DECODING BACTERIAL GENOME WITH HIGH-THROUGHPUT SEQUENCING: GENES AND GENETIC MARKERS

XIA ERYU

(B.S., UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA)

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

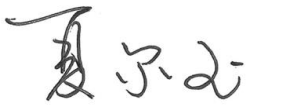
NUS GRADUATE SCHOOL FOR INTEGRATIVE SCIENCES AND ENGINEERING
NATIONAL UNIVERSITY OF SINGAPORE

2016

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.



Xia Eryu

2016

Acknowledgements

Finally, it is the day that I can write the acknowledgements down. I cannot resist my tears when looking back. Such a lot of things I feel grateful for. Such a lot of people I should thank.

Thank nature for showing its tremendous beauty which has attracted our eyes, for having much mystery which has fueled our impetus to explore, and for leaving hints here and there which has flattened our road to the truth.

Thank all the devotions of previous scientists who have raised the curtain of science, paved the road, and provided their shoulders for us to stand on.

Thank all those who have contributed to NCBI. Such a collaborative way of knowledge accumulation would enable the preservation of abundant data sources, on which the majority of my research was based.

Thank NGS for its effort to facilitate interdisciplinary research, which has enabled students from different background to communicate more, to think from other perspectives when facing a problem, and to find solutions with the help from others.

Thank USTC for forcing me to learn a lot of math and computing during my undergraduate study. At that time, I had no idea why I would need these as a biologist. But now, I cannot imagine how I can ever become a biostatistician without learning these. Thank those talented and industrious students in USTC who have inspired me to think more and work more.

Thank my supervisor YY for giving me, a girl with no experience in computing or statistics, a chance to try my hands in his lab, for always offering generous help when I need advice or support, and for offering me the freedom to explore what I'm interested in.

Thank Rick for his valuable help when I need advice, for his time when I need discussion, and for helping with my paper writing. I should also thank Rick for tolerating my aggressiveness, since I have always been very stubborn in research.

Thank my collaborators for offering me opportunities to get exposure to knowledge in different disciplines, for helping me understand the needs of clinicians and biologists in order to make my work more meaningful.

Thank my group members for their company. We all know how hard it is to do science, but we stand shoulder by shoulder along the way of pursuit. Thank Xuanyao, a girl quite similar to myself, for always being a considerate friend. Thank Zhou Jin, a girl who is direct and sincere, for sharing her experience about research, life, and work, and for having breakfast with me. Thank Wang Xu, Woei-Yuh, Wenting, Ruoying, Xinlu... for bringing so much happiness that has made my days colorful.

Special thanks go to my parents and friends for preserving me a place where I have no need to pretend to be a cheerful woman scientist, but can be the real me: a timid, sensitive, and sentimental girl. Thank my parents for their unconditional support whatever I choose to do. Thank Shuyuan for always spending time with me on Skype. Thank Yuxi for being a friend that I can bank on. Thank you for your love. I love you so much.

I should also thank myself for the persistent pursuit of science, for the unfailing child-like curiosity, and for always being optimistic. I'll carry on whatever is in front.

Table of Contents

Summary

Advancement in sequencing technology, which has significantly increased the throughput and decreased the cost, has made sequencing accessible to more clinical microbiology laboratories for both infection control and public health purposes. Some advantages of sequencing over traditional microbiology methods include providing more comprehensive information at a higher resolution in a single procedure, ability to make quick diagnoses and save human labor. In the thesis, my attempt to decode bacterial genome with high-throughput sequencing is summarized from two perspectives: genes and genetic markers.

Constructing a phylogenetic tree is one of the most useful tools for studying the evolutionary history of bacteria, and this genetic inference can be adversely affected by genetic recombination. In Chapter 3, I introduce and describe ReRCoP, a novel method for efficient identification and removal of recombination in large bacterial samples for accurate phylogenetic inference. The global dissemination of antibiotic resistance genes has posed a significant public health threat. In Chapter 4, the global dissemination and local transmission of the bla_{NDM} gene, which is capable of causing resistance to a broad range of beta-lactam antibiotics and of spreading to a wide range of Gram-negative bacteria, are examined at the genomic level to identify the means of dissemination which could provide insights for containment of its spread. New genes are continually emerging and discovered in bacteria, some offering increased fitness to survival while some causing antibiotic resistance. The emergence of new genes has been attributed to gene duplication and divergence. In Chapter 5, a new model called the IAID (Innovation-

Amplification-Innovation-divergence) model is proposed to explain gene evolution via duplication. Genetic markers have been widely used for bacterial molecular typing. In Chapter 6, SpoTyping, a fast and accurate *in silico* spoligotyping method for *Mycobacterium tuberculosis* from sequencing reads is described that can be used for fast disease diagnosis and correlating recent outbreaks with historical isolates.

In summary, the utility of high-throughput sequencing has been demonstrated in bacteria genomics study.

List of Tables

List of Figures

List of Abbreviations

CPE: carbapenemase-producing *Enterobacteriaceae*

CRE: carbapenem-resistant *Enterobacteriaceae*

CRISPR: clustered regularly-interspaced short palindromic repeats

DBSCAN: density-based spatial clustering of applications with noise

DR: direct repeat

HGT: horizontal gene transfer

kNN: k-nearest neighbors

MIRU-VNTR: mycobacterial interspersed repetitive units - variable numbers of tandem repeat

MLST: multi-locus sequence type

Mtb: *Mycobacterium tuberculosis*

NGS: next-generation sequencing

PCR: polymerase chain reaction

RFLP: restriction fragment length polymorphism

ROC: receiver operating characteristic

SMRT: single molecule, real-time

SNP: single-nucleotide polymorphism

TB: tuberculosis

Glossary

Antibiotics A type of antimicrobial for treating and preventing bacterial infection.

Antimicrobial An agent that can kill microorganisms or inhibit their growth.

Bacteria Microscopic single-celled organisms that live in enormous numbers broadly on Earth.

Bootstrapping A method to determine the confidence levels about the topology of an inferred phylogenetic tree using bootstrap resampling technique.

Carbapenem A class of broad-spectrum beta-lactam antibiotics, which is active against many bacteria by inhibiting cell wall synthesis.

Carbapenemase A class of enzymes produced by bacteria that can hydrolyze carbapenem antibiotics and thus provide resistance to them.

CRISPR Segments of prokaryotic DNA containing short sequence repetitions interspersed by short spacer DNA sequences from previous exposures to a bacterial virus or plasmid, which can be used to type bacteria like *Salmonella* and *Mycobacterium tuberculosis.*

Core genome A concatenation of the set of genes present in all members of the studied population.

DBSCAN A density-based clustering algorithm, which groups points in the high-density regions together while making points in the low-density regions outliers.

DNA A nucleic acid molecule that carries most of the genetic information used in the development, functioning and reproduction of all known living organisms and many viruses.

Genome The set of genetic information contained in an organism.

Homoplasy A phenomenon that identical character states (for example, the same nucleotide in genetic sequences) are not a result of shared ancestry but a result of convergent evolution from different ancestors.

Horizontal gene transfer (HGT) The process of genetic information transfer from one organism to another that is not its descendent.

kNN A useful, non-parametric method commonly used for classification and regression.

MIRU-VNTR A bacterial typing scheme for *Mycobacterium tuberculosis* complex, which classify bacterial strains by analyzing the variable number of tandem repeats.

MLST A technique for bacterial typing, which classifies bacterial isolates based on distinct alleles of internal fragments of multiple housekeeping genes.

Pan genome A concatenation of the set of genes present in at least one member of the studied population.

Phylogenetic tree A branching diagram representing the inferred phylogeny based on physical or genetic characteristics, where the taxa joined together are implied to have a common ancestor.

Phylogenetics The study of evolutionary history and relationships among a group of genetically related organisms.

Phylogeny The evolution of a group of genetically related organisms.

Plasmid A small, extra-chromosomal DNA that replicates independently and carries genes not essential but beneficial to the survival of the organism.

PCR A technology used in molecular biology to amplify one or a few copies of a piece of DNA to generate thousands to millions of copies of the DNA.

Receiver operating characteristic (ROC) A graphical plot created by plotting the true positive rate against the false positive rate of a binary classifier at various discrimination threshold settings to show the performance of the classifier under different thresholds.

Sequencing A process to determine the genetic sequence of a DNA.

Single-nucleotide polymorphism (SNP) A variation affects a single base pair between two DNAs.

Enterobacteriaceae A family of Gram-negative bacteria that includes, along with many harmless symbionts, many familiar pathogens, such as *Salmonella*, *Escherichia coli*, *Enterobacter cloacae*, *Klebsiella pneumonia*, and *Yersinia pestis*.

Enterobacter aerogenes A Gram-negative, rod-shaped bacterium of the family *Enterobacteriaceae* and the genus *Enterobacter* that causes opportunistic infections.

Enterobacter cloacae A Gram-negative, rod-shaped bacterium of the family *Enterobacteriaceae* and the genus *Enterobacter* that is commonly found in the guts of warm-blooded organisms and can cause infections at times.

Escherichia coli A Gram-negative, rod-shaped bacterium of the family *Enterobacteriaceae* and the genus *Escherichia* that is commonly found in the guts of warm-blooded organisms and can cause infections at times.

Klebsiella oxytoca A Gram-negative, rod-shaped bacterium of the family *Enterobacteriaceae* and the genus *Klebsiella* that can cause antibiotic-associated hemorrhagic colitis.

Klebsiella pneumoniae A Gram-negative, rod-shaped bacterium of the family *Enterobacteriaceae* and the genus *Klebsiella* that can cause destructive changes to human lungs.

Klebsiella variicola A Gram-negative, rod-shaped bacterium of the family *Enterobacteriaceae* and the genus *Klebsiella* that was previously regarded as a phylogroup of *Klebsiella pneumoniae*.

Mycobacterium tuberculosis A pathogenic bacterium of the family *Mycobacteriaceae* and the genus *Mycobacterium* that is the causative agent of tuberculosis.

Raoultella ornithinolytica A Gram-negative, rod-shaped bacterium of the family *Enterobacteriaceae* and the genus *Raoultella* that causes rare human infections.

Streptococcus pneumoniae A Gram-positive, lancet-shaped bacterium of the family *Streptococcaceae* and the genus *Streptococcus* that is a major cause of pneumonia.

Publications

The thesis is based on the following papers or manuscripts:

- **Xia E**, Teo YY, and Ong RT. "SpoTyping: fast and accurate in silico Mycobacterium spoligotyping from sequence reads." *Genome Medicine* 8.1 (2016): 1.

- Khong WX[*], **Xia E**[*], Marimuthu K[*], Xu W, Teo YY, Tan EL, Neo S, Krishnan PU, Ang BS, Lye DC, Chow AL, Ong RT, Ng OT. "Local Transmission and Global Dissemination of New Delhi Metallo-Beta-Lactamase (NDM): A Whole Genome Analysis." *BMC Genomics* 17.1 (2016): 1.

- Poh WT[*], **Xia E**[*], Chin-inmanu K, Wong LP, Cheng AY, Malasit P, Suriyaphol P, Teo YY and Ong RT. "Viral quasispecies inference from 454 pyrosequencing." *BMC bioinformatics* 14.1 (2013): 1.

- **Xia E,** Teo YY and Ong TR. "ReRCoP: core genome phylogeny of large bacterial population samples with recombination removal." *Manuscript in preparation.*

- **Xia E,** Teo YY and Ong TR. "Gene evolution by duplication: innovation, amplification, innovation and divergence." *Manuscript in preparation*.

Publications not included in this thesis:

- **Xia E**, Khong WX, Marimuthu K, Xu W, Ong RT, Tan EL, Krishnan PU, Ang BS, Lye DC, Chow AL, Teo YY, Ng OT. "Draft Genome Sequence of a Multidrug-Resistant New Delhi Metallo-β-Lactamase-1 (NDM-1)-Producing Escherichia coli Isolate Obtained in Singapore." *Genome Announcements* 1.6 (2013): e01020-13.

- Chee CB, Gan SH, Ong RT, Sng LH, Wong CW, Cutter J, Gong M, Seah HM, Hsu LY, Solhan S, Ooi PL, **Xia E**, Lim JT, Koh CK, Lim SK, Lim HK, and Wang YT. "Multidrug-resistant tuberculosis outbreak in gaming centers, Singapore, 2012." *Emerging Infectious Diseases* 21.1 (2015): 179.

- Marimuthu K, Ng OT, Khong WX, **Xia E**, Teo YY, Ong RT, Lye DC, Chow AL, Krishnan P, Ang BS. "Reactive Infection Control Strategy for Control of New Delhi Metallo-β-Lactamase (NDM)-Producing Enterobacteriaceae Analyzed Using Whole-Genome Sequencing: Hits and Misses." *Infection control and hospital epidemiology* (2016): 1.

- Marimuthu K, Ng OT, Khong WX, **Xia E**, Xu W, Teo YY, Tan EK, Ong RT, Lye DC, Chow AL, Krishnan P and Ang BS. "Halting NDM-producing enterobacteriaceae spread with the reactive infection control strategy: a real-world experience analyzed using a novel spatiotemporal epidemiologic risk measure (Epi-score) and whole-genome sequencing (WGS)." *Antimicrobial Resistance and Infection Control* 4.1 (2015): 1.

- Khong WX, Marimuthu K, Teo J, Ding Y, **Xia E**, Lee JJ, Ong RT, Venkatachalam I, Cherng B, Kaur PS, Choong WL, Smitasin N, Ooi S, Deepak R, Kurup A, Fong R, La MV, Tan TY, Koh TH, Tzer R, Tan EK, Krishnan P, Singh S, Pitout J, Teo YY, Yang L, Ng OT. "Tracking Inter-Institutional Spread of NDM and Identification of a Novel NDM-positive Plasmid, pSg1-NDM, Using Next-Generation Sequencing Approaches." *Journal of Antimicrobial Chemotherapy* (2016): dkw277.

- Teo JQ, Ong RT, **Xia E**, Koh TH, Lee SJ, Lim TP, Kwa AL. "Leading the way to a pre-antibiotic era: co-carriage of plasmid-mediated carbapenem and polymyxin resistance determinants." *Accepted.*

- Faksri K, Tan JH, Disratthakit A, **Xia E**, Prammananan T, Suriyaphol P, Teo YY, Ong RT, Chaiprasert A. "Whole-genome sequencing analysis of serially isolated multi-drug and extensively drug resistant Mycobacterium tuberculosis from Thai patients." *Accepted.*

- Faksri K*, **Xia E***, Tan JH, Teo YY, Ong RT. "RD-Analyzer: in silico regions of difference (RD) typing of Mycobacterium tuberculosis complex from sequence reads." *Under review.*

* These authors contributed equally to this work.

