CHAPTER TWO

METHODS

2.1 Ethical Considerations

Ethical approval was obtained from Kenya Medical Research Institute (KEMRI) and the Department of Livestock and Fisheries of Kenya government to conduct sampling and collect metadata information from participating farmers and their corresponding pigs in Kenya, and later for the shipment of the DNA and implementation of the study in the United Kingdom.

2.2 Study Design and *S. aureus* Isolation

Between October 2015 to August 2016, 248 *Staphylococcus aureus* isolates were collected from the nares of farmers and their corresponding pigs from different homesteads spread throughout 9 districts of Kiambu County, Kenya. Kiambu has a population of 1.5 million persons in a 2500 km^2 and is adjacent to north of Nairobi, capital city of Kenya. From each homestead, up to a maximum of 3 farmers and 3 pigs from herd size that range between 5 and 125 were randomly selected for sampling. About 55% of the homesteads sampled were keeping pigs for commercial purpose.

Written informed consent was obtained from each participating person. Swine isolates were excluded from the study if the handler or the owner refused to be sampled or to provide information such as age, history of the usage of antimicrobial agents, number of pigs in a farm, presence of pets and other livestock, and size of the farm. Sampling was done by inserting sterile cotton swabs moistened with 0.85% sodium chloride into the nose of the host.

The samples were sent within 24 hours to the Kenya Medical Research Institute/Center for Microbiology Research for processing. The nasal swabs were enriched in a 5 ml Tryptone soy broth and incubated at 37^{0}C for 24 hours. On the following day, they were transferred to a differential media where they were plated on Mannitol Salt Agar (MSA) and incubated overnight at 37^{0}C for 18-24 hours. The presumptive positive yellow colonies were subjected to Gram staining, and catalase (Oxoid, UK), oxidase (Oxoid, UK) and latex Staphaurex agglutination (Oxoid, UK) tests in order to confirm presence of *S. aureus* species.

Ninety-five isolates collected from 23 homesteads were randomly selected for DNA extraction, 47 from humans and 48 from pigs from 29 homesteads. All except 8 strains were paired by homesteads.

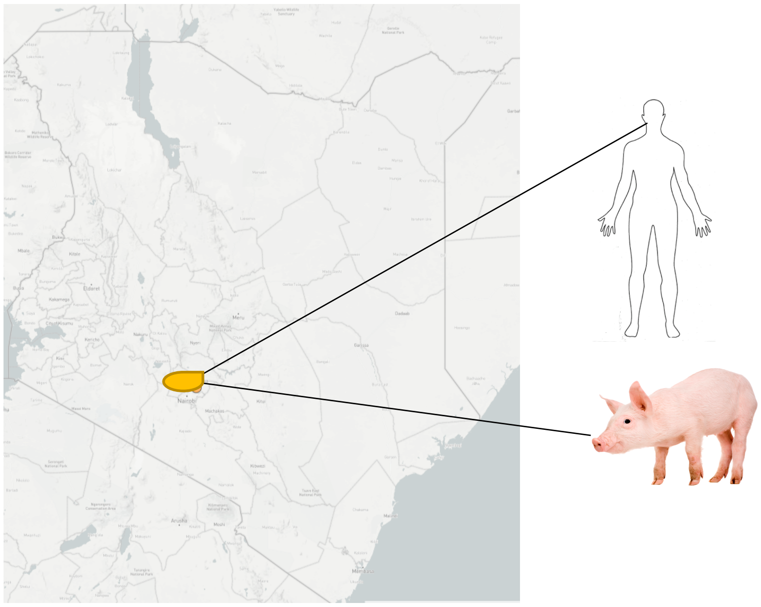

Fig. 1: Map of Kenya with the yellow dot showing Kiambu county where pigs and humans isolates were sampled

2.3 Antibiotic Susceptibility Testing

I determined antibiotic susceptibility test at KEMRI (September 2017) for 96 strains against a panel of 17 antibiotic drugs in three Mueller Hinton Agar (Oxoid, UK) plates per isolate using Kirby – Bauer Disk Diffusion. These drugs were; ampicillin (AMP), amoxicillin-clavulate (AMC), ceftazidime (CAZ), nalidixic acid (NA), gentamicin (CN), chloramphenicol (C), ciprofloxacin (CIP), cefoxitin (FOX), ofloxacin (OFX), trimethoprim-sulfamethoxazole (SXT), nitrofurantoin (F), linezolid (LZD), quinupristin-dalfopristin (QDA), amoxicillin (AML), imipenem (IPM), erythromycin (E), doxycycline (DO). Their diameter zone inhibition breakpoints results were interpreted with clinical and laboratory reference guidelines 2015.

