

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 Illumina-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 Pantone 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

ABSTRACT	III
ACKNOWLEDGEMENTS	VI
DECLARATION	VIII
PREFACE	IX
TABLE OF CONTENTS	X
LIST OF ABBREVIATIONS	XIII
LIST OF FIGURES	XIV
LIST OF TABLES	XVI
CHAPTER ONE	1
INTRODUCTION.....	1
THESIS SUMMARY	1
1.1 General Characteristics of the Genus <i>Staphylococcus</i>	1
1.2 Colonization, Infections and Diseases Caused by <i>S. aureus</i>	2
1.3 Treatment and Antibiotic Resistance	2
1.4 <i>S. aureus</i> as a Significant Public Health Problem in Africa.....	5
1.5 One Health Concept	6
1.6 Transmission of <i>S. aureus</i> Between Humans and Livestock	7
1.7 Conventional Molecular Typing Tools	9
1.8 Whole Genome Sequencing Technologies and Application.....	10
1.9 <i>S. aureus</i> Genome and Population Structure	13
1.10 Virulence Factors of <i>S. aureus</i>	14
1.11.1 Adherence factors (Surface Proteins).....	14
1.11.2 Exotoxins (Extracellular Enzymes).....	15
1.11.a The Staphylococcal Enterotoxins (SEs).....	15
1.11.b Immune Evasion Cluster (IEC) Genes	16
1.11.c Panton-Valentine Leukocidin	16
1.12 OBJECTIVES OF THE STUDY	17

CHAPTER TWO.....	19
METHODS	19
2.1 Ethical Considerations.....	19
2.2 Study Design and <i>S. aureus</i> Isolation	19
2.3 Antibiotic Susceptibility Testing	20
2.4 DNA Extraction	21
2.5 Library Preparation and Whole Genome Sequencing.....	21
2.6 <i>S. aureus</i> Assembly and Annotation	21
2.7 Mapping and Variant Calling	23
2.7.1 Using Pathogens Informatics pipeline.....	23
2.7.2 Using a Custom Script.....	24
2.8 <i>In Silico</i> Prediction of Recombination Regions in Pseudogenome Alignments	25
2.9 Phylogenetic Reconstruction	25
2.10 Determination of Antimicrobial Resistance and Virulence Genes using ARIBA	26
2.11 Multi-Locus Sequence Typing (MLST) and <i>spa</i> Typing	28
2.12 Pan Genome Analyses of ST398 and ST580 strains.....	28
2.13 Public Genomes Selection.....	29
2.14 Best-Hit Genes Investigation.....	30
2.15 Quality Control Checks.....	30
2.16 Study Project Number	30
CHAPTER THREE	32
RESULTS	32
OVERVIEW	32
SECTION A.....	32
3.1.1 Study Population.....	32
3.1.2 Genomic Metrics.....	32
3.1.3 <i>In Silico</i> Prediction of Multi-locus Sequence Types and <i>spa</i> typing	36
3.1.4 Phylogenetic Analyses	38

3.1.5	<i>In Vitro</i> Antibiotic Susceptibility Test Results.....	40
3.1.6	<i>In Silico</i> Prediction of Antimicrobial Resistance Genes	41
3.1.7	<i>In Silico</i> Prediction of Virulence Genes.....	43
SECTION B		46
3.2.1	Relatedness of Kenya Isolates with Publicly Available Genomes	46
3.2.2	Detailed Phylogenetic Analyses of Selected Sequence Types (STs).....	49
3.2.3	Phylogenetic Analyses of ST580 and Livestock-associated ST398 lineage ...	55
3.2.4	Accessory Genomes Analyses of ST580 and ST398.....	56
CHAPTER FOUR.....		59
DISCUSSION		59
CHAPTER FIVE.....		65
CONCLUSIONS.....		65
5.1 FUTURE PERSPECTIVES OF THE STUDY.....		66
5.1.1	High Resolution and Accuracy of Whole Genome Sequencing (WGS).....	66
5.1.2	Greater Genetic Diversity of Clonal Lineages.....	66
5.1.3	Sharing of Clonal Lineages Between Pigs and Humans	66
5.1.4	Prediction of Antimicrobial Resistance Genes using WGS	67
5.1.5	Genetic Relatedness of Kenyan Isolates with Strains of Other African Countries	67
APPENDICES:.....		69
REFERENCES:.....		80