Chapter 1

Introduction

1.1 Introduction to sequencing technologies

1.1.1 First-generation sequencing

The beginning of first-generation sequencing is marked by the chain-termination method developed by Sanger and Coulson in 1975 [1, 2], and a chemical sequencing method developed by Maxam and Gilbert [3] around the same time. Sanger sequencing, also known as enzymatic DNA sequencing, has been the most prevalent method, which is still used today. Sanger sequencing is based on using DNA polymerase to selectively incorporate chain-terminating dideoxynucleotides during in vitro DNA replication. Each reaction in Sanger sequencing can produce a sequence read of up to 800 to 1,000 base pairs (bp) in length. Sanger sequencing has the advantages of having high read accuracy and long read length, while suffering from the disadvantages of low throughput, high cost per base, and inefficiency in detecting low frequency variants compared to new generations of sequencing. Maxam-Gilbert sequencing is based on nucleobase-specific partial chemical modification of DNA followed by cleavage of the DNA backbone at sites adjacent to the modified nucleotides, and has become less favored due to technical complexity, extensive use of hazardous chemicals and difficulties to scale up.

1.1.2 Next-generation sequencing

Next-generation sequencing (NGS) is also known as second-generation sequencing and has the widest applications among all sequencing technologies in the current genomics study. The past 10 years have witnessed dramatic

improvements in NGS technology, which included the significant increase in sequencing throughput and the rapid drop in sequencing cost.

The year 2004 marks the beginning of NGS by having the first NGS equipment available. Since then, many new sequencing platforms have been introduced such as the 454, SOLiD, Illumina, Ion Torrent PGM, Ion Proton, and so on. Different from classical sequencing methods that amplify one amplicon from one sample and produce a single sequence, NGS chemistries have the amplicons amplified clonally, separated spatially and read in cyclic parallel [4].

Several steps of DNA sequencing are shared regardless of the platform: library preparation, clonal amplification, and sequencing chemistry.

The first step is library preparation. The DNA sample to be sequenced is first fragmented into pieces either with mechanical forces like sonication or nebulization or by enzymatic digestion. The target fragment size varies depending on the platform and chemistry and can be selected with gel or beads. Short adaptors, which provide priming sequences for amplification and sequencing, are then ligated to the ends of the fragments. If multiplexing is needed, barcode sequences are also ligated to provide information about the DNA identity. If a mate pair library is to be prepared, apart from the adaptors and barcodes mentioned above, an internal adaptor is used to separate two DNA fragments.

The second step is clonal amplification. Each fragment in the prepared library needs to be amplified clonally before sequencing to enhance the signal in the sequencing process for accurate detection. Two approaches are available:

bridge PCR used by Illumina and emulsion PCR introduced by 454 Life Sciences and used also by SOLiD, Ion Torrent PGM and Ion Proton.

The next step is central to sequencing: the sequencing chemistry that performs base interrogation on all DNA fragments in parallel and detects signals that are later translated into DNA bases. Different platforms have different sequencing chemistry. Several examples are: 454 pyrosequencing, Illumina sequencing by synthesis, SOLiD sequencing by ligation, and Ion Torrent semiconductor sequencing. Pyrosequencing determines the DNA sequence based on the light emitted upon incorporation of the next complementary nucleotide. It detects the activity of DNA polymerase with another chemoluminescent enzyme. Illumina sequencing by synthesis uses only DNA polymerase, and is based on reversible dye-terminators that enable the identification of single nucleotides as they are introduced into DNA strands. SOLiD sequencing by ligation does not use DNA polymerase but uses DNA ligase, whose preferential ligation for matching sequences results in a signal to identify the nucleotide on a given position. Ion Torrent semiconductor sequencing is based on the detection of pH alteration caused by hydrogen ions that are released during the polymerization of DNA, which is different from the optical methods used in other sequencing systems.

Apart from the merits, which include high throughput, high accuracy, and low cost, NGS has two major weaknesses, which are: (1) the read length is shorter compare to Sanger sequencing; and (2) the use of PCR can introduce bias in the amplification process [5].

1.1.3 Third-generation sequencing

No consensus has been reached on the definition of third-generation sequencing (also known as next-next-generation sequencing). It has been suggested that single molecule sequencing without the need to halt, enzymatically or otherwise, between read steps should be called the third-generation sequencing, where each read represents the sequence of a single molecule of DNA [6]. The technologies fall into three main categories: (1) sequencing by synthesis, where single molecules of DNA polymerase are monitored as a single molecule of DNA is synthesized; (2) sequencing with nanopores, where single molecules are directed through or positioned next to a nanopore and are sequenced base by base as they pass the nanopore; and (3) sequencing by direct imaging, where advanced microscopies are used to sequence individual DNA molecules [6]. The single-molecule real-time (SMRT) sequencing developed by Pacific Biosciences represents a first third-generation technique that has been applied to genomics study. SMRT sequencing is marked by two key innovations: zero-mode waveguides which allow light to illuminate only the bottom of a well where a DNA polymerase/template complex is immobilized, and phospholinked nucleotides which allow observation of the immobilized complex when the DNA polymerase produces a completely natural DNA strand. While the long sequencing reads and rapid turnaround time attracts great attention to third-generation sequencing, efforts are still needed to increase the throughput, increase the read accuracy and decrease the cost.

1.1.4 Next-generation sequencing, high-throughput sequencing and whole genome sequencing

While the title of the thesis defines its scope to high-throughput sequencing, next-generation sequencing and whole genome sequencing are also terms frequently referred to. The three terms, while all used to describe sequencing technologies, view technologies from different perspectives. Next-generation sequencing, as elaborated above, refers to the sequencing technologies that are developed during a time period, and perform sequencing by having amplicons clonally amplified, spatially separated and read in cyclic parallel. High-throughput sequencing was initially coined to describe the first commercial 96 capillary sequencers, but the concept has changed as the sequencing throughput increases with time [7]. It is now used to refer to sequencing technologies that outperform Sanger sequencing in their daily throughput, which include both next-generation sequencing and third-generation sequencing. Whole genome sequencing, different from the two mentioned above, has little to do with the sequencing technology. Also known as full genome sequencing, complete genome sequencing, and entire genome sequencing, whole genome sequencing refers to any process that determines the DNA sequences of an organism's genome in a single procedure. In bacteria studies, this entails sequencing of the bacterial chromosomal DNA as well as the extra-chromosomal DNA such as the plasmid DNA. Though not in itself a technical term, whole genome sequencing has been made easier as sequencing throughput increases, which is made possible by high-throughput sequencing techniques like next-generation sequencing.

1.2 Introduction to bacteria genomics

1.2.1 Bacteria

Bacteria are microscopic single-celled prokaryotic organisms, which live in enormous numbers and constitute a large domain of prokaryotic microorganisms. They are found in every habitat on Earth, and some live in other organisms like plants and animals including humans. There are a lot of bacterial cells in the human body, with the largest number in the human gut. While the majority of the bacteria in human body are harmless or even beneficial to our health, some species are pathogenic and can cause infectious disease. Several examples of bacteria that are found in human body are: *Escherichia coli* (*E. coli*), which is commonly found in the human gut and can cause infections at times; *Klebsiella pneumoniae* (*K. pneumoniae*), which can cause destructive changes to the human lung; and *Mycobacterium tuberculosis* (*Mtb*), which is the causative agent of tuberculosis (TB).

1.2.2 Bacterial genome

Bacteria have simple cell structures. There is neither nucleus nor membrane-bound organelles, and the genetic information is usually carried by a single loop of chromosomal DNA. For some bacteria, there are extra-chromosomal DNAs called plasmids. Bacterial genome, defined as the complete set of genetic information, thus includes both chromosome(s) and plasmids.

Unlike most eukaryotes whose DNAs are linear, most bacteria have a single circular chromosome, the size of which ranges from about 0.13 million base pairs (Mbp) as symbionts in nutrient-provisioning environment in several insect lineages [8] to over 14 Mbp [9] due to genome expansion in different

environmental conditions. The genome of *E. coli* is about 5.1 Mbp with about 4,900 genes. The genome of *K. pneumoniae* is about 5.6 Mbp with about 5,500 genes. The genome of *Mtb* is about 4.4 Mbp with about 4,000 genes. These bacterial genomes are only about 0.1% the size of the human genome, while having about 10% as many genes. This is a result of the differences between bacterial chromosome and human chromosome from three perspectives: (1) bacterial genes, on average, have fewer codons than human genes; (2) bacterial genes have no introns; and (3) length of non-coding DNA between bacterial genes is shorter.

Plasmids are extra-chromosomal DNAs that are usually circular, self-replicating, and play important roles in maintaining and disseminating novel genetic elements in the bacterial population. Plasmids carry genes encoding adaptive traits such as antibiotic resistance, pathogenesis, or the ability to exploit new environments or compounds. Bacterial chromosomes, as they represent features necessary for the survival of bacteria, show a relatively high conservation of the structure with many universally shared genes. Plasmids, on the other hand, are more variable in terms of the gene content and gene organization, even at very short genetic distances [10].

1.2.3 Genomic features of bacteria

Horizontal gene transfer (HGT), an important mechanism for the evolution of microbial genomes, refers to the transfer of genetic material to a non-offspring cell, which is different from vertical gene transfer that passes genetic material from an ancestor to a descendent. Mobile genetic elements like plasmids, bacteriophages and pathogenicity islands can mediate HGT that transfers

genes often involved in infection [11]. There are different mechanisms explaining HGT: transformation, transduction and conjugation [11]. Transformation causes genetic alteration by directly uptaking and incorporating foreign DNA from its surroundings through the cell membrane. Transduction causes genetic alteration by introducing foreign DNA via a virus or viral vector. Bacterial conjugation causes genetic alteration by transferring DNA between bacterial cells via direct contact or via a bridge-like connection between two cells.

Apart from the traditional view that prokaryotes evolve by clonal divergence and periodic selection, bacterial genome evolution is shaped by three main forces: gene acquisition via HGT, gene loss by deletion events and gene change like mutations or rearrangements [12]. Different bacterial pathogens adopt different scale of the forces, leading to different genomic dynamics. Three main genomic dynamics have been reported: (1) Some bacteria have genetically uniform lineages. These are usually reproductively isolated bacteria, for example, *Mtb* and *Bacillus anthracis*, and are thus "clonal" in the genome evolution. (2) Some bacteria recombine extensively between closely related sequences in closely related strains. These are usually competent mucosal pathogens by nature, for example, *Haemophilus influenza* (*H. influenza*) and *Streptococcus pneumoniae* (*S. pneumoniae*). (3) Some bacteria have widespread HGT that introduces genetic sequences into the genome, thus bringing in large blocks of foreign gene sequences in a single event. This is common in certain pathogens like many enterobacteria, some staphylococci and streptococci [11].

1.2.4 Bacteria genomics

NGS has become widely used for clinical microbiology research due to improvements that have made it faster, cheaper and more accurate, and can now replace many laboratory tests with a single sequencing run. Three tasks essentially performed by NGS are: (1) species identification of a bacterial isolate; (2) determination of properties such as antibiotic resistance and virulence; and (3) detect the emergence and control the spread of pathogens [13]. Various studies have been conducted, which have showcased the application of NGS in bacteria genomics on species like *Clostridium difficile* [14], *E. coli* [15–17], *K. pneumoniae* [18], Methicillin-resistant *Staphylococcus aureus* (MRSA) [19, 20], and so on. Some other researches have focused on metagenomics problems like identifying mixed infections [21, 22], investigating intra-host bacteria diversity [23] and assembling genomic sequences from metagenomics data [24] , which studied the bacteria communities.

Traditional laboratory tests are usually multiple-step, labor-intensive, complex and sometimes slow, which may take days for fast-growing bacteria like *E. coli* and months for slow-growing bacteria like *Mtb*. Genomics approaches with NGS, however, enable the results to be achieved in a single step after culturing and sequencing. Moreover, they can provide information not achievable with current molecular typing methods, which are usually of single-nucleotide resolution.

1.3 Introduction to basic bioinformatics approaches in bacteria genomics

1.3.1 Sequencing data format, quality control, and pre-processing

FASTA format is a text-based format in bioinformatics to represent nucleotide or peptide sequences, in which each sequence begins with a description line distinguished by '>' at the beginning, followed by lines of sequences where each nucleotide or amino acid is represented by a single letter. FASTQ format is a text-based format that bundles a FASTA sequence with its quality data, which is the current standard format of raw reads in high-throughput sequencing. Each sequence in a FASTQ file has four lines, where: (1) the first line begins with '@', and bears the sequence identifier and description; (2) the second line is the sequence read; (3) the third line begins with '+', and is optionally followed by the same information as in the first line; and (4) the fourth line encodes the respective quality values for each character in the sequence read in the second line.

Several metrics can be used for quality control of the raw sequencing reads, which can usually be computed with FastQC [25]. The first thing to consider is the quality scores in the FASTQ file, where low quality scores indicate low sequencing quality and less reliable reads. Another important thing to inspect is the presence of contamination from sequencing adapters, PCR primers, contaminant DNA and other artifacts. Pre-processing needs to be conducted when quality issues like low quality scores, adaptor contamination, or other contaminations occur, the first two of which can be performed using Trimmomatic [26].

1.3.2 *De novo* assembly

NGS, though having high throughput, produces sequencing reads of short length. Decoding bacterial genome requires the genomic sequences to be determined, making it necessary to assemble the sequencing reads into larger fragments. In sequence assembly, two methods are used: mapping assembly and *de novo* assembly. Mapping assembly uses a known sequence as the backbone, conventionally called the reference sequence, and assembles sequencing reads against the reference sequence. *De novo* assembly refers to the process of assembling short sequencing reads to create full-length (sometimes novel) sequences without prior knowledge about the sequence backbone or the reference sequence. Since bacteria have quite diverse and flexible genomes subject to HGT, duplication, inversion, and large scale structural rearrangements, using reference-based methods may cause inaccurate interpretation of the genomic features. Thus, *de novo* assembly approaches are favored in bacteria genomics study. Barriers, however, exist for such approaches, which include: (1) long repeat sequences; and (2) special genetic context such as extreme GC contents or palindromic sequences. Thus, gaps are left where the genomic sequence cannot be resolved, resulting in draft-quality genomes with hundreds of contigs instead of complete genomes. Some examples of *de novo* assembly tools useful in bacteria genomics are Velvet [27], SPAdes [28], and SOAPdenovo [29].