2.4 DNA Extraction

The colonies that were identified to be *S. aureus* in step 2.2 were revived by streaking on Mueller Hinton Agar plates and incubating for 18-24 hours at 37^0C. Genomic DNA was extracted using the GenElute DNA kit (Sigma-Aldrich, USA) following the manufacturer's instructions. DNA were prepared ready for shipment on 96 barcoded wells according to the specifications of the WTSI sequencing facility. The DNA extraction and preparation for shipment of DNA were carried out by collaborators at KEMRI (December 2017).

2.5 Library Preparation and Whole Genome Sequencing

The WTSI core sequencing department prepared index-tagged multiplex paired-end libraries, and then sequencing of the samples were completed on the Illumina-B HiSeq X platform generating 150 bp paired-end reads, following previously described protocols (Quail et al., 2008, Harris et al., 2010b).

2.6 *S. aureus* Assembly and Annotation

Following Illumina sequencing of the isolates, the raw reads were *de novo* assembled and annotated automatically with the Sanger Pathogen Informatics pipeline as described by Page et al. 2016. In brief, multiple assemblies of the sequence data of each isolate were first generated using Velvet (v1.2) (Zerbino and Birney, 2008) and VelvetOptimiser (v2.2.5)(http://bioinformatics.net.au/software.velvetoptimiser.shtml). The SMALT(http://www.sanger.ac.uk/science/tools/smalt-0) was used to identify assembly reads that were mapped to different contigs or aligned to the same contigs but in improper orientation or that had different insertion sizes or were totally unmapped, to be subjected to further assembly improvement steps. This involved scaffolding of contigs of best N50 using SSPACE v2.0 (Boetzer et al., 2011) followed by closure of the gaps (1 or N's) with GapFiller v1.11 (Boetzer and Pirovano, 2012) that were generated during scaffolding of the contigs. The closure was done through cycling of BWA (Li and Durbin, 2009) and Bowtie (Langmead et al., 2009). The contigs that were shorter than the targeted fragment size length (300-500 bases) were removed in the final assembly. Finally, statistics of the assembly quality were produced by mapping back the sequence reads to the final assembly using SMALT.

The final assembly sequences in FASTA format proceeded automatically to the annotation step that was performed using PROKKA (v1.11) (Seemann, 2014). PROKKA used PRODIGAL (Hyatt et al., 2010) to identify the coding sequences and compared them to genus specific databases from NCBI Reference Sequences (RefSeq) (Pruitt et al., 2012). The BLAST+ method (Camacho et al., 2009) was then used to search for the highest similarity results that were transferred for annotation