LIST OF ABBREVIATIONS

ACCTAN	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 <i>Staphylococcus aureus</i>
MRSA	Methicillin Resistant <i>Staphylococcus aureus</i>
NCBI	National Centre for Biotechnology
PVL	Panton-Valentine leucocidin
RAxML	Randomized Accelerated Maximum Likelihood
SAM	Sequence Alignment Map
SNPs	Single Nucleotides Polymorphisms
<i>Spa</i>	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

Fig. 1: Map of Kenya _____	20
Fig. 2: A schematic diagram summarizing the assembly _____	22
Fig. 3: Graphs showing distribution of 94 genomes before the QC filtering parameters were applied. _____	34
Fig. 4: A schematic diagram of a summary of general flow description of species identification and contamination of sequence data _____	35
Fig. 5: Prevalence of STs (A) and distribution of STs per host (B) _____	37
Fig. 6: <i>Spa</i> type distribution between humans and swine. _____	37
Fig. 7: Midpoint rooted maximum likelihood phylogenetic tree based on 57010 core genome SNPS _____	38
Fig. 8: Histogram showing phenotypic distribution of antibiotic drugs between pigs and humans _____	40
Fig. 9: A midpoint rooted maximum likelihood phylogenetic tree with heatmap distribution of AMR genes and heavy metals _____	42
Fig. 10: A Midpoint rooted phylogenetic tree with a heatmap of virulence genes _____	43
Fig. 11: Artemis visualization of enterotoxins gene clusters in ST25 (A) and ST22 (B) _	44
Fig. 12: Regression lines of each ST in terms of SNPs pairwise difference over time (in years) in relation to TW20 (ST239). _____	47
Fig. 13: A Midpoint rooted maximum likelihood phylogenetic tree that has been optimized with ACCTRAN algorithms _____	48
Fig. 14: Phandango visualization of 2 regions of repetitive regions (Red) in pseudogenome alignment of ST188 _____	50
Fig. 15: Midpoint rooted phylogenetic tree of ST188 based on 2552 core genome SNPs on mapping to HongKong draft genome reference _____	51
Fig. 17: Midpoint rooted ML phylogenetic tree of ST6 based on 1283 core genome SNPs on alignment to <i>Staphylococcus aureus</i> TW20 reference _____	52
Fig. 18: Midpoint rooted ML phylogenetic tree of ST152 based on 1446 core genome SNPs on alignment to <i>Staphylococcus aureus</i> BB155 reference _____	53
Fig. 19: Multiple sequence alignment of lukF-PV showing non-synonymous SNPs _____	54

Fig. 20: Maximum Likelihood phylogeny of ST398 and ST580 based on 8575 core genome SNPs that has been optimized with ACCTRAN algorithms _____ 55

Fig. 21: Heatmap based on the number of shared genes in the accessory genomes _____ 57

LIST OF TABLES

Table 1: Different references used in mapping and SNP-calling analyses	23
Table 2: Quality control thresholds for SNPs calling	24
Table 3: Filtering parameters for Custom script.....	25
Table 4: Default parameters in ARIBA	27
Table 5: Showing heavy metals and antiseptic	27
Table 6: Showed number of strains of public genomes	30
Table 7: Quality control filtering parameters	34
Table 8: Genome metric characteristics	34
Table 9: showing pairing of strains in the homesteads	35
Table 10: Different types of IECs with their prevalence	45
Table 11: shows SNPs effect on ACCTAN reconstruction of phylogeny	49

