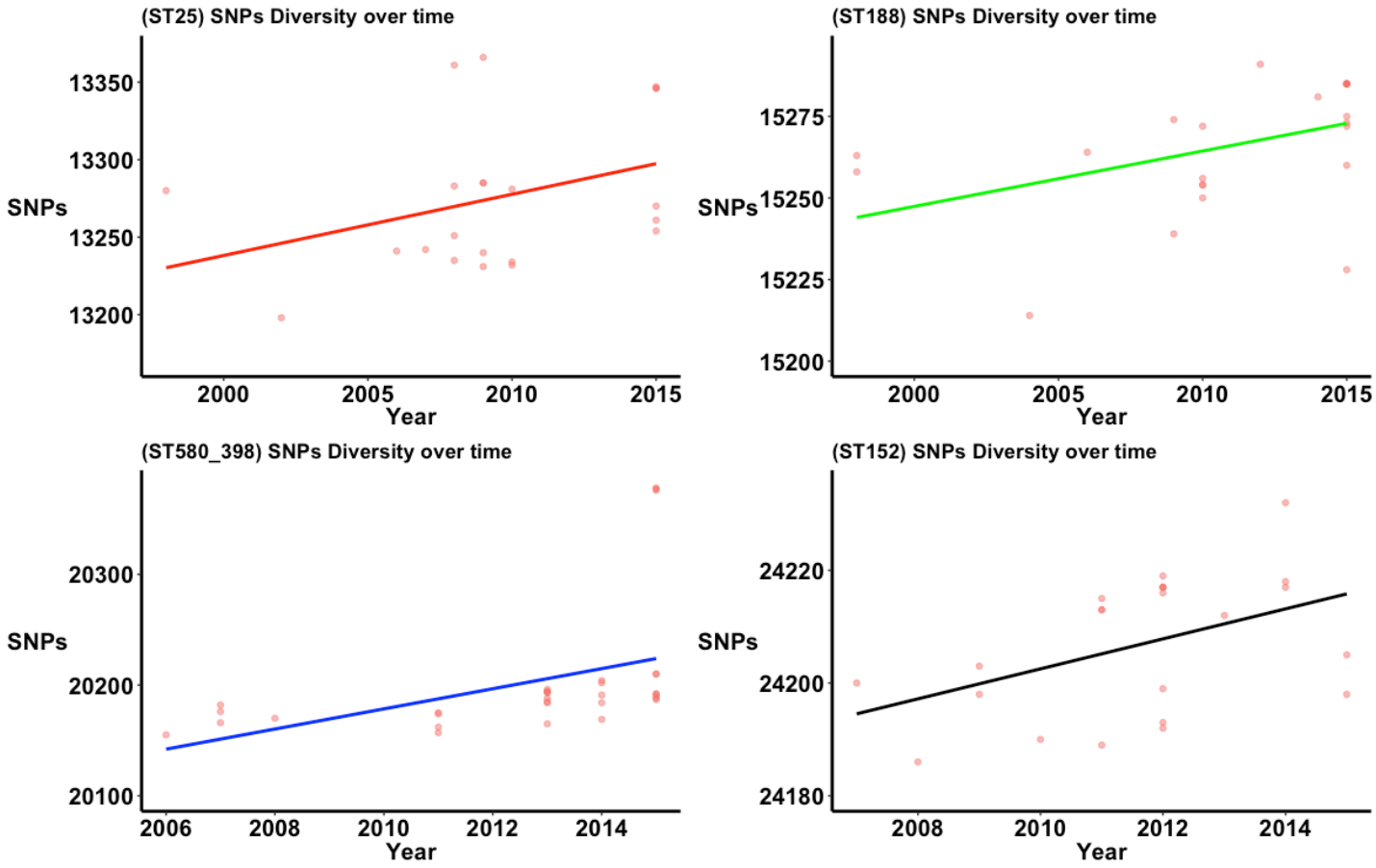


Fig. 12: Regression lines of each ST in terms of SNPs pairwise difference over time (in years) in relation to TW20 (ST239). The dot represents the number of SNPs of individual strain while the line represents the line of best fit distribution of isolates generated in ggplot2 in RStudio.