1.3.3 Reads mapping and variant calling

While *de novo* assembly requires no additional information besides the sequencing reads, reference-based methods require a DNA sequence known to

be similar to the DNA that has been sequenced. Reference-based methods are most useful for studies of highly conservative bacterial genomes like *Mtb*, or studies of less conservative genomes when they are believed to be genetically similar such as being sampled from the same disease outbreak. Reads mapping is a process of aligning short sequencing reads to the reference sequence, which attempts to assign sequencing reads to the most likely location in the reference sequence. Various sequence alignment tools have been developed for reads mapping, some of the widely used ones include Bowtie2 [30], BWA [31], Novoalign, and SSAHA [32].

Genetic variants are differences between the studied DNA sequence and the reference sequence, which are genetic differences and may bring about phenotypic differences. Types of genetic variants include single-nucleotide polymorphism (SNP) that affects a single nucleotide, small-scale sequence variation like insertion and deletion of several consecutive bases, and large-scale sequence variation like copy number variation and rearrangement. SNP is the best studied and described among the variations. SNP calling refers to the process of determining single-nucleotide variants from the reference sequence, which generally processes the sequence alignments from reads mapping, recalibrates the quality scores, calls and filters the variants. A combinatory use of SAMtools [33] and GATK [34] proves to yield higher accuracy in SNP calling.

1.3.4 Phylogenetic tree

Phylogeny is the evolutionary relationships exhibited by different species, different strains of a same species, or other entities. A phylogenetic tree is a

tree-like diagram whose branches show the inferred phylogeny based on physical or genetic distances measured by similarity and difference. Taxa joined in the tree have an implication of descending from a common ancestor.

Two methods are usually used to construct phylogenetic trees from genetic sequences: distance-based methods like Neighbor-Joining and character-based methods like maximum parsimony and maximum likelihood. Distance-based methods first calculate the pair-wise distances from the sequence alignments, based on which a tree would be constructed. Character-based methods use individual substitutions along the sequences to determine the most likely underlying phylogenetic relationship. While character-based methods are usually more accurate than distance-based methods, the characteristic that they are highly computationally expensive makes them hard to be applied to studies with more than a few dozens of sequences.

Phylogenetic trees can also be classified based on the relative size of the branches: (1) additive trees are trees whose branch lengths are accurate representations of the accumulated differences; (2) scaled trees are trees whose branch lengths are not accurate, yet proportional to the differences between pairs of neighboring nodes; and (3) unscaled trees are trees that only convey kinship information.

Phylogenetic trees can be either rooted or unrooted. In rooted trees, one node is designated as the common ancestor, which is often artificially assigned to an outgroup (a sequence that separates early from the other sequences in the study). In unrooted trees, only interrelations are shown without indication of the evolution direction.

Trees are often tested for their reliability with bootstrapping, which offers information about the stability of the tree topology. Bootstrap generally randomly samples the columns from the sequence alignments so that some columns are not used while some used more than once. The bootstrap value, presented as a count of how many times each branch exists in exactly the same topology in all the resampled trees, is used to indicate the potential bias. While high bootstrap values are indicative of the reliability of the constructed phylogeny tree, no rule of thumb exists to define a tree as reliable using a threshold.

Various programs are available for constructing phylogenetic trees. Some of the most frequently cited programs include MrBayes [35], PAUP* [36], RAxML [37], Phyml [38], MEGA [39] and PHYLIP [40].

1.3.5 Core genome and pan genome

While phylogenetic trees are widely used for sequence analysis, which can be used to describe non-independent sequence evolution due to a common ancestor, their application to plasmid study is limited mainly by two factors: (1) massive HGT events happen; and (2) few homologous regions exist for non-clonal plasmids. The first factor is also applicable to some plastic bacterial chromosomes like the *E. coli* chromosome, where a substantial number of distinct genes exist though a set of housekeeping genes are shared.

A bacterial core genome consists only of core genes, which refer to genes shared by all individual genomes in the studied population. A bacterial pan genome, however, is made up of all non-redundant genes present in at least one of the studied genomes. Phylogenetic trees constructed using core

genomes are called the core genome trees, which are based primarily on sequence alignments, while those constructed using pan genomes are called pan genome trees, which are based primarily on the presence and absence of genes and the similarity of the genes present.

If a pan genome tree is constructed based only on the presence or absence of genes, the genetic information in the gene sequences are overlooked and hidden paralogies are ignored by using the BLAST reciprocal best hit definition of orthology [41]. A modified version of a pan genome tree is to base not only on the genes' presence or absence among the studied genomes, but also on the similarity of the genes using a distance measure. This reforms the pan genome tree if the divergences of the genes are large and thus the similarity level implies phylogenetic relationship. However, when considering the concept of phylogenetic study as the study of evolutionary relationships, a pan genome tree is actually more of a distance-based clustering pattern rather than a phylogenetic tree. In fact, phylogenetic study is not well suited for plasmid relationship analysis due to the absence of universally shared genes, which is a prerequisite for phylogenetic analysis.

If a core genome tree is constructed from a concatenation of all the core gene sequences, genes that are shared among all sequences in the studied population are considered. Evidence has been reported [42] that informational genes, in contrast to operational genes, have more macromolecule interactions and are less likely to be transferred, which is supported by the findings of Daubin, *et al* [41]. It is therefore possible that a set of genes are more closely correlated in the long run and thus may form the core genome. One study reported the core genome tree of *E. coli* correlates well with the phylotypes

and multi-locus sequence types (MLSTs), thus supporting the use of core genome tree to infer *E. coli* phylogeny [43].

In bacteria genomics, if we want to study the relatedness of different plasmids, a pan genome approach would be appropriate since that the divergence is so high that plasmids may share no genes in common and that the differences between genes are so large that distances calculated from the similarity level can well reflect the phylogenetic relationship. If, however, we are investigating the phylogenetic relationships of bacterial chromosomes, a core genome approach is preferred due to the large portion of genes shared and the biological explanation of the existence of a core genome.

Chapter 2

Aims

2.1 Chapter 3 ReRCoP: core genome phylogeny of large bacterial population samples with recombination removal

Phylogenetic study is a most useful approach for evolutionary history inference in bacteria genomics, which can be adversely affected by recombination caused by HGT or homologous recombination. In Chapter 3, I would describe ReRCoP, a novel method for identifying and removing recombination in bacterial genomes, which possesses the following features: (1) efficiently processes whole genome sequences of a large number of bacterial isolates; (2) automatically identifies and extracts the core genome; (3) robust to mutational hotspots and coldspots; and (4) accepts both complete and draft-quality assembled genomes. Simulations, comparisons, and analysis were conducted to assess its performance and utility.

2.2 Chapter 4 Local transmission and global dissemination of New Delhi metallo-beta-lactamase (*bla*$_{\mathrm{NDM}}$): a whole genome analysis

The New Delhi metallo-beta-lactamase (*bla*$_{\mathrm{NDM}}$) gene, a plasmid-borne carbapenemase gene that encodes an enzyme to make bacteria resistant to a broad range of beta-lactam antibiotics, has been found in extremely diverse bacterial strains globally, thus causing serious public health concerns worldwide. In Chapter 4, a whole genome analysis was conducted to investigate the local transmission and global dissemination of the *bla*$_{\mathrm{NDM}}$ gene. To investigate the local transmission pattern, whole genome sequencing data

of 11 *bla*$_{NDM}$-positive bacteria isolated in a local hospital was analyzed to: (1) identify and compare the *bla*$_{NDM}$-positive plasmids; and (2) study the phylogenetic relationships of the bacterial chromosomes. The global analysis was conducted by analyzing 2,749 complete plasmid sequences (including 39 *bla*$_{NDM}$-positive plasmids) in the NCBI database, where: (1) the plasmids were clustered based on the gene composition similarity and clusters with *bla*$_{NDM}$-positive plasmids were identified to be of special concern; (2) phylogenetic study was conducted for each *bla*$_{NDM}$-positive plasmid cluster to infer the phylogenetic relationships within each cluster; (3) gene transposition events introducing *bla*$_{NDM}$ into different plasmid backbones were identified; and (4) clustering pattern was correlated with the plasmids' incompatibility groups and the geographical distribution. The analysis has revealed the complex genetic pathways of *bla*$_{NDM}$ spread, where the global dissemination is mainly by introduction into different backbones via gene transposition and the subsequent local transmission is a result of plasmid conjugation and bacteria spread.

2.3 Chapter 5 Gene evolution by duplication: innovation, amplification, innovation and divergence

Gene duplication is an important mechanism for gene evolution and new gene generation. In Chapter 5, the IAID (Innovation-Amplification-Innovation-Divergence) model is proposed to explain the generation of new genes by

duplication, especially in bacteria. In this model, a gene with side functions generated by microevolution get amplified, after which microevolution still brings about innovations for each copy as they diverge from each other under selection pressure. One example is the *LamB* gene that is duplicated in *Klebsiella pneumoniae* and other related species. Using 34 complete genome sequences from NCBI, it is shown that the duplication arising by tandem duplication and passing on to different genomes is stably maintained and the copies are driven to diverge from each other by selection pressures. Haplotype reconstruction of whole genome sequences from 22 clinical isolates pictured the gene in each isolate as a population of similar sequences. The results suggest the efficacy of the IAID model in explaining gene evolution by duplication in bacteria.

2.4 Chapter 6 SpoTyping: fast and accurate in silico *Mycobacterium* spoligotyping from sequencing reads

Spoligotyping is a widely used genotyping method for *Mycobacterium tuberculosis.* In Chapter 6, I described SpoTyping, a fast and accurate program for *in silico* spoligotyping of *Mycobacterium tuberculosis* isolates from next-generation sequencing reads. This novel method achieves high accuracy for reads of both uniform and varying lengths, and is about 20-40 times faster than SpolPred. SpoTyping also integrates the function of producing a report

summarizing associated epidemiological data from a global database of all isolates having the same spoligotype.

Chapter 3

ReRCoP: core genome phylogeny of large bacterial population samples with recombination removal

3.1 Background

Homoplasy refers to the situation where two organisms are genetically similar despite not descending from a common ancestor. A major reason for homoplasy in bacteria is genetic recombination [44], which is the exchange of genetic materials between two DNA molecules. While some bacterial species like *Mtb* have genetically uniform lineages [45], others can experience more extensive genomic changes due to recombination. Some bacterial species, *H. influenzae* and *S. pneumoniae*, for example, have extensive homologous recombination between similar sequences from closely related strains [46]. Some bacterial species go through widespread HGTs that introduce large blocks of foreign genetic sequences into the genome, which is common in certain pathogens like many enterobacteria, some staphylococci and streptococci [11, 47–49]. Three mechanisms account for bacterial genetic recombination: conjugation [50, 51], transformation [51, 52], and transduction [53]. Unlike point mutations that are inherited vertically and accumulated gradually, genetic recombination introduces large fragments of foreign sequences instantaneously. Since genetic recombination has no implication for common ancestry or descendant, removing recombination can help to eliminate any confounding effect it has on evolutionary history reconstruction [54–56] and molecular clock inference [56–58].

Many methods have been proposed to detect recombination from genomic sequences [59], which can be broadly classified into two categories: similarity-search methods and SNP density change detection methods. Similarity-search methods view recombination as the transmission of genetic material from a donor sequence to a recipient sequence and thus explicitly

search for high levels of similarity between genetically divergent sequences. These methods can be either block-based, which search for ‘mosaic structures’ in genomic sequences [60–62], or position-based, which search for homoplasmic sites [63] or incongruent phylogenetic partitions [64]. Homoplasy test [63], for example, describes true homoplasy as the same sites mutated independently in different phylogenetic tree branches. However, such similarity-search methods rely on the assumption that both donor and recipient sequences are available for analysis, and this is not always possible owing to the large size of bacterial populations and the limited number of sequences that are usually sampled.

Methods that detect SNP density change view recombination as introducing genetic regions with a different density of SNPs compared to the background level [54]. Many of such methods detect abnormal distributions of discordant sites [65, 66], such as analyzing the distribution of variable sites and searching for clustering or non-random distributions of genetic variants [66]. Methods such as ClonalFrame [67], BratNextGen [68] and Gubbins [54] search for genomic regions with higher mutation rates than the background rate, or search closely related sequences for highly divergent regions. However, methods that rely on detecting changes in SNP density or mutation rates typically do not consider the possibility that mutation sites can be unevenly distributed across the genomes, particularly ignoring the presence of mutational hotspots and coldspots [69].

Existing methods to detect recombination thus possess the following limitations: (1) For bacterial species affected by HGT and with highly plastic genomes, rightfully the phylogenetic study should be confined to the core

bacterial genome rather than with genome alignments against a reference genome [16, 17, 43]. However, many of the existing recombination removal tools for analyzing whole bacterial genomes either cannot be applied to core genomes, or require substantial user pre-processing. (2) Advancements in high-throughput sequencing technologies have enabled large number of bacterial isolates sequenced and assembled in draft quality. Many of the existing analytical methodologies either cannot handle large numbers of bacterial samples or cannot handle draft-quality genomes in the absence of a highly similar complete genome as the reference sequence. (3) Mutation rates are assumed to be constant across the entire genome, ignoring the presence of mutational hotspots and coldspots.

In this chapter, ReRCoP (Recombination Removal for Core genome Phylogeny), a novel method for identifying and removing recombination in the core genomes of bacterial isolates is described. ReRCoP relies on detecting changes in SNP density as an indicator of recombination, except it does this at the gene-level rather than at regular fixed intervals of the genomic sequence. This allows a different mutation rate for each gene which is expected to be conserved across different genomic sequences. The presence of abnormally high or low number of SNPs in a gene segment for a genomic sequence is thus an indication that recombination is likely to have occurred to introduce a gene segment of dissimilar SNP density. This thus changes the nature of identifying recombination to one of detecting outliers in SNP density. ReRCoP comes with three different approaches to detect outliers, and we benchmarked the sensitivity and specificity of ReRCoP with a series of simulations to detect HGT in *E. coli* and homologous recombination in *S.*

pneumoniae. ReRCoP performed particularly well in detecting recombination with inter-lineage donors in closely related bacterial strains. ReRCoP was also compared against Gubbins in detecting homologous recombination in *S. pneumoniae*, demonstrating that ReRCoP achieved higher sensitivity and was more computationally efficient in memory and time taken, albeit at lower specificity. A comparison of the phylogenetic trees obtained for 94 diverse *E. coli* chromosomes and 91 ST131 *E. coli* isolates before and after recombination removal revealed striking differences between the trees, especially for closely related strains.