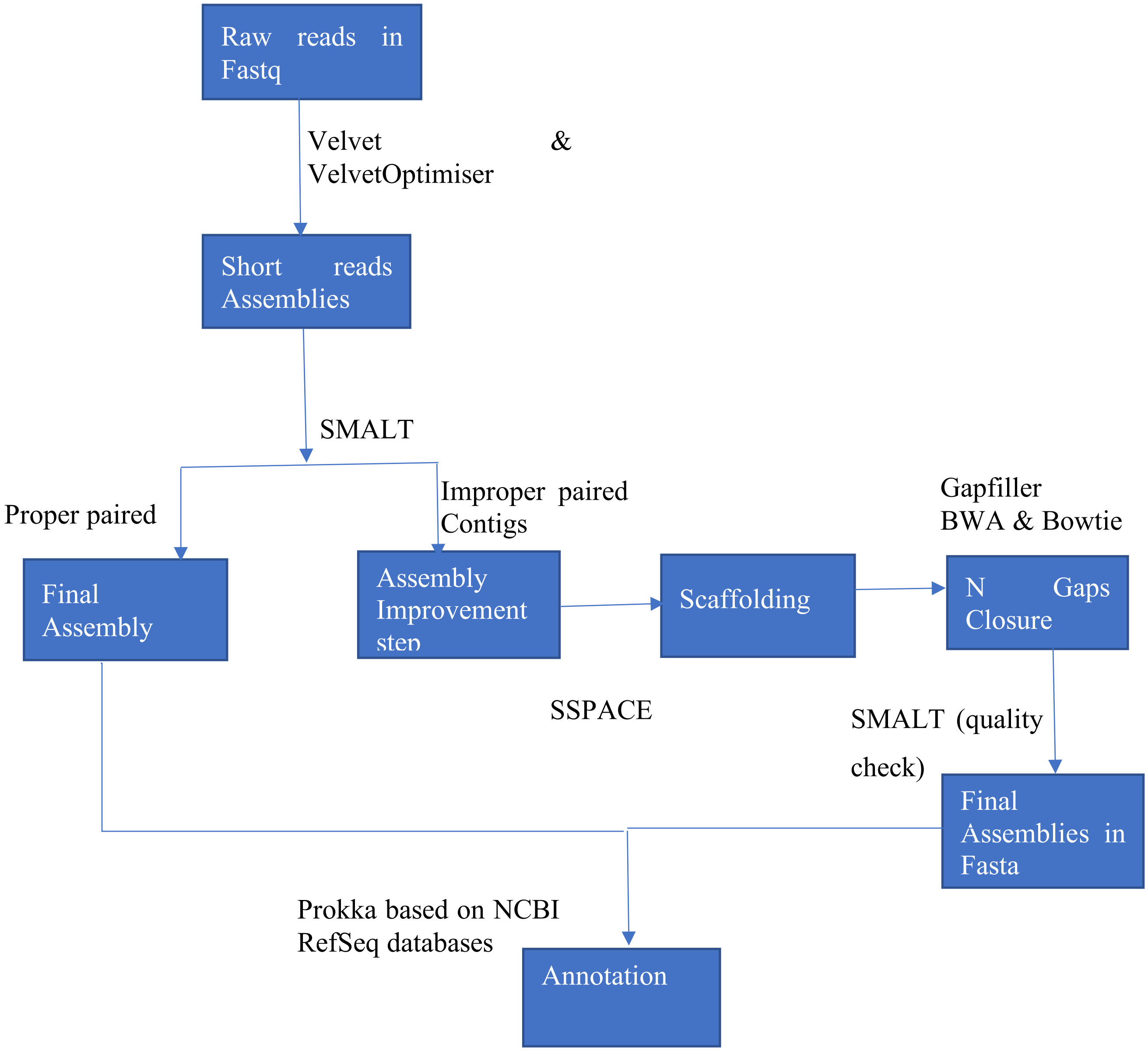

Fig. 2: A schematic diagram summarizing the assembly of *S. aureus* genomes by the Pathogen Informatics Pipeline as described by Page et al. 2016

2.7 Mapping and Variant Calling

2.7.1 Using Pathogens Informatics pipeline

Using the Pathogen Informatics pipeline, Illumina raw reads were mapped onto different reference genomes (table 1) depending on the analyses with Burrows Wheeler Alignment (BWA v0.7.15-r1140) (Li and Durbin, 2009). Briefly, alignment of the raw reads to the reference was done under default settings of BWA. After mapping, base call substitutions were identified using SAMtools mpileup v0.1.19 (Li et al., 2009) with Picard v1.92 detecting and marking duplicates in the BAM file. Bcftools v0.1.19 was used to filter out low quality SNPs that failed to meet the stringent threshold of the parameters in Table 2

Table 1: Different references used in mapping and SNP-calling analyses.