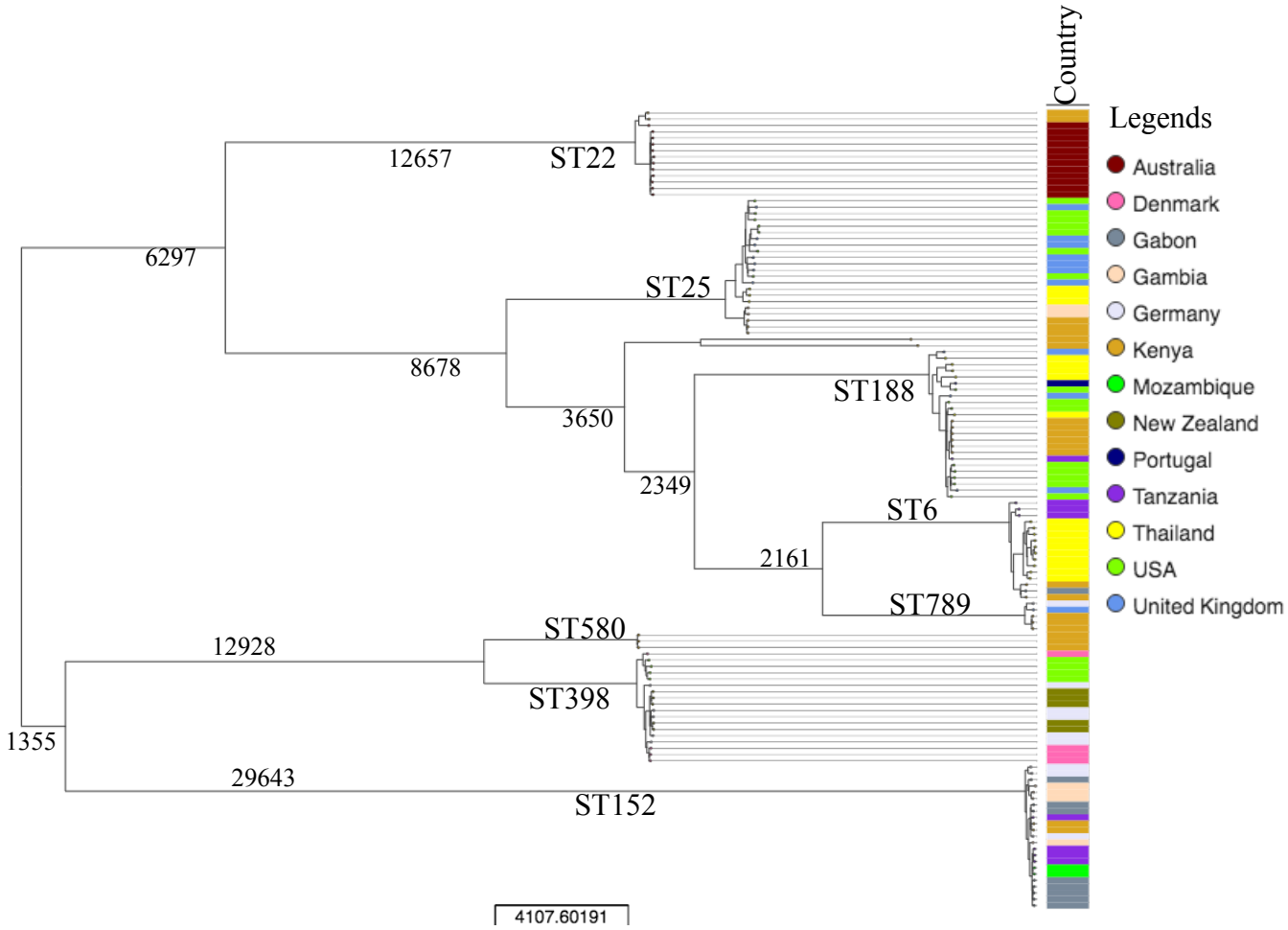


Fig. 13: A Midpoint rooted maximum likelihood phylogenetic tree that has been optimized with ACCTRAN algorithms. The tree tips are aligned to country of origin. The legend indicates the color of country. The number in the branches indicate number of SNPs in the branch

A midpoint rooted maximum likelihood phylogeny (fig 13) segregated into MLST lineages to form the same 7 distinct clusters as shown before in phylogenetic tree of Kenya isolates (fig 7) suggesting associations of lineages circulating in humans and swine isolates of Kenya with these global lineages. Strikingly, ST152 clusters were characterized with long deep branch as compared to other MLST clusters suggesting that they are distantly related to other clusters.

3.2.2 Detailed Phylogenetic Analyses of Selected Sequence Types (STs)

To better understand the genetic clustering of Kenyan isolates in relation to global genomes, individual subtrees of individual ST (ST25, ST789, ST22, ST6) were reconstructed based on their respective core genome SNPs on alignment to TW20 reference. Nevertheless, ST188 and ST152 were mapped to respective ST references using custom script. Fourteen regions of 132 kb size that were associated with repetitive elements of the reference were excluded from the alignment in ST152 resulting in 1446 core genome SNPS for maximum likelihood reconstruction of phylogeny. To examine the effect of these core genome SNPs in the genomes of ST152, accelerated transformation was used to reconstruct maximum likelihood phylogeny with pseudogenome alignment and annotated reference file. This led to identification of non-synonymous, synonymous SNPs and intergenic (table 11).