ReRCoP is written in Python which can be used on Linux, Mac OS, and Windows systems and is freely downloadable from https://github.com/xiaeryu/ReRCoP.

3.2 Methods

3.2.1 Description of algorithm

ReRCoP requires an input file in FASTA format, where each sequence is a genomic sequence from the studied population. The sequences can be aligned, as a result of reference-based consensus sequence building, or as a result of multiple sequence alignment. They could also be unaligned, each of which could be a complete genome or a concatenation of assembled contigs. A GenBank file of genome information is required if extracting core genes from aligned sequences is required. A FASTA file of gene coding sequences is requited if core genome identification and extraction is required.

ReRCoP is composed of four components: (1) pre-processing; (2) difference calculation; (3) recombination detection; and (4) post-processing.

The pre-processing step differs based on input files and user preference. If input genomic sequences are aligned, and phylogenetic study is to be conducted on core genomes, a GenBank file containing the genome information of the aligned sequences is required. Here, a gene is called to be 'present' in a genomic sequence if the coverage of the gene in the genomic sequence is above a threshold (*covCut*, default = 0.7). A gene recorded in the GenBank file is a core gene if it is present in all studied genomic sequences. Core genes would be extracted from each genomic sequences and each concatenated to form core genomes, which would be used as input for recombination removal. If input genomic sequences are aligned, and phylogenetic study is to be based on complete genomes, a sliding-window approach would be used to divide the genomic sequences into smaller fragments (used in a similar manner to genes used in core genome approaches, and are also included in the referred 'genes' below) for recombination removal based on a window size and a step size. If input genomic sequences are not aligned, gene coding sequences from any one of the genomic sequences are required for core genome identification. Each gene coding sequence would be searched and located in each genomic sequence using nucleotide BLAST [70]. A similarity value is calculated for each gene in each sequence from BLAST output file as [71]: (length of the matching sequence) × (BLAST identity) / (length of the reference sequence). Here, a gene with a similarity value above 0.49 is considered to be 'present' in the genome (similar to described in [72], and assessed in 3.2.8.1 and 3.3.7.1). Genes present in all genomic sequences would be extracted, aligned based on BLAST alignment, and further concatenated for each studied isolate.

For difference calculation, a consensus sequence is first built based on the resulting genomic sequences from the pre-processing step, and the number of SNPs compared to the consensus sequence would be calculated for each gene in each genomic sequence. These numbers would be scaled to make the total number of SNPs in each genomic sequence to be the same (the median of all total number SNPs), in order to better compare the number of SNPs in each gene.

Three methods are available for detecting recombination in ReRCoP: Grubbs' test (referred to as 'Grubbs'' below), k-nearest neighbors (kNN), and density-based spatial clustering of applications with noise (DBSCAN). Recombination test is conducted for each gene, where the number of scaled SNPs in this gene in each genomic sequence would be used as data points for outlier detection. If a data point is detected to be an outlier, the corresponding sequence would be recognized as recombinant at this gene.

Grubb's test [73] is a statistical test for outlier detection in univariate, normally-distributed datasets, which is also known as the maximum normed residual test, or the extreme studentized deviate test. The null hypothesis of Grubbs' test is no outliers in the dataset, while the alternative hypothesis is there being at least one outlier in the dataset. The test statistics is the largest absolute deviation from the sample mean in units of the sample standard deviation, which, for the two-sided test, can be defined as:

$$G = \frac{max_{i=1,2,\ldots,N}|X_i - \bar{X}|}{s} \qquad (1)$$

, where $\bar{X}$ and s denote the sample mean and standard deviation, respectively. The null hypothesis of no outliers is rejected at significance level α if:

$$G > \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\frac{\alpha}{2N},N-2}}{N-2+t^2_{\frac{\alpha}{2N},N-2}}} \quad (2)$$

, where $t^2_{\frac{\alpha}{2N},N-2}$ denotes the upper critical value of the t-distribution with a degree of freedom of N-2 and a significance level of $\frac{\alpha}{2N}$. For the Grubbs' outlier detection in ReRCoP, the Grubbs' statistics would be calculated for each data point as:

$$G_j = \frac{|X_j - \bar{X}|}{s} \quad (3)$$

, for j in 1, 2, …, N. A data point would be detected as an outlier if it satisfies equation (2) at a user-specified significance level (*alpha*, default = 0.05).

The kNN algorithm is a useful, non-parametric method commonly used for classification and regression, where *k* is a user-defined number and nearest neighbors are defined according to the closeness quantified by a similarity measure (distance measures, for example). It has also been proposed as a formulation for distance-based outlier detection, where each point is ranked based on its distance to its k^{th}-nearest neighbor and the top *n* points in this ranking are declared to be outliers [74]. ReRCoP thereby derives its kNN method. Absolute difference is used to measure the distances between data points in this univariate dataset. Any data point whose distance to its k^{th}-nearest (*k*, default = 0.2 (in the unit of total number of points)) neighbor is larger than a distance threshold (*radius*, default = 1.5 (in the unit of standard deviation of data points)) would be detected as an outlier.

DBSCAN is a density-based clustering algorithm, which groups points in the high-density regions together while making points in the low-density regions outliers. In DBSCAN, points are classified into core points, reachable

points and outliers based on the maximum distance to be called in the same neighborhood (*eps*) and the minimum number of points to form a dense region (*minPts*). A point is a core point if more than *minPts* points lie within its neighborhood. A point is a reachable point if it lies in the neighborhood of at least one core point. Outliers are defined as points that are not reachable from any other points. ReRCoP makes use of this algorithm in DBSCAN outlier detection method using parameters *eps* (default = 0.2 (in the unit of total number of points)) and *minPts* (default = 1 (in the unit of standard deviation of data points)).

In the post-processing step, genes detected as recombinant in a genomic sequence would have all their bases in this genome set to '-' and would thus be excluded from downstream phylogenetic analysis.

Selection of kNN parameters is discussed in detail in 3.2.8.2 and 3.3.7.2, while selection of DBSCAN parameters is discussed in 3.3.7.3.

3.2.2 Outlier detection method comparison

To compare different outlier detection methods, Grubbs', kNN, and DBSCAN as implemented in ReRCoP were each performed on simulated sequences. A typical round of simulation experiment was conducted as follows: (1) an ancestral sequence was defined, from which a specified number of sequences (*nSeq*) would be generated; (2) non-recombinant sequences were generated from the ancestral sequence by mutating each base at a specified probability (*base rate*); (3) a recombinant sequence generated from the ancestral sequence by mutating each base at a specified probability (*special rate*) would replace a non-recombinant sequence at a specified probability (*rec rate* = 0.05); and (4)

Grubbs', kNN, and DBSCAN were each used to detect recombination using default parameters in ReRCoP.

Different simulation scenarios were proposed using different *base rate*, *special rate*, and *nSeq*. Closely related bacterial strains were simulated by setting *base rate* to 0.002, with the *special rate* 2X, 5X, and 10X the *base rate* (0.004, 0.01, and 0.02, respectively). Diverse bacterial strains were simulated by setting *base rate* to 0.01, with *special rate* set to 0.1X, 0.2X, 0.5X, 2X, 5X, and 10X the *base rate* (0.001, 0.002, 0.005, 0.02, 0.05, and 0.1, respectively). These, altogether, added up to 9 pairs of mutating rates. For each pair of mutating rates, different numbers of sequences (*nSeq* = 10, 30, 50, 75, 100, 150, and 200) were simulated to evaluate the effect of sample size on outlier detection, leading to altogether 63 simulation scenarios. For each scenario, 50 iterations were conducted using different gene coding sequences as the ancestral sequence, each randomly selected from *E. coli* NA114 genome [GenBank:CP002797.2].

Outliers detected by each of the three algorithms were compared with the simulated recombination to assess the sensitivity and specificity.

3.2.3 Simulation of horizontal gene transfer on *E. coli* genomes

The ancestral genome was ST131 *E. coli* NA114 genome. Eighteen donor genomes were used in this study (Table 1). Among the 18 donor genomes, 9 are inter-lineage donors, which are complete *E. coli* genomes archived in NCBI that are different from the NA114 genome. The other 9 are intra-lineage donor genomes, each of which is a concatenation of assembled contigs from sequencing reads of an ST131 *E. coli* isolate randomly selected from an ENA

study [ENA:ERP001354]. Core genes were identified and extracted using ReRCoP with the 18 donor genomes and the ancestral genome as the input genomic sequences, and the gene coding sequences of the NA114 genome as the input gene sequences, resulting in 3,366 core genes. Core genes in the ancestral genome and donor genomes were each concatenated into an ancestral sequence and 18 donor sequences.

Each simulated sequence had a specific mutating rate which was1 randomly sampled from a uniform distribution on the interval of [0, 2**base rate*), and was generated from the ancestral sequence by creating point mutations at this mutating rate. The parameter *base rate* was set to 0.002 for simulation of closely related strains and 0.01 for simulation of diverse strains. For each gene in each simulated sequence, there is a probability of 0.01 that the gene was selected to be a recombinant gene, where the sequence was replaced by the corresponding gene sequence from a randomly selected donor. One hundred sequences were simulated in each iteration, and 100 iterations were each generated for *base rate* of 0.002 and 0.01.

Recombination detection was conducted with ReRCoP using Grubbs', kNN, and DBSCAN as outlier detection methods using default parameters.

Sensitivity was calculated as the percentage of SNPs brought in by simulated recombination that were captured by ReRCoP. False positive rate was calculated as the number of bases falsely detected as recombination compared to the total number of bases that were not simulated to be recombination. Specificity was calculated as one minus the respective false positive rate.

Table 1. Information of sequences used in simulation of horizontal gene transfer on *E. coli* genomes.

Accession	Name	Instance
CP002797.2	*E. coli* NA114	Ancestor
CP000802.1	*E. coli* HS	Inter-host donor
AP009240.1	*E. coli* SE11 DNA	Inter-host donor
CU928163.2	*E. coli* UMN026	Inter-host donor
AP010958.1	*E. coli* O103:H2 str. 12009	Inter-host donor
FN649414.1	*E. coli* ETEC H10407	Inter-host donor
CP002729.1	*E. coli* UMNK88	Inter-host donor
CP003289.1	*E. coli* O104:H4 str. 2011C-3493	Inter-host donor
BA000007.2	*E. coli* O157:H7 str. Sakai	Inter-host donor
U00096.3	*E. coli* str. K-12 substr. MG1655	Inter-host donor
ERR161234	*E. coli* ST131 lineage	Intra-host donor
ERR161235	*E. coli* ST131 lineage	Intra-host donor
ERR161236	*E. coli* ST131 lineage	Intra-host donor
ERR161237	*E. coli* ST131 lineage	Intra-host donor
ERR161238	*E. coli* ST131 lineage	Intra-host donor
ERR161239	*E. coli* ST131 lineage	Intra-host donor
ERR161240	*E. coli* ST131 lineage	Intra-host donor
ERR161241	*E. coli* ST131 lineage	Intra-host donor
ERR161242	*E. coli* ST131 lineage	Intra-host donor

Table 2. Information of sequences used in simulation of homologous recombination on *S. pneumoniae* genomes.

Accession	Name	Instance
FM211187.1	*S. pneumoniae* ATCC 700669	Ancestor
FQ312029.1	*S. pneumoniae* INV200	Inter-host donor
AE005672.3	*S. pneumoniae* TIGR4	Inter-host donor
AE007317.1	*S. pneumoniae* R6	Inter-host donor
CP003357.2	*S. pneumoniae* ST556	Inter-host donor
CP001993.1	*S. pneumoniae* TCH8431/19A	Inter-host donor
CP000921.1	*S. pneumoniae* Taiwan19F-14	Inter-host donor
CP001015.1	*S. pneumoniae* G54	Inter-host donor
CP000919.1	*S. pneumoniae* JJA	Inter-host donor
CP000410.1	*S. pneumoniae* D39	Inter-host donor
ERR023428	*S. pneumoniae* clone PMEN1	Intra-host donor
ERR023430	*S. pneumoniae* clone PMEN1	Intra-host donor
ERR023432	*S. pneumoniae* clone PMEN1	Intra-host donor
ERR023434	*S. pneumoniae* clone PMEN1	Intra-host donor
ERR023436	*S. pneumoniae* clone PMEN1	Intra-host donor
ERR023438	*S. pneumoniae* clone PMEN1	Intra-host donor
ERR023451	*S. pneumoniae* clone PMEN1	Intra-host donor
ERR023453	*S. pneumoniae* clone PMEN1	Intra-host donor
ERR023455	*S. pneumoniae* clone PMEN1	Intra-host donor

3.2.4 Simulation of homologous recombination on *S. pneumoniae* genomes

The ancestral sequence was *S. pneumoniae* ATCC 700669 genome [GenBank:FM211187.1]. Eighteen donor sequences were used in the simulation (Table 2) with half inter-lineage donors and half intra-lineage donors, whose sequence alignments were generated as described before [54]. Point mutations were created similarly as described in 'Simulation of horizontal gene transfer on *E. coli* genomes' to generate simulated sequences from the ancestral sequence. Closely related strains were simulated with *base rate* of 0.002 while diverse strains were simulated with *base rate* of 0.01. For each simulated sequence, recombination was simulated to replace a part of the original sequence at a specified probability (*rec rate*, set to 0.3, 0.6, and 0.9, respectively) with a randomly selected donor, a random start position, and a per-base probability of 0.00016 to stop recombination as suggested before [54]. One hundred sequences were simulated in each iteration, and 100 iterations were each generated for *base rate* of 0.002 and 0.01 at *rec rate* of 0.3, 0.6, and 0.9. Recombination detection was conducted using ReRCoP with Grubbs', kNN, and DBSCAN as outlier detection methods using default parameters in a sliding-window manner. Sensitivity and specificity were calculated the same as described above.

3.2.5 Performance comparison of ReRCoP and Gubbins

Recombination detection was conducted using Gubbins in comparison with ReRCoP using the simulated dataset described in the section 'Simulation of homologous recombination on *S. pneumoniae* genomes'. Both programs were run on a 64-bit Fedora Linux server workstation having a 2.0GHz quad

processor and 32GB RAM. Gubbins crashed due to insufficient free memory while processing 100 simulated sequences, each 2,221,315 bp in length. As a compromise, 60 simulated sequences were used as the input sequences at *base rate* of 0.002, and 20 simulated sequences were used at a *base rate* of 0.01, both of which were the maximum number of sequences that did not cause crash. Default parameters were used. Sensitivity and specificity were calculated the same as using ReRCoP and were compared correspondingly.