Dataset	**Reference strain (complete genomes)**	**Accession**	**Mapping pipeline**	**Identification of Recombination regions**
Kenyan isolates	*Staphylococcus aureus* subspecies aureus MSSA476	GCF_000011525_1	Pathogens Informatics	Identified by Mathew Holden
Global context of Kenyan isolates	*Staphylococcus aureus* subspecies aureus TW20	FN433596	Pathogens Informatics	Provided by Silvia Argimon
ST152 datasets	*Staphylococcus aureus* BB155	GCF_900004855_1	Custom script	Identified by Matthew Holden
ST188 strains	*Staphylococcus aureus* FORC_039	GCF_002140115_1	Custom script	Gubbins detection
Genetic comparison ST580 and ST398	*Staphylococcus aureus* subspecies *aureus*	GCF_001887075_1	Custom script	Gubbins detection for ST398 strains

	LA_MRSA_ST398			

Table 2: Quality control thresholds that variant sites must pass for them to be called SNPs

Parameter	**Thresholds**
Bcf variant quality score	>50
Mean Phred quality score	>30
Same allele frequency (af) of the base call as the reference	0
Allele frequency of SNP Variant base call of the reference	>0.95≤1
Read mapping of the base call	Ratio > 0.75
Mapping depth	> 4 reads
Mapping depth per strand	> 2 reads
Strand_bias, mapped_bias, tail bias	P value < 0.001

The pipeline produced VCF files that had unfiltered variants, filtered variants sites and pseudogenome alignment in FASTA format. A pseudogenome that was reconstructed had Ns in addition to SNPs, to represent the variant sites that failed filtering thresholds and bases that were deleted with respect to the reference. However, it lacked any bases that were inserted with respect to the reference.

2.7.2 Using a Custom Script

A custom script (developed by Simon Harris) employed a similar approach to the Pathogen Informatics pipeline in mapping and calling SNPs as described in step 2.7.1 such as use of BWA for alignment, SAMtools v-1.2 mpileup for detection of variants, BCFtools for quality control filtering of SNPs and Picard for marking duplications. However, the default settings for calling SNPs in this custom script are shown in Table 3

Table 3: Filtering parameters with cut-off values in the custom mapping and SNP-calling script.

Filtering parameter	**Cut-off value**
Minimum base call quality	50
Minimum mapping quality	20
Minimum depth reads matching SNPs	8
Minimum depth reads matching SNPs per strand	3
Minimum quality ratio of SNP/mapping reads	0.8

2.8 *In Silico* Prediction of Recombination Regions in Pseudogenome Alignments

Genealogies Unbiased By recomBination in Nucleotides Sequences (Gubbins) (https://github.com/sanger-pathogens/gubbins) (Croucher et al., 2015) was used with default parameters to predict recombination regions. These were identified as regions with a higher density of base substitutions in a pseudogenome alignment generated in step 2.7. In summary, phylogenetic reconstruction of the tree was performed with RAxML using the pseudogenome alignment, followed by reconstruction of SNPs on the phylogeny using Phylogenetic Analyses by Maximum Likelihood (PAML). This was followed by successive iterations of phylogenetic reconstruction with RAxML based on a reduced alignment after detection and subsequent removal of recombination events. At least 3 base call substitutions were used in identification of recombination regions. The output files include a newick tree file and the predicted recombination regions in GFF format, obtained upon detection of similar recombination regions on two successive repeats, and were visualized with the web-based application, Phandango (Hadfield et al., 2018).

2.9 Phylogenetic Reconstruction

Regions such as insertion sites, phages, transposons and mobile genetic elements defined in EMBL tab-delimited format were first removed from the pseudogenome alignment with a python script, *remove_blocks_from_aln* (https://github.com/sanger-pathogens/remove_blocks_from_aln). Variable positions that were exclusively 'ACTG' (without Ns) in all samples were extracted using the *SNP sites* tool (Page et al., 2016). Finally, the phylogenetic tree was inferred with RAxML-HPC v7.0.3 (Randomized

Axelerated Maximum Likelihood for High Performance Computing) using the SNP alignment. The RAxML-HPC used the general time reversible (GTR) model with a gamma correction for nucleotide substitution and 100 rapid bootstrap replicates. The resulting tree was mid-point rooted in FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/) and exported to Microreact (https://microreact.org/) (Argimon et al., 2016), a web-based visualization tool, and combined with metadata.