Table 11: shows SNPs effect of core genome of ST152 as determined by ACCTRAN reconstruction of phylogeny with original pseudogenome alignment and annotated reference file

No. of SNPs	Type of SNPs
512	Intergenic
990	Synonymous
1066	Nonsynonymous
23	Non-stop codon to stop codon

The recombination regions were predicted using Gubbins in ST188 where it identified approximately 71kb size repetitive regions of the pseudogenome alignment (fig 14) and were removed before SNP sites extraction in genome alignment and RAxML phylogenetic tree reconstruction. These recombination regions were phages with various genes that annotate for functions such as phage proteins of unknown functions, phage major capsid, putative phage PVL protein, phage tail fiber proteins, hypothetical phage related proteins, putative phi ETA-like proteins, bacteriophage protein of unknown functions among other phage functions.

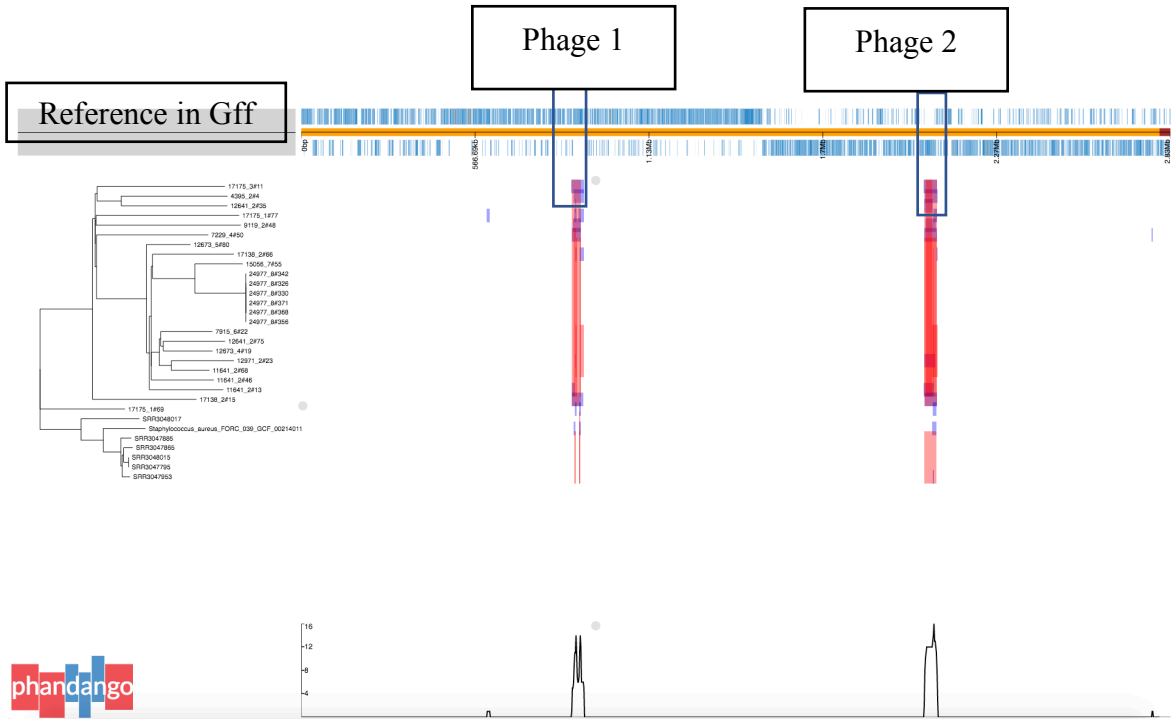


Fig. 14: Phandango visualization of 2 regions of repetitive regions (Red) in pseudogenome alignment of ST188 as identified by Gubbins alongside RAxML phylogenetic tree reconstructed by Gubbins without these regions.