3.2.6 Core genome analysis with recombination removal of 94 diverse *E. coli* chromosomes

Ninety-four complete *E. coli* chromosomes were downloaded from GenBank, which are diverse in phylotype (determined *in silico* based on [75]) and MLST (determined *in silico* based on [76]) (Table 3). They were used as input genomes for ReRCoP, with gene coding sequences from *E. coli* str. K-12 substr MG1655 [GenBank:U00096.3], after removing duplication, as input gene coding sequences.

ReRCoP was conducted with default parameters using Grubbs', kNN, and DBSCAN as outlier detection methods. Maximum-likelihood phylogenetic trees were constructed using RAxML [37] using 'GTRCAT' model each for the core genomes without outlier removal, after Grubbs', kNN, or DBSCAN outlier removal. Consensus networks [77] were constructed using SplitsTree [78] to compare phylogenetic trees before and after outlier removal, where incompatible splits were highlighted in red.

Table 3. Information of 94 diverse *E. coli* chromosomes used in core genome analysis with recombination removal.

Accession	Name	MLST	Phylotype
AGTD01000001.1	*E. coli* UMNF18	10	A
AKBV01000001.1	*E. coli* str. K-12 substr. MG1655	10	A
AKVX01000001.1	*E. coli* str. K-12 substr. MG1655	10	A
AP009048.1	*E. coli* str. K12 substr. W3110	10	A
AP012306.1	*E. coli* str. K-12 substr. MDS42	10	A
CM000960.1	*E. coli* str. K-12 substr. MG1655star	10	A
CP001396.1	*E. coli* BW2952	10	A
CP002291.1	*E. coli* P12b	10	A
CP006698.1	*E. coli* C321.deltaA	10	A
CP008801.1	*E. coli* KLY	10	A
CP009273.1	*E. coli* BW25113	10	A
CP009644.1	*E. coli* ER2796	10	A
CP009685.1	*E. coli* str. K-12 substr. MG1655	10	A
CP009789.1	*E. coli* K-12 strain ER3413	10	A
HG738867.1	*E. coli* str. K-12 substr. MC4100	10	A
U00096.3	*E. coli* str. K-12 substr. MG1655	10	A
CP004009.1	*E. coli* APEC O78	23	A
CP000802.1	*E. coli* HS	46	A
FN649414.1	*E. coli* ETEC H10407	48	A
AM946981.2	*E. coli* BL21(DE3)	93	A
CP000819.1	*E. coli* B str. REL606	93	A
CP001509.3	*E. coli* BL21(DE3)	93	A
CP001665.1	*E. coli* 'BL21-Gold(DE3)pLysS AG'	93	A
CP002729.1	*E. coli* UMNK88	100	A
CP007265.1	*E. coli* strain ST540	540	A
CP007390.1	*E. coli* strain ST540	540	A
CP007391.1	*E. coli* strain ST540	540	A
AP012030.1	*E. coli* DH1 (ME8569)	1060	A
CP000948.1	*E. coli* str. K12 substr. DH10B	1060	A
CP001637.1	*E. coli* DH1	1060	A
CP000946.1	*E. coli* ATCC 8739	3021	A
AP010960.1	*E. coli* O111:H- str. 11128	16	B1
AP010958.1	*E. coli* O103:H2 str. 12009	17	B1
AP010953.1	*E. coli* O26:H11 str. 11368	21	B1
CP005998.1	*E. coli* B7A	94*	B1
AP009240.1	*E. coli* SE11	156	B1
CP009578.1	*E. coli* FAP1	453	B1
CP009106.1	*E. coli* strain 94-3024	672	B1
CP003289.1	*E. coli* O104:H4 str. 2011C-3493	678	B1
CP003297.1	*E. coli* O104:H4 str. 2009EL-2050	678	B1
CP003301.1	*E. coli* O104:H4 str. 2009EL-2071	678	B1
CU928145.2	*E. coli* 55989	678	B1
CP002185.1	*E. coli* W	1079	B1
CP002516.1	*E. coli* KO11	1079	B1
CP002967.1	*E. coli* W	1079	B1
CP002970.1	*E. coli* KO11FL	1079	B1
CP006584.1	*E. coli* LY180	1079	B1
CU928160.2	*E. coli* IAI1	1128	B1
CP000800.1	*E. coli* E24377A	1132	B1

CP009104.1	*E. coli* strain RM9387	2773	B1
AE014075.1	*E. coli* CFT073	73	B2
CP001671.1	*E. coli* ABU 83972	73	B2
CP002211.1	*E. coli* str. 'clone D i2'	73	B2
CP002212.1	*E. coli* str. 'clone D i14'	73	B2
CP007799.1	*E. coli* Nissle 1917	73	B2
CP009072.1	*E. coli* ATCC 25922	73	B2
CP000243.1	*E. coli* UTI89	95	B2
CP000468.1	*E. coli* APEC O1	95	B2
CP001969.1	*E. coli* IHE3034	95	B2
CU928161.2	*E. coli* S88	95	B2
NZ_HG428755.1	*E. coli* PMV-1	95*	B2
AP009378.1	*E. coli* SE15	131	B2
CP002797.2	*E. coli* NA114	131	B2
CP006784.1	*E. coli* JJ1886	131	B2
CP001855.1	*E. coli* O83:H1 str. NRG 857C	135	B2
CU651637.1	*E. coli* LF82	135	B2
CP002167.1	*E. coli* UM146	643	B2
CP000247.1	*E. coli* 536	4727*	B2
FM180568.1	*E. coli* 0127:H6 E2348/69	4728*	B2
CU928162.2	*E. coli* ED1a	4731*	B2
AE005174.2	*E. coli* O157:H7 EDL933	11*	D
BA000007.2	*E. coli* O157:H7 str. Sakai	11	D
CM000662.1	*E. coli* O157:H7 str. TW14588	11	D
CP001164.1	*E. coli* O157:H7 str. EC4115	11	D
CP001368.1	*E. coli* O157:H7 str. TW14359	11	D
CP001925.1	*E. coli* Xuzhou21	11	D
CP008805.1	*E. coli* O157:H7 str. SS17	11	D
CP008957.1	*E. coli* O157:H7 str. EDL933	11	D
CP010304.1	*E. coli* O157:H7 str. SS52	11	D
CP006027.1	*E. coli* O145:H28 str. RM13514	32	D
CP007136.1	*E. coli* O145:H28 str. RM12581	32	D
CP003034.1	*E. coli* O7:K1 str. CE10	62	D
CU928164.2	*E. coli* IAI39	62	D
CP001846.1	*E. coli* O55:H7 str. CB9615	335	D
CP003109.1	*E. coli* O55:H7 str. RM12579	335	D
CP000970.1	*E. coli* SMS-3-5	354	D
FN554766.1	*E. coli* 042	414	D
CU928163.2	*E. coli* UMN026	597	D
CP009859.1	*E. coli* strain ECONIH1	648	D
CP006262.1	*E. coli* O145:H28 str. RM13516	4729	D
CP007133.1	*E. coli* O145:H28 str. RM12761	4729	D
CP007392.1	*E. coli* strain ST2747	4730	D
CP007393.1	*E. coli* strain ST2747	4730	D
CP007394.1	*E. coli* strain ST2747	4730	D

* The most similar sequence type. No matching sequence type with 100% identity.

3.2.7 Recombination removal using a sliding-window approach of Illumina sequencing reads of 91 ST131 *E. coli* isolates

Sequencing reads of 91 ST131 *E. coli* isolates [ENA:ERP001354] were included in the analysis. Complete genomic sequence of ST131 *E. coli* NA114 [GenBank:CP002797.2] was used as the reference genome, against which sequencing reads were mapped using BWA [31] and consensus sequences were built using SAMtools [33]. The constructed consensus sequences were used as input genomes for ReRCoP. Recombination removal was conducted with default parameters using Grubbs', kNN, and DBSCAN in a sliding-window manner. Maximum-likelihood phylogenetic trees were constructed and compared, in the same manner as in the 94 diverse *E. coli* chromosomes.

3.2.8 Choice of parameters

3.2.8.1 Choice of parameter in core gene identification

One parameter needs to be optimized in core gene identification, which is the similarity value threshold to classify a gene as 'present' or 'absent' in the genome. A similarity value is calculated from nucleotide BLAST output file as: (length of the matching sequence) × (BLAST identity) / (length of the reference sequence). The choice of the similarity value threshold was thus assessed in the following experiment. All gene coding sequences of the 94 diverse *E. coli* isolates (Table 3) were downloaded from NCBI (438,159 genes in total). Each gene coding sequence was compared with every other gene coding sequence using nucleotide BLAST, from which a similarity value was calculated as described above. The threshold was selected within the region that has the lowest number of gene pairs having such similarity values.

3.2.8.2 Choice of parameters in kNN outlier detection

Simulations were conducted to evaluate selection of parameters in kNN (*k*: number of nearest neighbors to consider, and *radius*: distance threshold). The simulations were conducted on a dataset composed of: (1) 10,000 non-outliers that were randomly generated from the standard normal distribution (mean = 0, standard deviation = 1); and (2) 10,000 outliers that were randomly generate from uniform distributions (half on the interval of [-4, -2], the other half on the interval of [2, 4]). Eight distance thresholds (d_{thresh}) were considered in the simulation: 0.5, 1, 1.5, 2, 2.5, 3, 3.5, and 4 (all in the unit of the standard deviation of the data points). Respective simulation was conducted using each of the distance thresholds (d_{thresh}) as such: (1) For each data point labeled as either 'non-outlier' or 'outlier' in the dataset, the percentage of data points having an absolute distance smaller than the distance threshold was calculated ($p_{neighbor}$); and (2) A set of different percentages (p_{thresh}) was used in an attempt to predict the label of the data point (non-outlier if $p_{neighbor} > p_{thresh}$, and outlier otherwise), where the sensitivity and specificity of prediction were calculated for each p_{thresh}. Here, distance thresholds d_{thresh} represent *radius* in kNN outlier detection in the unit of standard deviation, and the percentages p_{thresh} represent *k* in kNN outlier detection in the unit of total number of points.

3.3 Results

3.3.1 Comparison of outlier detection methods in ReRCoP

Simulations were conducted to assess and compare performance of Grubbs' test (referred to as Grubbs' below), kNN, and DBSCAN outlier detection

under different circumstances. After formulating recombination detection into an outlier detection problem, factors potentially affecting detection performance were varied, which include the mutating rate from the ancestral sequence in non-recombinant sequences (*base rate*), the mutating rate from the ancestral sequence in recombinant sequences (*special rate*), and the number of sequences in the simulation (*nSeq*). Program parameters in ReRCoP were optimized separately, thus were not varied in this simulation and default settings were used. Two different *base rate*s were used: 0.002 to simulate closely related strains, and 0.01 to simulate diverse strains. For closely related strains, *special rate* was set to 2X, 5X, and 10X of the *base rate* (0.004, 0.01, and 0.02, respectively) to simulate recombination that brings in more SNPs compared to the background level. It is not surprising that recombination can not only lead to more SNPs but also fewer SNPs from the ancestral sequence. This is not discriminatory in closely related strains due to the already limited number of SNPs in non-recombinant sequences, but it is discriminatory in diverse strains. As a result, for simulation of diverse strains, *special rate* was set to 0.1X, 0.2X, 0.5X, 2X, 5X, and 10X of the *base rate* (0.001, 0.002, 0.005, 0.02, 0.05, and 0.1, respectively). For each pair of mutating rates (*base rate* and *special rate*), *nSeq* was set to 10, 30, 50, 75, 100, 150, and 200 to simulate different number of studied isolates. Fifty iterations were conducted in each scenario, each using a different ancestral sequence.

For simulation of closely related strains (*base rate* = 0.002; results summarized in Figure 1), the overall detection sensitivity increased with the increase of *special rate*. Both kNN and DBSCAN had similar sensitivity, while Grubbs' had a relatively lower sensitivity. Number of sequences (*nSeq*)

was not consequential to the sensitivity. The specificity was mostly above 0.95, which increased with *special rate* and *nSeq*. Grubbs' had the highest specificity, with kNN second to it, and DBSCAN the lowest.

For simulations of diverse strains (*base rate* = 0.01; results summarized in Figure 1), when *special rate* is larger than 1, the overall sensitivity increased with *special rate*. Similarly, kNN and DBSCAN had higher sensitivity compared to Grubbs'. Again, *nSeq* was not consequential to the sensitivity. The sensitivity was always about 1 with *nSeq* larger than 30. However, when *special rate* is smaller than 1, although having high specificity, the sensitivity was low, which is a result of the relatively smaller SNP number differences.

In summary, DBSCAN and kNN had relatively higher sensitivity and lower specificity, while Grubbs' did the opposite. The performance increased when recombination brought a greater increase in the number of SNPs.