In some instances, optimization of SNPs in the phylogeny and quantification of number of SNPs at the branches and the effect of SNPs on annotation were completed with a custom script (developed by Simon Harris). The script required pseudogenome alignment generated in 2.7 and best phylogenetic tree as the inputs for homoplastic parsimony, and the annotation of the reference in embl format for SNP effect prediction. It then performed optimization based on accelerated transformation (ACCTRAN) algorithms.

2.10 Determination of Antimicrobial Resistance and Virulence Genes using ARIBA

The presence of antimicrobial resistance (AMR) genes was determined directly from sequence reads using ARIBA (Hunt et al., 2017) with Resfinder (Zankari et al., 2012) and Arg-annot (Gupta et al., 2014) reference public databases, and custom databases of known AMR genes and mutations (developed by Matthew Holden and Sandra Reuter). In summary, coding sequences of the similar genes for the reference database were first clustered together using CD-HIT on default settings (Table 4) (Huang et al., 2010). This was followed by alignment of the paired raw reads of each strain to each cluster of the reference database with *Minimap* (Li, 2016). The reads that mapped to each cluster of reference sequences were assembled separately using Fermi-lite (https://github.com/lh3/fermi-lite). Then contig sequences of the assemblies with highest percentage to the cluster reads of the reference were identified with *nucmer* and variant information between the sequences and indels were detected with the *show-snps* program of the MUMmer package (Kurtz et al., 2004). In order to verify the completeness of the assemblies to the reference and identified known variants, the cluster reads were mapped back to the contigs with Bowtie2 (Langmead and Salzberg, 2012), and the depth coverage of the contigs and variants sites were determined with SAMtools mpileup (Li et al., 2009).

Heavy metals and antiseptic resistance protein genes sequences for the accession numbers in table 5, were downloaded from ENA and their presence among 23 strains were predicted in ARIBA as described before in AMR determination.

The virulence genes were predicted with a custom database of 102 virulence genes (provided by Matthew Holden and Ewan Harrison) using ARIBA as highlighted in the AMR determination.

Table 4: Default parameters in ARIBA.

CD-HIT clustering settings in preparation of reference databases	**Thresholds**
Sequence identity thresholds	≥0.9
Length sequence cut-off	0
Length of reference genes nucleotides	>6 ≤10,000
Number of genetic code to use	11
Number of threads	1
Nucmer default settings in selection of best match sequences	
Identity of sequence alignment	≥90
Length of sequence alignment	≥20
Depth coverage of reads for assembly	≥50
Number of reads for scaffolding of two contigs	≥10

Table 5: Showing heavy metals and antiseptic and their accession numbers

Heavy Metals and antiseptic	Accession numbers
arsB	AWW93856
arsC	PZH89330
cadX	AXE42895
cadD	PZL84176
qacA	AXE42913
qacB	BAG12275
qacC	SRE49840

2.11 Multi-Locus Sequence Typing (MLST) and *spa* Typing

The alleles of seven housekeeping genes of *S. aureus* used for multi-locus sequence typing (MLST), (https://pubmlst.org/saureus/), *arcC*, *aroE*, *glpF*, *gmk*, *pta*, *tpi* and *yqiL,* (supplementary file 1) were predicted *in silico* from assemblies using MLSTcheck script (https://github.com/sanger-pathogens/mlst_check). The nucleotide sequences of Protein A (*spa*) genes were extracted from the annotated assemblies using an in-house script (developed by Pathogen Informatics) and were uploaded to a web tool for assigning *spa* types (http://spatyper.fortinbras.us/). The tool uses Ridom and Kreiswirth nomenclatures for identification of repeat units and types of polymorphic X region of *spa* nucleotides.

2.12 Pan Genome Analyses of ST398 and ST580 strains

The 33 annotated assemblies of ST398 and ST580 strains produced as described in section 2.6 were used as input files (in GFF format) for Roary (Page et al., 2015), a pan-genome analysis tool. Roary was run using default settings, except that options -n to align each core gene family one by one with MAFFT and -e to generate a core genome alignment with PRANK. In summary, the initial step involved converting the nucleotide sequences in the annotated assemblies to protein sequences and subsequently filtering out the partial sequences. This was followed by iterative clustering of the remaining sequences with CD-HIT (Fu et al., 2012) starting with 100% identical sequences down to a default of 98% sequence identity. The last step involved pairwise comparison of each of the cluster sequences with BLASTP using the default parameter of 95% sequence identity and 100% matching sequence length. Subsequently, genes were then ordered and grouped into respective core or accessory regions based on their occurrence in the input sequences.