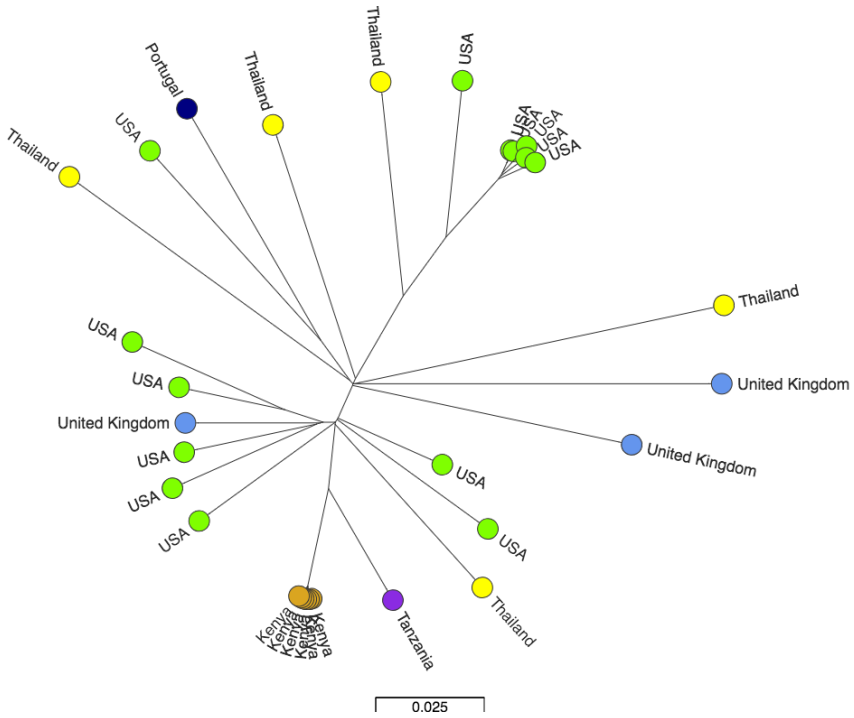


Fig. 15: Midpoint rooted phylogenetic tree of ST188 based on 2552 core genome SNPs on mapping to HongKong draft genome reference. Nodes indicate country of origin.

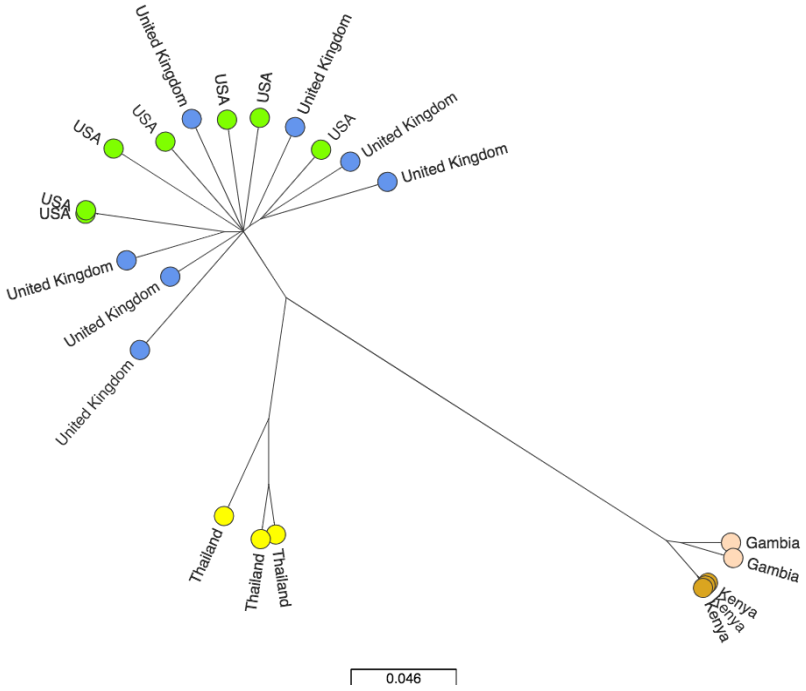


Fig. 16: A Midpoint rooted ML phylogeny of ST25 based on 3011 core genome SNPs on alignment to TW20 reference