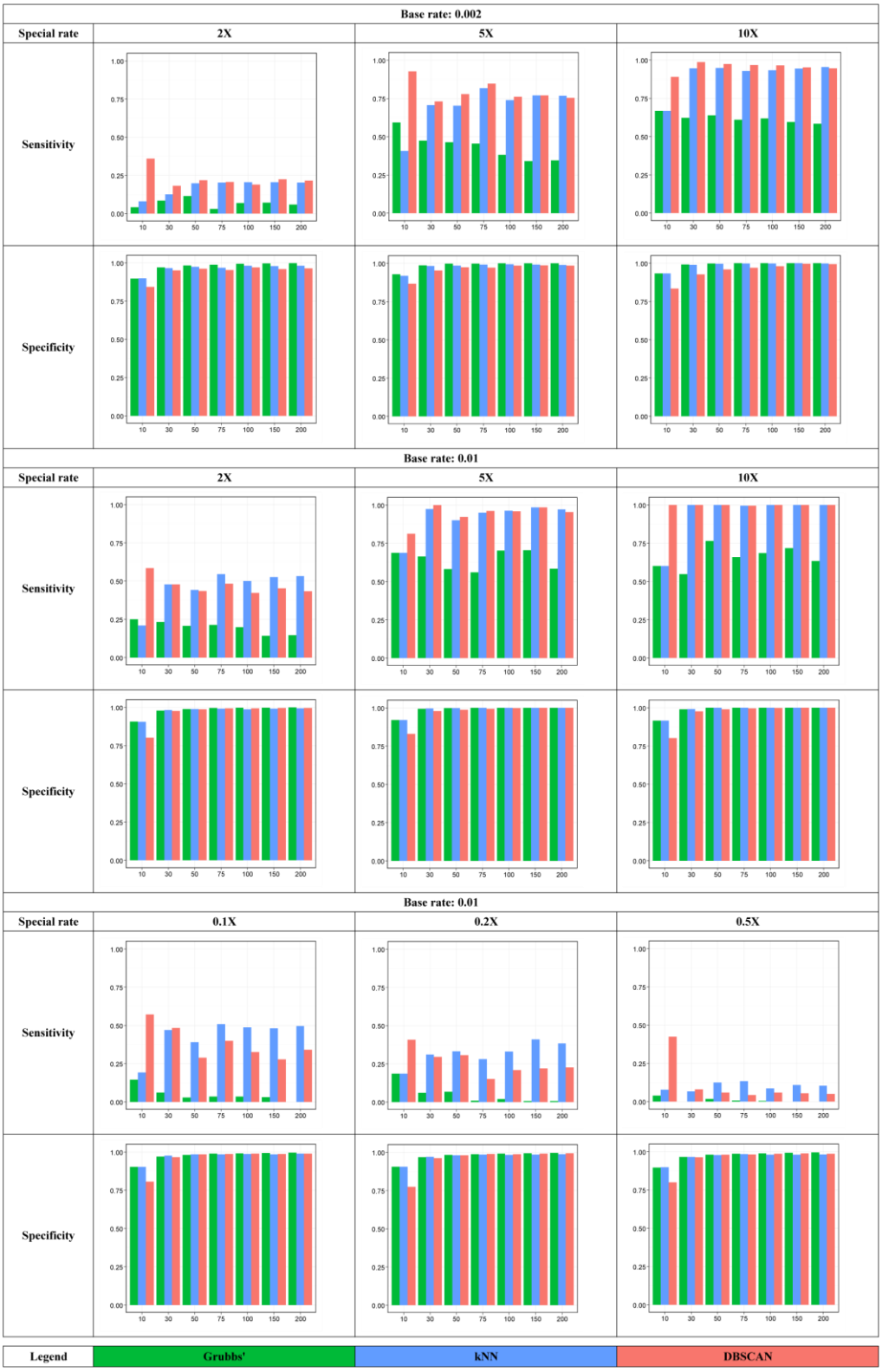


Figure 1. Comparison of outlier detection methods. Different scenarios were simulated to compare sensitivity and specificity of Grubbs' (green), kNN (blue), and DBSCAN (pink) outlier detection under different circumstances (*base rate*: mutating rate in non-recombinant sequences, *special rate*: mutating rate in recombinant sequences, and *nSeq*: number of simulated sequences). The x-axis indicates the number of sequences while the y-axis indicates the respective sensitivity or specificity. In simulation of closely related strains (*base rate* = 0.002), detection

sensitivity increased with the increase of *special rate*. Both kNN and DBSCAN had similar sensitivity, while Grubbs' had a relatively lower sensitivity. Detection specificity was mostly above 0.95, which increased with *special rate*. Grubbs' had the highest specificity, with kNN second to it, and DBSCAN the lowest. In simulation of diverse strains (*base rate* = 0.01), when *special rate* is larger than 1, the overall sensitivity increased with *special rate*. Similarly, kNN and DBSCAN had higher sensitivity compared to Grubbs'. However, when *special rate* is smaller than 1, the detection sensitivity was low for all three methods. Detection specificity was mostly about 1. In simulations of both close and diverse bacterial strains, increase in the number of sequences (*nSeq*) was not consequential to the sensitivity, but helped to increase detection specificity. In summary, DBSCAN and kNN had relatively high sensitivity and low specificity, while Grubbs' did the opposite.

3.3.2 Simulation of horizontal gene transfer on *E. coli* genomes

Genetic sequences were generated from an ancestral sequence with a mutating rate (*base rate*). HGT was simulated by replacing certain simulated gene sequences with foreign gene sequences from either intra-host donors, which are sequences having the same sequence type as the ancestral sequence, or inter-host donors, which are sequences quite different from the ancestral sequence.

ReRCoP was run on a 64-bit Fedora Linux server workstation having a 2.0GHz quad processor and 32GB RAM in all experiments. It took an average of 6.02 min (standard deviation = 0.37 min) to complete running an analysis of 100 sequences, each of 3,119,466 bp in length. Sensitivity and specificity were summarized in Figure 2. For simulations of closely related strains (*base rate* = 0.002), ReRCoP detected recombination with intra-lineage donors at a sensitivity around 5% whichever method was used. For recombination with inter-lineage donors, the sensitivity differed with the method used, where DBSCAN had the highest average sensitivity of 89.01%, kNN followed with an average sensitivity of 84.56%, and Grubbs' the lowest of 75.61%. All

methods had specificity above 97%, where DBSCAN had an average specificity of 97.27%, kNN of 98.15%, and Grubbs' of 99.33%. For simulations of diverse strains (*base rate* = 0.01), ReRCoP detected recombination with intra-lineage donors better than it did in closely related strains though having a larger variation, with an average sensitivity of 27.27% while using kNN, 20.42% while using DBSCAN, and 3.42% while using Grubbs'. Detection sensitivity of recombination with inter-lineage donors, however, was lower than in closely related strains, where kNN and DBSCAN performed similarly with a sensitivity of about 53%, while Grubbs' did relatively lower at 40.84%. Grubbs' had an average specificity of 99.32%, DBSCAN had a specificity of 97.52%, while kNN had the lowest specificity of 96.81%.

In summary, detection sensitivity for recombination with intra-lineage donors was not high due to the limited number of SNPs brought in by the recombinant sequence. Detection sensitivity for recombination with inter-lineage donors was higher, especially in closely related strains. Specificity was consistently above 96%. In terms of methods, Grubbs' had lower sensitivity and higher specificity, while kNN and DBSCAN had higher sensitivity and lower specificity.

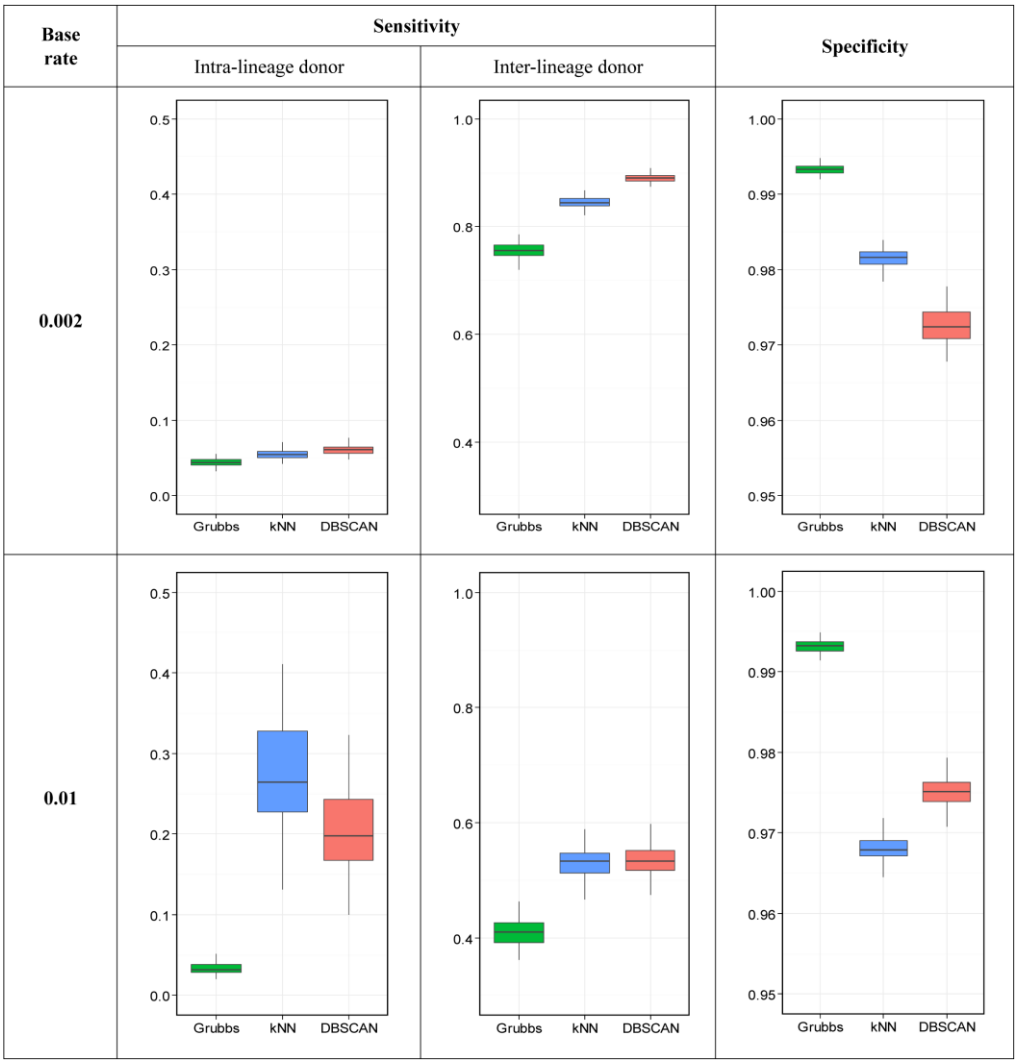


Figure 2. Performance of ReRCoP recombination detection in simulations of horizontal gene transfer on *E. coli* genomes. Simulations of HGT were conducted on *E. coli* genomes to assess the detection sensitivity and specificity of ReRCoP under different circumstances (*base rate*: mutating rate in non-recombinant sequences, and donor sequences: either inter-host donor or intra-host donor). The y-axis indicates the respective sensitivity and specificity indicated as the column names. For simulations of closely related strains (*base rate* = 0.002), ReRCoP detected recombination with intra-lineage donors at low sensitivity, while for recombination with inter-lineage donors, the sensitivity was much higher, where DBSCAN had the highest sensitivity, followed by kNN and Grubbs'. Contrary to the sensitivity, DBSCAN, kNN, and Grubbs' had decreasing specificity. For simulations of diverse strains (*base rate* = 0.01), ReRCoP detected recombination with intra-lineage donors better than it did in closely related strains though less consistent. Detection of recombination with inter-lineage donors, however, was lower than in close strains, where kNN and DBSCAN performed similarly, while Grubbs' did relatively lower. Grubbs' had the highest specificity, followed by DBSCAN and kNN. In terms of methods, Grubbs' had the lowest sensitivity and the highest specificity, while kNN and DBSCAN had higher sensitivity and lower specificity.

3.3.3 Simulation of homologous recombination on *S. pneumoniae* genomes

After generating genetic sequences from an ancestral sequence with a mutating rate (*base rate*), homologous recombination was simulated in a certain percentage (*rec rate*) of the sequences by replacing random regions of genetic sequences with corresponding foreign sequences from either intra-host donors or inter-host donors.

It took an average of 5.64 min (standard deviation = 0.83 min) to complete an analysis of 100 sequences, each of 2,221,315 bp in length. Sensitivity and specificity were summarized in Figure 3. The overall performance of ReRCoP was better on simulated datasets of closely related strains in terms of sensitivity, specificity and consistency. For simulated datasets of closely related strains (*base rate* = 0.002), the performance was consistent regardless of *rec rate*. Grubbs' had lower sensitivity and higher specificity compared to kNN and DBSCAN, both of which had similar sensitivity while DBSCAN had slightly higher specificity. The average sensitivity was always around 15% using all three methods in detecting recombination from intra-lineage donors, and was around 70% using Grubbs', 78% using kNN and DBSCAN in detecting recombination from inter-lineage donors. The average specificity was around 98% using Grubbs', over 95% using kNN and DBSCAN. When considering simulated datasets of diverse strains (*base rate* = 0.01), using a *rec rate* of 0.9 would slightly decrease the sensitivity and increase the specificity, while results using 0.3 and 0.6 were very similar and are used in the following description of performance. Grubbs' still had the lowest average sensitivity (7% for intra-lineage donors, and 46% for inter-host donors), the highest average specificity (above 98%) and the most consistent performance.

DBSCAN had medium average sensitivity (22% for intra-lineage donors, and 48% for inter-host donors) and medium average specificity (above 95%). kNN exhibited the best average sensitivity (46% for intra-lineage donors, and 51% for inter-host donors), especially a much better average sensitivity in detecting recombinant genes from intra-host donors with a wider range of sensitivity values in different iterations, though it has the lowest specificity (above 93%).

In summary, *rec rate* did not have a large effect on detection performance. Detection sensitivity was higher for recombination with inter-lineage donors, especially in closely related strains. Sensitivity was lower for recombination with intra-lineage donors due to the limited number of SNP change. When comparing the three methods, Grubbs' had the highest specificity, with DBSCAN second to it, and kNN the lowest.