Genes that were present in more than 90% of the strains were discarded and the remaining 1863 genes were used for accessory genomes analyses. A python script was used to create a pairwise comparison of the number of accessory genes that were shared between any two strains of these 33 isolates. The pairwise comparison output was then converted to a matrix and subsequently clustered in R studio v1.1383 with ‘heatmap.2’ and coloured with ‘RColorBrewer’ from the ‘gplots’ package. The clustering was based on the proportion of shared genes in the accessory genomes between any two isolates.

2.13 Public Genomes Selection

117 genomes from a collection of over 10,000 *S. aureus* public genomes that were downloaded by Centre for Genomic Pathogen Surveillance from the European Nucleotide Archive (ENA) were randomly selected for inclusion in this study. In addition, seven genomes from the Washington National Primate Research Center in USA were downloaded from NCBI (Soge et al., 2016) using fast-dump package of the SRA toolkit program v2.92. The selection of genomes (Table 6) was based on the multi-locus sequence typing (MLST) results of the Kenyan *S. aureus* isolates from this study.

Table 6: Showed number of strains of public genomes from different study project numbers and the publications reference

No. of Genomes	Project Study No.	Reference
7	PRJEB12419	(Senghore et al., 2016)
18	PRJEB11627	(Strauß et al., 2016)
5	PRJEB12552	(Goncalves da Silva et al., 2017)
1	PRJEB12240	(Smith et al., 2016)
4	PRJEB12818	(Uhlemann et al., 2013)
1	PRJEB18560	(Edslev et al., 2018)
6	PRJEB2655 & PRJEB2944	(Jamrozy et al., 2017)
4	PRJEB2755	(Warne et al., 2016)
5	PRJEB2756	(Reuter et al., 2016)
3	PRJEB6236	(Kaas et al., 2014)
5	PRJEB7089	(Bletz et al., 2015)
2	PRJEB8084	(Mellmann et al., 2016)
15	PRJEB1915	(Moradigaravand et al., 2017)
11	PRJEB2096	(Holden et al., 2013a)
1	PRJEB2510	(Holden et al., 2013b)
18	PRJEB9575	(Moradigaravand et al., 2017)
6	PJNA306753	(Soge et al., 2016)
3	PRJEB9644	Permissions from Dr. Dorota Jamrozy
1	PRJEB11177	(Tosas Auguet et al., 2016)

6	PRJEB11281	(Nielsen et al., 2016)
3	PRJEB14187	(Julia Bünter 2016)

2.14 Best-Hit Genes Investigation

The genes of interest in Fasta format were individually compared to blast databases of nucleotides sequences of each assembly strain using an in-house python script (written by Sophia David) that employed NCBI Blastn (Altschul et al., 1990) for pairwise comparison algorithms. This produced the best hits (100% identical) genes presence and their locations in the genome for easy identification in ARTEMIS visualization. In some cases, the nucleotides sequences of the best hit genes were aligned with muscle and exported to SeaView v4.3 in order to identify mutations and truncations.

2.15 Quality Control Checks

Statistics generated by Pathogens Informatics Pipeline were used in filtering out contaminated strains. They were produced based on mapping of raw reads to TW20 reference, de novo assemblies and annotation as described in step in 2.6, and Kraken pipeline for assigning reads to taxon on mapping to bacterial databases of NCBI RefSeq.

2.16 Study Project Number

The genomes of this study together with their metadata will be deposited in European Nucleotides Archives (ENA) (www.ebi.ac.uk/ena/) under project number ERP105373.
The results for the analyses for the 23 Kenyan isolates together with phylogenetic tree are found in microreact (https://microreact.org/project/B1WId9zXQ). The results of 23 isolates in combination with public genomes are found in this microreact link (https://microreact.org/project/rk0AaZBjz).