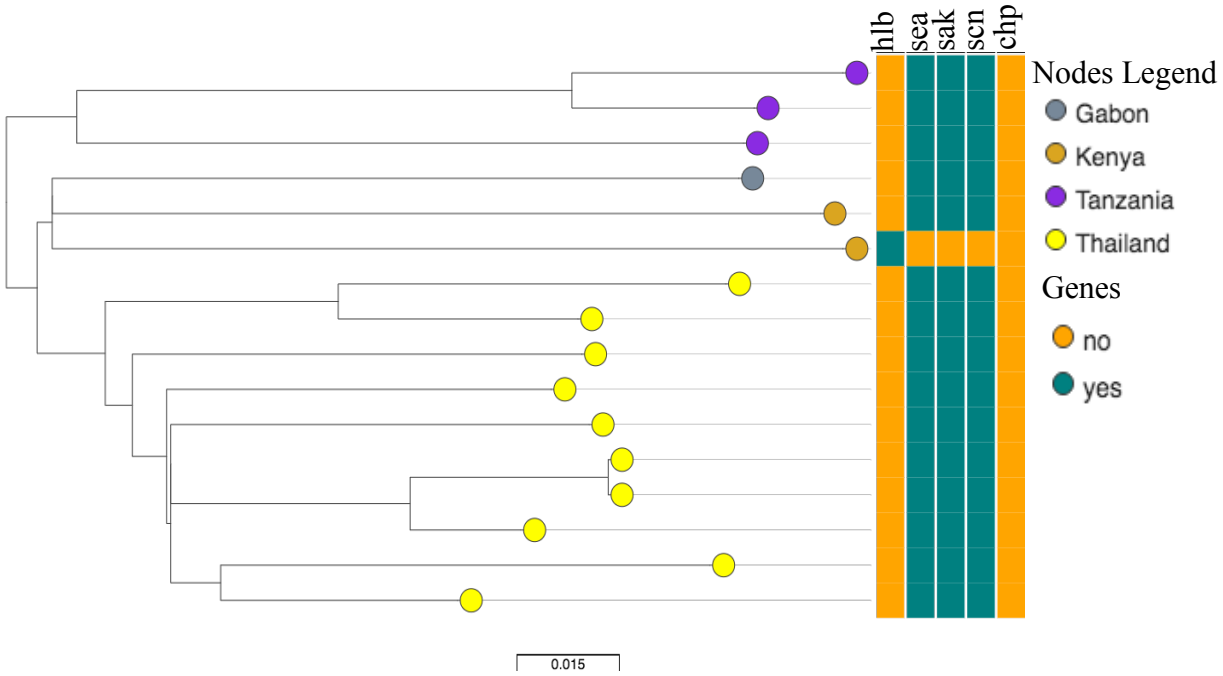


Fig. 17: Midpoint rooted ML phylogenetic tree of ST6 based on 1283 core genome SNPs on alignment to *Staphylococcus aureus* TW20 reference. The tree tips are aligned to presence or absence of integrase prophage group 3 genes. Green color in the heatmap indicate presence and orange color absence

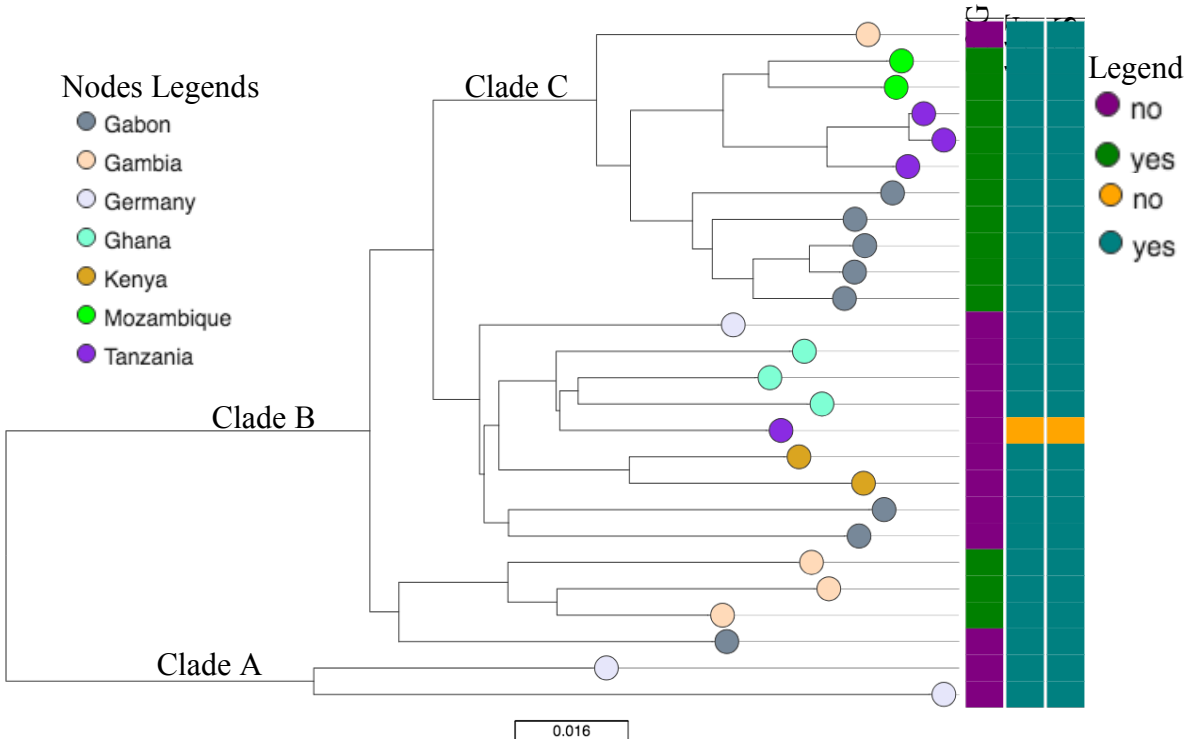


Fig. 18: Midpoint rooted ML phylogenetic tree of ST152 based on 1446 core genome SNPs on alignment to *Staphylococcus aureus* BB155 reference. The tree tips are linked to dfrG and PVL presence and absence genes

Generally, Kenyan isolates were phylogenetically related to other African *S. aureus* strains. For instance, Kenyan isolates of ST188 were closely related to an isolate from Tanzania among 20 genomes collected from six countries representing 4 continents (Fig. 15). A similar observation was demonstrated in ST25 phylogeny (Fig. 16), where three Kenyan and two Gambian strains formed distinct clades among three phylogenetic clades of 22 genomes from across three other continents.