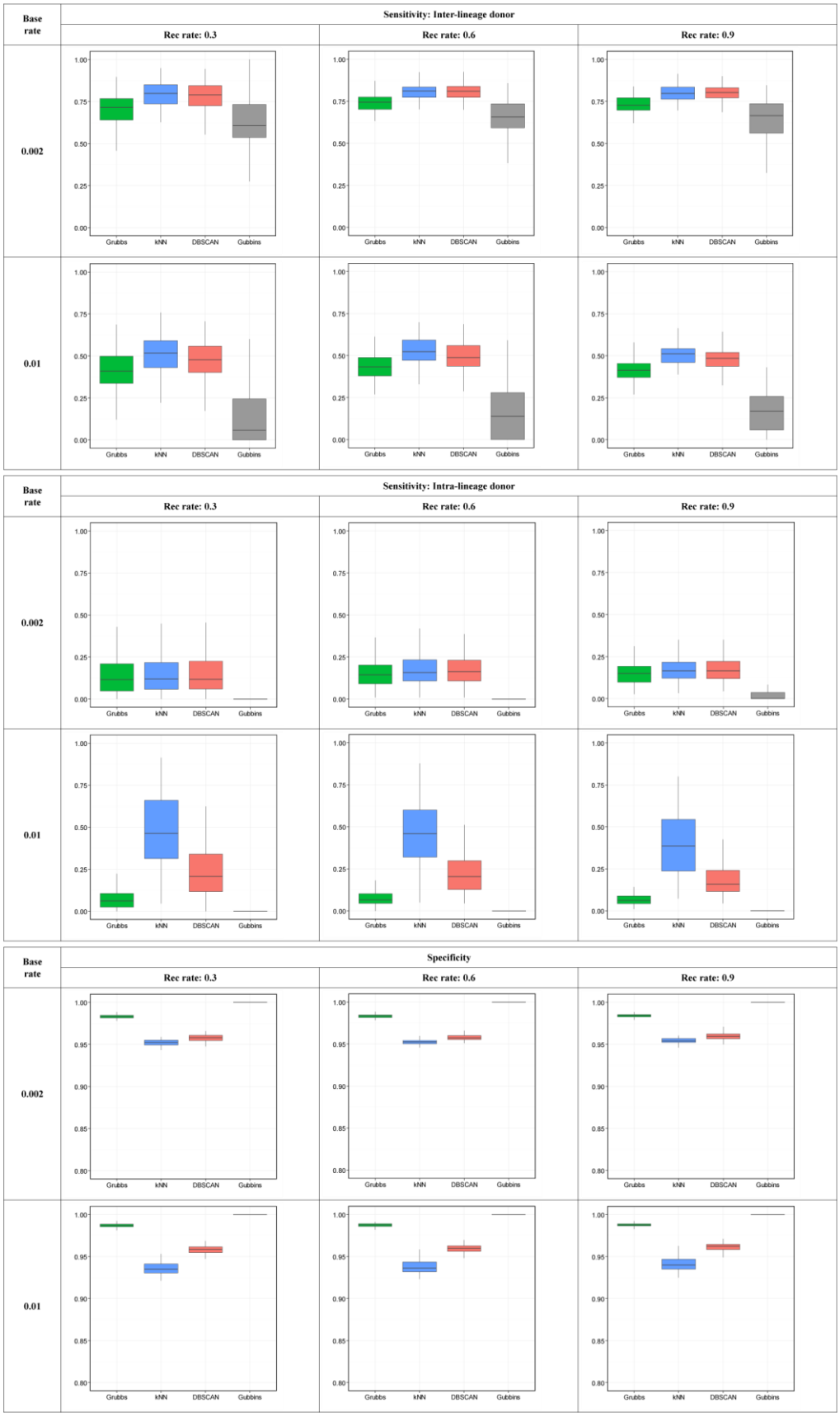


Figure 3. Performance of ReRCoP in simulations of homologous recombination on *S. pneumoniae* genomes in comparison with Gubbins. Simulations of homologous recombination were conducted on *S. pneumoniae* genomes to assess the

detection sensitivity and specificity of ReRCoP in comparison with Gubbins under different circumstances (*base rate*: mutating rate in non-recombinant sequences, *rec rate*: percentage of sequences with homologous recombination, and donor sequences: either inter-host donor or intra-host donor). The overall performance of ReRCoP and Gubbins was better on simulated datasets of closely related bacteria in terms of sensitivity, specificity and consistency. Detection of recombination with inter-host donors was more sensitive than intra-host donors due to the less number of SNPs brought in. Generally, *rec rate* did not have a large effect on the detection performance. When comparing different methods in ReRCoP, Grubbs' had relatively lower sensitivity and higher specificity. kNN and DBSCAN had similar sensitivity while DBSCAN had slightly higher specificity. When comparing ReRCoP and Gubbins, ReRCoP was more memory and time efficient, more sensitive, and less specific than Gubbins.

3.3.4 Performance comparison of ReRCoP and Gubbins

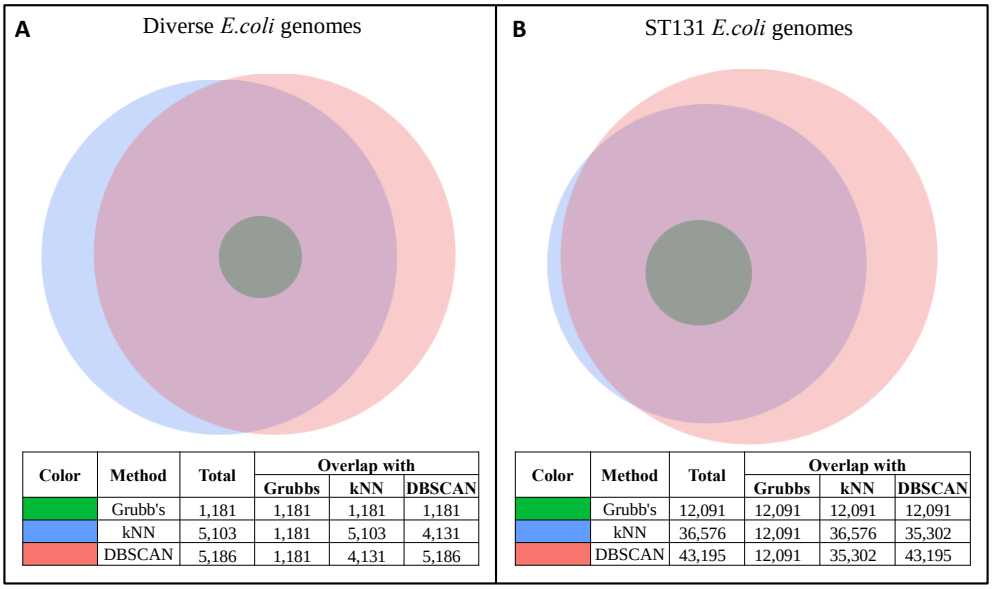
Since Gubbins was mostly described to be used for detecting homologous recombination, performance comparison of ReRCoP and Gubbins was conducted on the datasets used in the simulation of homologous recombination on *S. pneumoniae* genomes described above. Gubbins returned an error message indicating insufficient free memory when processing 100 sequences. I thus decided on using 60 sequences in simulations of closely related strains and 20 sequences in simulation of diverse strains instead of 100 sequences to guarantee successful execution and the most number of sequences used. Gubbins required both much free memory and time to run. For simulation of closely related strains (*base rate* = 0.002), Gubbins took an average of 312.04 min to process 60 sequences (standard deviation = 196.41 min). For simulation of diverse strains (*base rate* = 0.01), Gubbins took an average of 14.05 min to process 20 sequences (standard deviation = 3.76 min). Sensitivity and specificity were plotted on Figure 3 next to ReRCoP. Gubbins showed lower sensitivity and higher specificity than any of the three methods used in ReRCoP in all simulation scenarios. Gubbins performed its own best in detecting recombination with inter-lineage donor in closely related strains,

where the sensitivity was close to, though still lower than, Grubbs' outlier detection in ReRCoP with a larger variation. The sensitivity decreased much in diverse strains with an even larger variation. Both Gubbins and ReRCoP showed significantly lower sensitivity in detecting recombination from intra-lineage donors. While ReRCoP still detected some recombination, Gubbins was almost not detecting any such recombination. Coming along with the lower sensitivity was the higher specificity of Gubbins, where nearly no false positive hits were identified. In summary, ReRCoP was more memory and time efficient, more sensitive, and less specific than Gubbins.

3.3.5 Core genome analysis with recombination removal of 94 diverse *E. coli* chromosomes

In this analysis, input genomes were 94 complete *E. coli* genomes with different phylotypes and MLSTs, thus representing a diverse collection of bacterial chromosomes. A core genome approach was applied based on the facts that: (1) the sequences were not aligned, and (2) gene composition and organization were different. Gene coding sequences from *E. coli* str. K-12 substr MG1655 [GenBank:U00096.3] were used as input gene sequences for identifying core genes to comprise the core genome. Among the 3,769 input genes, 2,720 were identified as core genes, adding up to a core genome size of 2,618,529 bp. The three methods in ReRCoP were each used for recombination removal. The running time was 97 min, the majority of which was spent on core genome identification and alignment. The number of genes identified (out of the total of 255,680 genes) as recombinant was 1,181 for Grubbs', 5,103 for kNN, and 5,186 for DBSCAN. Number of overlapped

genes identified is shown in Figure 4A, showing that genes identified by Grubbs' was a subset of genes identified by kNN or DBSCAN, and that kNN and DBSCAN had more than 80% of the identified genes in common, which was consistent with the simulation result that Grubbs' is a more conservative method. Maximum-likelihood trees were constructed using sequences before and after recombination removal. Phylogenetic trees built from sequences after using the three recombination removal methods were each compared with the tree built from the sequence before recombination removal by constructing a consensus network, where incompatible splits were highlighted in red to show the difference. Consensus networks showed that recombination removal did not affect the major branching of the phylogenetic tree, but had an impact on the topology within branches (Figure 5).



Color	Method	Total	Overlap with		
			Grubbs	kNN	DBSCAN
	Grubb's	1,181	1,181	1,181	1,181
	kNN	5,103	1,181	5,103	4,131
	DBSCAN	5,186	1,181	4,131	5,186

Color	Method	Total	Overlap with		
			Grubbs	kNN	DBSCAN
	Grubb's	12,091	12,091	12,091	12,091
	kNN	36,576	12,091	36,576	35,302
	DBSCAN	43,195	12,091	35,302	43,195

Figure 4. Overlap of recombinant genes detected by Grubbs', DBSCAN, and kNN. Overlap of recombinant genes detected by Grubbs' test, DBSCAN, and kNN were summarized in recombination removal of 94 diverse *E. coli* chromosomes (A) and recombination removal of 91 ST131 *E. coli* isolates (B). In A, genes identified by Grubbs' were a subset of genes identified by kNN or DBSCAN, and that kNN and DBSCAN had more than 80% of the identified genes in common. In B, the results showed that Grubbs' identified a subset of genes of kNN or DBSCAN, that 96.5% of genes identified by kNN were also identified by DBSCAN, and that DBSCAN identified the largest number of genes.

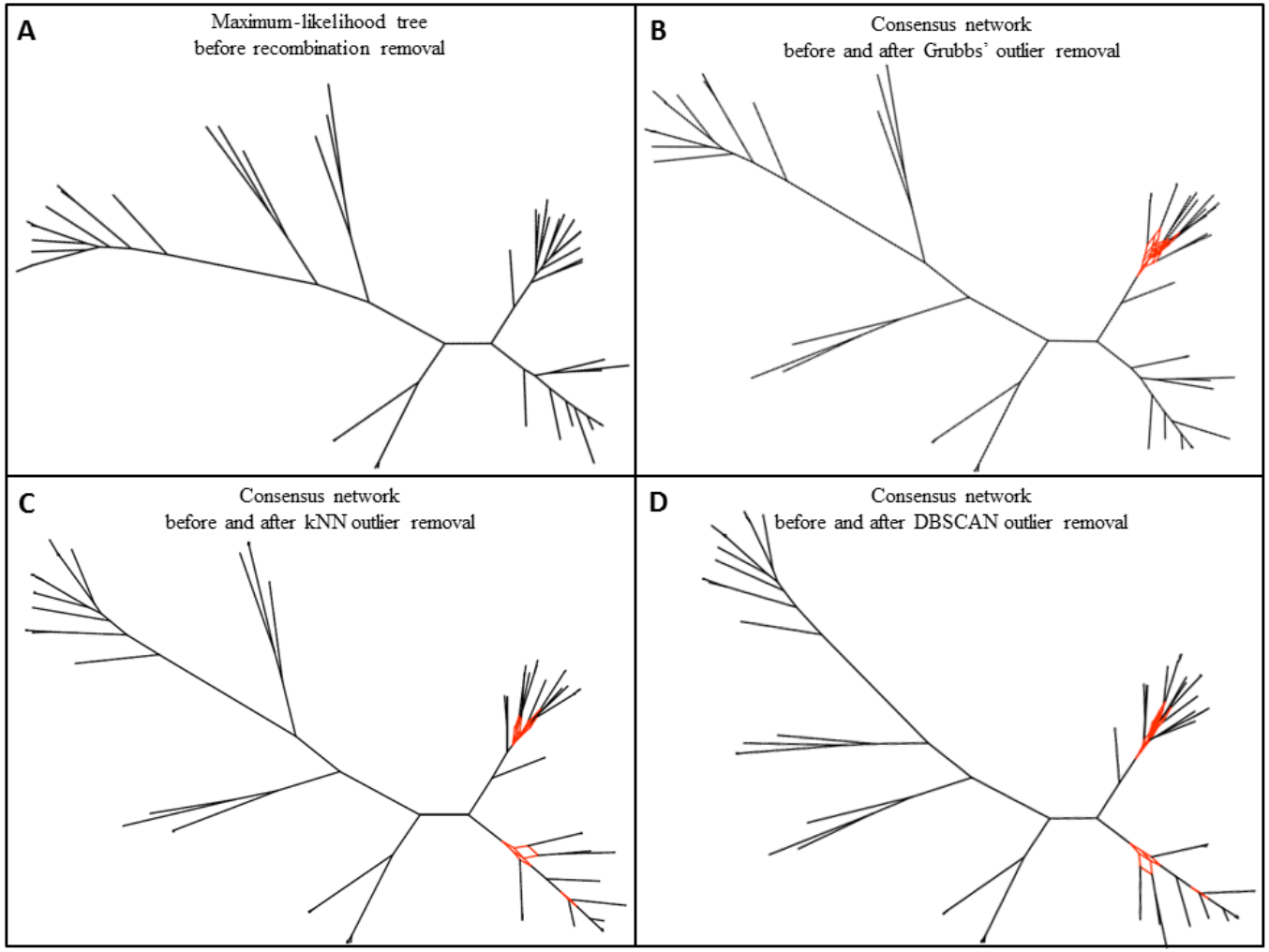


Figure 5. Phylogenetic tree change after recombination removal in 94 diverse *E. coli* isolates. Maximum-likelihood trees were constructed using sequences before and after recombination removal by ReRCoP with each of the three methods, and were compared by constructing a consensus network, where incompatible splits were highlighted in red to show the difference. Consensus networks show that recombination removal did not affect the major branching of the phylogenetic tree, but had an impact on the topology within branches.

3.3.6 Recombination removal using a sliding-window approach of Illumina sequencing reads of 91 ST131 *E. coli* isolates

The 91 ST131 *E. coli* isolates represent closely related bacteria: of the same sequence type and some may be belong to one or more outbreaks, and can use complete genomes instead of the core genomes due to their similar gene composition and organization. Consensus sequence for each isolate was constructed against the ST131 *E. coli* NA114 genome. A sliding-window approach was used for recombination removal using all three outlier detection methods in ReRCoP. The job finished within 11 min. ReRCoP identified (out of the total of 904,722 genes) more recombinant genes than in diverse bacterial genomes: 12,091 for Grubbs', 36,576 for kNN, and 43,195 for DBSCAN, which was consistent with the observed higher detection sensitivity in closely related bacterial populations in the simulations. Number of overlapped genes identified (Figure 4B) showed that Grubbs' identified a subset of genes of kNN or DBSCAN, that 96.5% of genes identified by kNN were also identified by DBSCAN, and that DBSCAN identified the largest number of genes. The results are consistent with the simulation results that in closely related bacterial strains, Grubbs', kNN, and DBSCAN had increasing sensitivity and decreasing specificity. Consensus networks were built on maximum-likelihood trees to visualize the differences generated by recombination removal (Figure 6). More extensive differences were observed compared to diverse bacterial strains. This is a result of more significant changes in the relative distances, which can be due to the larger number of recombinant genes detected and removed, and the smaller differences between closely related strains before removal.

