Epidemiology and Genomic Diversity of *Staphylococcus aureus* in Humans and Pigs in Kenya

VINCENT KIPLANGAT BETT

Fitzwilliam College, University of Cambridge

Wellcome Trust Sanger Institute

This dissertation is submitted for the degree of Master of Philosophy

August, 2018

ABSTRACT

Background

Staphylococcus aureus is an important pathogen of public health concern because of the emergence of methicillin resistant and multidrug resistant strains that can colonize and cause infections in humans and animals. However, there is little knowledge of pathogen characteristics and the circulating clones in low and middle-income countries where the burden of staphylococcal disease is often high. Here, whole genome sequencing data was used to determine the presence of shared clonal lineages, antibiotic resistance, virulence genes and the phylogenetic relatedness of 23 Kenyan strains. In addition, the correlation of certain resistance and virulence genes with phylogenetic lineages and the genetic relatedness of Kenyan isolates with respect to those of other countries were assessed.

Methods

Ninety-four isolates sampled from Kiambu county, Kenya, between October 2015 and August 2016, were randomly selected for sequencing using the llumina-B HiSeq X10 150 bp platform at the Wellcome Trust Sanger Institute. The 23 genomes (from 12 farmers and 11 pigs in 15 different homesteads) that passed minimum quality control thresholds, were used for analyses, in combination with 126 public genomes from the same lineages as defined by multi-locus sequence typing (MLST) results.

Results

The collection of *Staphylococcus aureus* isolates from Kiambu county was highly diverse represented by 9 Sequence Types (STs). The four major STs are ST188, ST789, ST25 and ST580, and were present in both the hosts. There were no genetic clonal lineage or specific classes of antimicrobial resistance genes that could be associated with either host. Notably, the presence of phages that carried human immune evasion gene clusters (IECs) in the majority of the strains suggests that the colonization by *S. aureus* could be of human origin. Comparison with public genomes revealed that the Panton Valentine leucocidin (lukF/S-PV) genes and enterotoxins genes clusters (*egc*) could be conserved in ST152 and ST25 lineages respectively. Furthermore, analysis of ST580 genomes with the double locus variant, livestock-associated ST398 lineage identified ST580 (2 pigs and 1 human) strains co-segregating with human associated clade of ST398 lineage based on the distribution of their accessory genes. The global population structure analysis also showed that the humans and

pig strains of this study are related to other African genomes, suggesting that Kenyan isolates could represent lineages circulating in the continent. Taken together, this study provides the first glimpse into the genomic diversity of *S. aureus* in Kenya and highlights the need for future genomic epidemiological surveillance using large datasets sampled from multiple hosts across many other collaborating countries.

ACKNOWLEDGEMENTS

It is with great pleasure and honor to take this special opportunity to thank my fantastic supervisor David Aanensen for his continual advice, encouragement and generous support throughout my Masters project. I feel privileged to be with his excellent team in learning and getting guidance in world of genomics. I extend my deepest gratitude to Silvia Argimon for her enthusiastic day-to-day guidance, attention to details and her commitment in assisting me to end this Master. I am also greatly indebted to Monica Abrudan and Sophia David for their kind guidance and generous time in transferring bioinformatics skills to me. I owe a great debt of gratitude to thesis committee members Mark A. Holmes, Sam Kariuki and Stephen Bentley for their expertise insights in genomics.

I am grateful to Wellcome Trust Sanger Institute Masters studentship, graduate office for giving an opportunity to realize not only my childhood dreams of studying abroad but also being in top globally excellent University. Special thanks to Annabel Smith, Christina Hedberg-Delouka and Sanger committee of graduate studies for their consistent advises, support and close monitoring of my progress.

I am greatly indebted to KEMRI staffs led by John Ndemi Kiiru for the generation of the study isolates and carrying out the DNA extraction. Many thanks to WTSI core sequencing department for library preparation and sequencing of the samples. I would like to express much appreciation to Pathogen Informatics team for their pipeline tools that were used in most analyses of this thesis. Many thanks to WTSI IT department for their services offered during the course of the study. I appreciate insights feedback of Matthew Holden and Ewan Harrison.

My deepest appreciation goes to Centre for Genomics Pathogen Surveillance team; Carol Churcher, Dawn Muddyman, Victoria Cohen, Corin Yeats, Ben Taylor, Simon Harris, Khalil AbuDahab and Richard Goater, who were very instrumental in my stay and learning at Sanger. I am very grateful to Fitzwilliam College for the accommodation and organizing social event activities which gave me good experiences about norms, culture and traditions of the University. Special thanks to my college tutor Jonathan Cullen and Fitzwilliam college chapel priest Revd Helen Arnold for their concern support during the terms.

Finally, I would like to offer my special thanks to my mother Gladys Chebet, family members, relatives and friends for their continual encouragement, prayers and support. And

special dedication of this thesis to my grandmother Grace Byomdo for her constant blessings and support during my childhood life.

DECLARATION

I declare that this thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text.

I declare this thesis is not substantially the same as any that I have submitted or is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the preface and specified in the text. I further state that no substantial part of dissertation has already been submitted or is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution except as declared in the Preface and specified in the text

I declared that this thesis does not exceed 20,000 words prescribed in Special Regulation for MPhil in Biological Science.

Sign …….. Date ……..

PREFACE

The collection of study samples and DNA extraction were carried out by our collaborators at KEMRI led by Dr. John Ndemi Kiiru, KEMRI Kenya. WTSI core sequencing department completed sequencing of the DNA of the sampled strains. The analyses for the results of this dissertation were done by myself using mostly the pipeline developed by WTSI Pathogen Informatics pipeline as referenced in the texts. Public genomes that were included to extend the study were all downloaded by CGPS team except 6 strains which I retrieved myself from NCBI and 3 genomes that were donated by Dr. Dorota Jamrozy.

Table of Contents

LIST OF ABBREVIATIONS

ACCTRAN	Accelerated Transformation
AMR	Antimicrobial Resistance
AST	Antibiotic Susceptibility Tests
BAM	Binary Alignment Map
CC	Clonal Complex
CDS	Coding Sequences
DNA	Deoxyribonucleic Acid
G + C	Guanine-Cytosine
HIV	Human Immunodeficiency Virus
STs	Sequence Types
KEMRI	Kenya Medical Research Institute
Mb	Megabytes
MLST	Multi-locus Sequence Types
MSSA	Methicillin Susceptible *Staphylococcus aureus*
MRSA	Methicillin Resistant *Staphylococcus aureus*
NCBI	National Centre for Biotechnology
PVL	Panton-Valentine leucocidin
RAxML	Randomized Axelerated Maximum Likelihood
SAM	Sequence Alignment Map
SNPs	Single Nucleotides Polymorphisms
Spa	Protein A
VCF	Variant Call Format
WHO	World Health Organization
WGS	Whole Genome Sequencing
WTSI	Wellcome Trust Sanger Institute
GFF	General File Format

LIST OF FIGURES

LIST OF TABLES