Furthermore, phage type D that are knowns to carry staphylococcal enterotoxin a (*sea*), staphylokinase (*sak*) and staphylococcal complement inhibitory protein (*scn*) (van Wamel et al., 2006b) were present in almost all strains 92% (n=13) of ST6 except one Kenyan isolate that had intact *hlb* although it was sampled from human (Fig. 17) .

Another important observation in the phylogenetic tree of ST152 (Fig. 18) was clade C that had two sub-clusters; one for trimethoprim resistant (*dfrG*) and another for susceptible genomes. The Kenyan strains being trimethoprim susceptible were closely related to Tanzania and Ghana genomes that were also susceptible to the drug. The temperate bacteriophage that harbor *luk*S-PV and *luk*F-PV seem to be conserved in this lineage ST152. The distribution of nucleotide sequences of PVL genes in ST152 were identical with exception of two novel non-synonymous mutations at position 532 and 668 of *lukS-PV*, and position 153 and 789 for *lukF-PV* (Zhao et al., 2016), (Fig. 19) of two Germany strains of clade A (Fig. 18) in the phylogenetic tree.

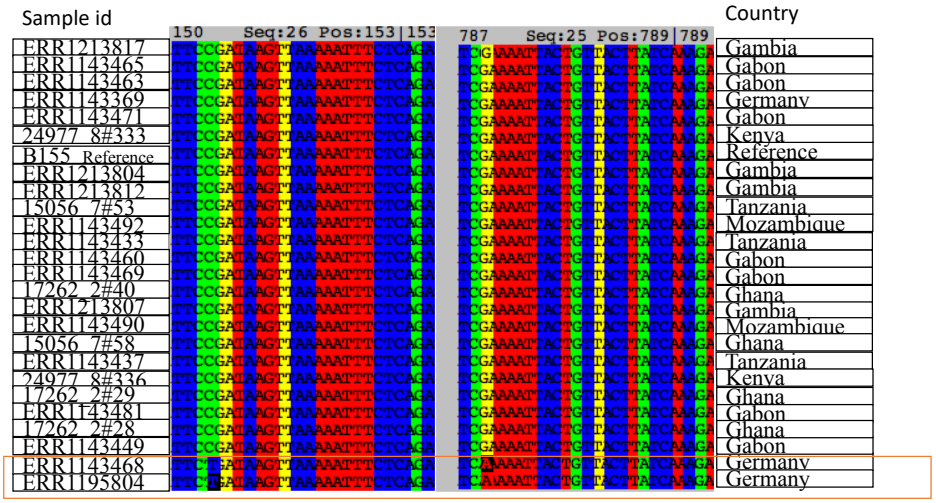


Fig. 19: Multiple sequence alignment of lukF-PV showing non-synonymous SNPs at positions 153 and 789 of two German genomes (bottom highlighted rectangular shape)

3.2.3 Phylogenetic Analyses of ST580 and Livestock-associated ST398 lineage

To find out whether three Kenyan isolates (2 swine and 1 human) of ST580 could be livestock adapted or human adapted lineage, I analyzed their phylogenetic relatedness with 30 global genomes of ST398 that were isolated from 7 countries between 2006 to 2015. The ST580 lineage (ST profile 3-35-48-19-20-26-39) is a double locus variant of ST398 (3-35-19-2-20-26-39) because they differ by only two allelic loci of the seven housekeeping genes used in determination of multi-locus sequence types (MLST) (http://saureus.mlst.net). These 33 strains were mapped to livestock-associated MRSA ST398 reference and subsequently reconstructed phylogenetic based on 8575 core genome SNPs.

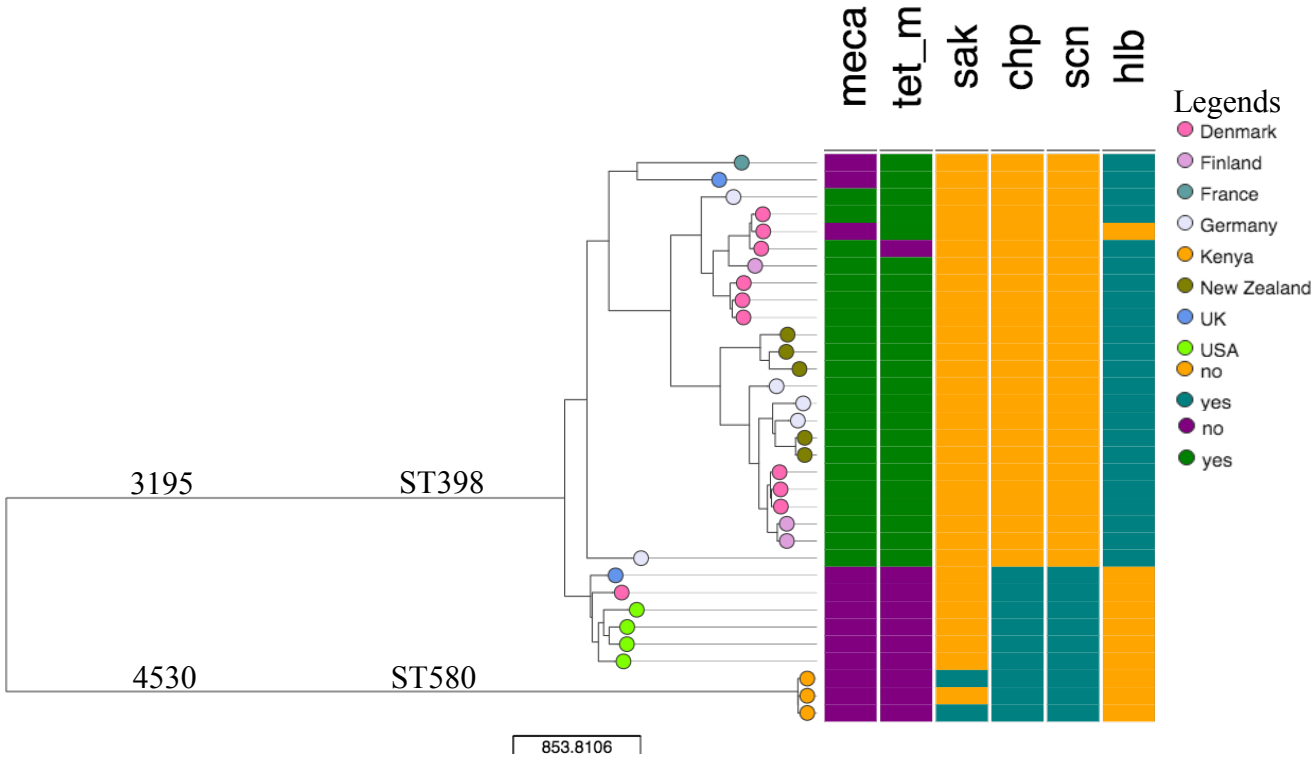


Fig. 20: Maximum Likelihood phylogeny of ST398 and ST580 based on 8575 core genome SNPs that has been optimized with ACCTRAN algorithms. The tree is linked to the heatmap of AMR and IECs genes. The number in the branched show the total number of SNPs in the branch after parsimony.

As expected, the genomes in the phylogeny grouped into 2 distinct clusters (Fig. 20), one coinciding with ST398 and the other with long branch length for 3 strains of ST580. The ST580 Kenyan isolates differ with the reference by an average of 6925 SNPs, suggesting that they are distantly related. Interesting, ST398 global collection strains formed two major clades; the basal clade with strains that were methicillin susceptible and had ß-toxin converting phage type C (positive for *chp* and *scn*), and the top clade with strains that were mostly methicillin (*mec*A) and tetracycline (*tet*M) resistant and had lost phages that carried human immune evasions cluster genes (*scn*, *sak*, *chp*) that are located in ß-hemolysin (*hlb*). This followed a similar observation by Price et al. (Price et al., 2012b) that the human adapted MSSA of ST398 with phages that carried human modulatory proteins, methicillin susceptible, tetracycline susceptible were forming ancestral basal clades. Another striking feature of the phylogenetic tree was that Kenyan ST580 strains were relatively closer (average of 6945 SNPs) to human adapted strains in the basal clade compared to livestock adapted isolates (average of 6970 SNPs differences pairwise) of ST398. This is further supported by strains of ST580 with similar characteristics of isolates of basal clade of ST398 cluster such as methicillin and tetracycline susceptible, negative for *hlb* genes and harboring IECs genes.

3.2.4 Accessory Genomes Analyses of ST580 and ST398

In order to understand further the genetic relatedness of ST580 and ST398, I explored their distribution of genes in their accessory genomes using Roary which bin the genes into core and accessory genomes. The number of genes in the pan-genome was 4084, in which 1910 genes were core genes and 2174 genes were accessory. 2221 genes were discarded that were present in more than 90% of the strains and created a pairwise matrix (supplementary file 7) based on 1863 accessory genes using a custom python script. In order to visualize this output pairwise matrix, R Studio were used to generate heatmap and clustering tree based on the pairwise proportions of number of shared genes in the accessory genomes (fig 21). The darker the color in the heatmap, the higher the number of shared genes between isolates.

The analysis of the accessory genomes inferred by Roary showed that Kenyan isolates of ST580 seemed to share a higher number of accessory genes with human adapted MSSA than livestock adapted MRSA strains of ST398 lineage (Fig. 21). The annotated genes that the

ST580 strains and human-adapted genomes of ST398 shared include autolysin, some groups of transposon-related proteins, hypothetical proteins, lipoproteins, some phages that encode for *chp*, *scn* and *sak* genes, serine aspartate repeat proteins *sdrC*. They differ mostly with livestock adapted genomes in the ST398 lineage in terms of genes that encode for phage proteins, transposase, recombinases proteins, pathogenicity islands proteins, hypothetical proteins among many other mobile genetic elements.

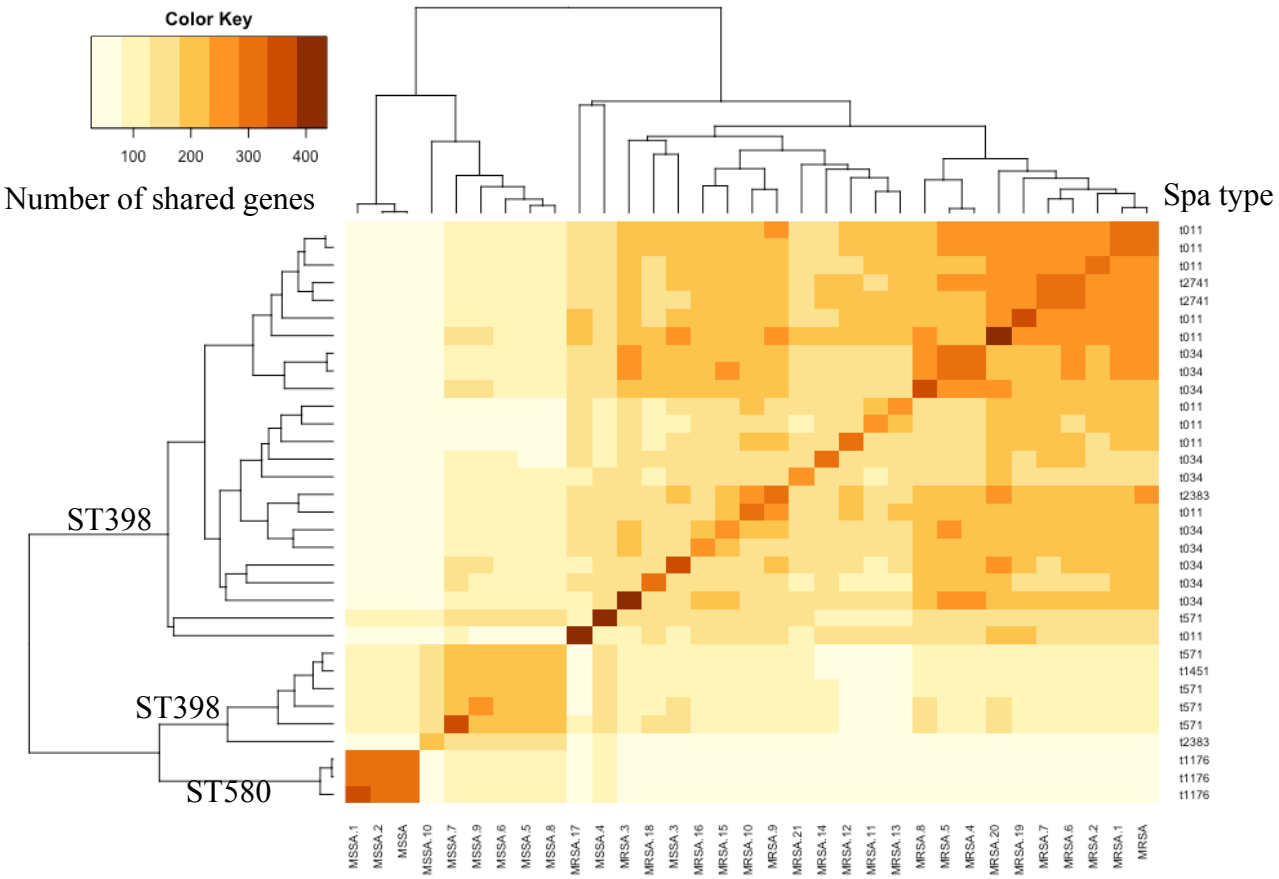


Fig. 21: Heatmap based on the number of shared genes in the accessory genomes. The x-axis labels indicate presence of either MSSA or MRSA while y-axis indicate the spa type of the STs. The color key indicates the proportion of number of shared genes. The genomes were clustered based on the accessory gene distribution