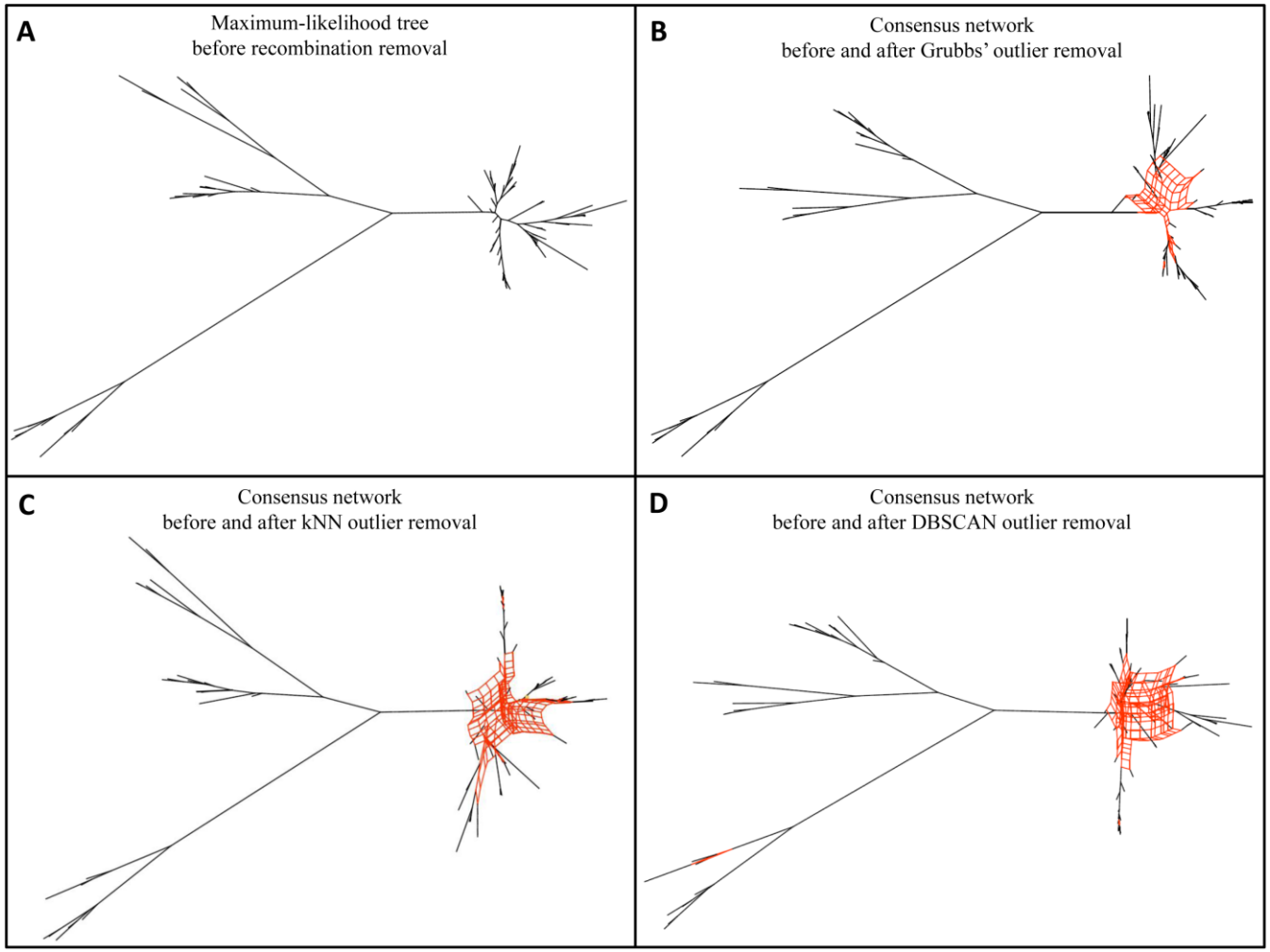


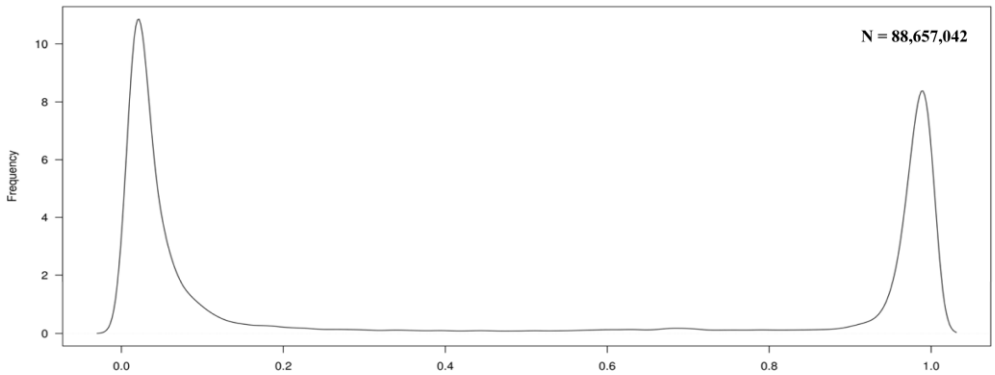
Figure 6. Phylogenetic tree change after recombination removal in 91 ST131 *E. coli* isolates. Maximum-likelihood trees were constructed using sequences before and after recombination removal by ReRCoP with each of the three methods, and were compared by constructing a consensus network, where incompatible splits were highlighted in red to show the difference. More extensive differences were observed compared to diverse bacterial samples. This is a result of more significant changes in the relative distances, which can be due to the larger number of recombinant genes detected and removed, and the small differences between closely related bacterial strains before removal.

3.3.7 Choice of program parameters

3.3.7.1 Choice of parameter in core gene identification

In core gene identification, a gene coding sequence would be searched in each genomic sequence using nucleotide BLAST, where a similarity value would be calculated from the BLAST output file. Based on the experiment, 438,159 genes were compared with each other using nucleotide BLAST, after which similarity values would be calculated from the output. Of the 191,982,871,122 similarity values, 191,894,214,080 (99.95%) were 0. Distribution of the non-zero similarity values was summarized with a density plot in Figure 7, showing two clear peaks of similarity values, one suggesting potentially same gene, and the other potentially different genes. When breaking down the similarity values into intervals, the interval (0.45, 0.5] had the least number of similarity values. The default threshold was thus set to be 0.49, a value within this interval.

While we do BLAST in ReRCoP, one gene is taken as the query sequence, while the other as the reference sequence. Which is used as the reference sequence affects the similarity value by affecting the length of the reference sequence, whose effect is thus assessed. For each pair of gene coding sequences, two similarity values were calculated, using either member as the reference sequence, respectively. Only 0.00089% of the pairs had one similarity value larger than 0.49, while the other smaller than 0.49, which is a strong indication that under this similarity value threshold, which sequences is used as the query sequence does not have a large effect on core gene identification.



Interval	(0,0.05]	(0.1,0.15]	(0.2,0.25]	(0.3,0.35]	(0.4,0.45]	(0.5,0.55]	(0.6,0.65]	(0.7,0.75]	(0.8,0.85]	(0.9,0.95]
Count	36,775,338	2,270,777	778,762	493,827	400,361	418,301	581,157	575,707	527,530	2,244,437
Interval	(0.05,0.1]	(0.15,0.2]	(0.25,0.3]	(0.35,0.4]	(0.45,0.5]	(0.55,0.6]	(0.65,0.7]	(0.75,0.8]	(0.85,0.9]	(0.95,1]
Count	8,086,089	1,185,728	612,774	437,979	363,971	503,926	714,053	535,400	673,318	30,477,607

Figure 7. Summary of similarity value distribution by density plot and interval breakdown. Non-zero similarity values were summarized with a density plot and statistics of the values. Two clear peaks of similarity values were observed, one suggesting potentially same gene, and the other potentially different genes. When breaking down the similarity values into intervals, the interval (0.45, 0.5] had the least number of similarity values.

3.3.7.2 Choice of parameters in kNN outlier detection

In kNN outlier detection in ReRCoP, absolute difference is used to measure the distances between data points in the univariate dataset. Any data point whose distance to its k^{th}-nearest neighbor is larger than a distance threshold (*radius*) would be detected as an outlier. Simulations were conducted to evaluate selection of parameters (*k*: number of nearest neighbors to consider, and *radius*: distance threshold). The simulation results were summarized in Table 4. Here, distance thresholds (d_{thresh}) represent *radius*, and the percentages (p_{thresh}) represent *k.* By default, kNN outlier removal in ReRCoP uses parameters of 0.2 for *k* (in the unit of total data points) and 1.5 for *radius* (in the unit of standard deviation of data points), which, based on the simulation, gives sensitivity of 0.89 and specificity of 0.98. Though similar performance can also be achieved by using larger *k* and *radius*, a smaller *k* was chosen to allow non-outliers to be in more than one tight cluster while only one was simulated.

Table 4. Sensitivity and specificity of kNN outlier detection using different *k* and *radius*.

p_{thresh}	d_{thresh}: 0.5		d_{thresh}: 1		d_{thresh}: 1.5		d_{thresh}: 2	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
0.05	0.9678	0.9646	0.7886	0.9919	0.6202	0.9983	0.4529	0.9997
0.1	1.0000	0.9129	0.9094	0.9778	0.7425	0.9947	0.5747	0.9989
0.15	1.0000	0.8470	0.9903	0.9579	0.8227	0.9890	0.6562	0.9976
0.2	1.0000	0.7661	1.0000	0.9339	0.8873	0.9813	0.7214	0.9957
0.25	1.0000	0.6638	1.0000	0.9042	0.9427	0.9706	0.7772	0.9927
0.3	1.0000	0.5341	1.0000	0.8687	0.9921	0.9574	0.8265	0.9887
0.35	1.0000	0.3390	1.0000	0.8264	1.0000	0.9406	0.8734	0.9835
0.4	1.0000	0.0000	1.0000	0.7775	1.0000	0.9208	0.9167	0.9763
0.45	NA	NA	1.0000	0.7188	1.0000	0.8956	0.9597	0.9667
0.5	NA	NA	1.0000	0.6490	1.0000	0.8656	0.9972	0.9548
0.55	NA	NA	1.0000	0.5623	1.0000	0.8283	1.0000	0.9389
0.6	NA	NA	1.0000	0.4499	1.0000	0.7849	1.0000	0.9195
0.65	NA	NA	1.0000	0.2857	1.0000	0.7283	1.0000	0.8932
0.7	NA	NA	1.0000	0.0001	1.0000	0.6593	1.0000	0.8593
0.75	NA	NA	NA	NA	1.0000	0.5685	1.0000	0.8138
0.8	NA	NA	NA	NA	1.0000	0.4439	1.0000	0.7511
0.85	NA	NA	NA	NA	1.0000	0.2271	1.0000	0.6603
0.9	NA	NA	NA	NA	1.0000	0.0001	1.0000	0.5132
0.95	NA	NA	NA	NA	NA	NA	1.0000	0.1669
1	NA	NA	NA	NA	NA	NA	1.0000	0.0002

p_{thresh}	d_{thresh}: 2.5		d_{thresh}: 3		d_{thresh}: 3.5		d_{thresh}: 4	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
0.05	0.2859	1.0000	0.1193	1.0000	0.0000	1.0000	0.0000	1.0000
0.1	0.4072	0.9998	0.2398	1.0000	0.0747	1.0000	0.0000	1.0000
0.15	0.4884	0.9995	0.3216	0.9999	0.1554	1.0000	0.0011	1.0000
0.2	0.5535	0.9992	0.3865	0.9999	0.2189	1.0000	0.0542	1.0000
0.25	0.6094	0.9985	0.4425	0.9997	0.2750	1.0000	0.1090	1.0000
0.3	0.6597	0.9975	0.4920	0.9995	0.3253	0.9999	0.1592	1.0000
0.35	0.7071	0.9962	0.5399	0.9993	0.3725	0.9999	0.2055	1.0000
0.4	0.7511	0.9942	0.5842	0.9988	0.4163	0.9998	0.2489	1.0000
0.45	0.7938	0.9914	0.6257	0.9982	0.4589	0.9997	0.2921	1.0000
0.5	0.8350	0.9879	0.6692	0.9973	0.5011	0.9994	0.3344	0.9999
0.55	0.8773	0.9829	0.7113	0.9960	0.5438	0.9992	0.3764	0.9999
0.6	0.9194	0.9756	0.7536	0.9941	0.5870	0.9988	0.4189	0.9998
0.65	0.9636	0.9657	0.7976	0.9911	0.6294	0.9981	0.4628	0.9997
0.7	0.9995	0.9521	0.8437	0.9871	0.6779	0.9971	0.5098	0.9994
0.75	1.0000	0.9323	0.8936	0.9802	0.7279	0.9955	0.5604	0.9991
0.8	1.0000	0.9024	0.9495	0.9690	0.7841	0.9922	0.6160	0.9984
0.85	1.0000	0.8564	1.0000	0.9505	0.8475	0.9868	0.6816	0.9970
0.9	1.0000	0.7765	1.0000	0.9146	0.9284	0.9740	0.7628	0.9935
0.95	1.0000	0.6035	1.0000	0.8238	1.0000	0.9368	0.8834	0.9820
1	1.0000	0.0005	1.0000	0.0012	1.0000	0.0025	1.0000	0.0972

- d_{thresh} corresponds to *k*, and is in the unit of standard deviation of the data points
- p_{thresh} corresponds to radius, and is in the unit of total number of data points

3.3.7.3 Choice of parameters in DBSCAN outlier detection

Parameter selection in DBSCAN can be based on simulations in the kNN outlier detection with the aim that non-outliers are either core points or reachable points and outliers are neither core points nor reachable points. The probability of every non-outlier point to be identified as a core point with parameters *minPts* and *eps* is the same as the specificity of kNN outlier detection with parameters *k=minPts, radius=eps,* thus can be figured out from Figure 7, from which the *minPts* and *eps* can be selected based on the desired specificity. To allow some non-outlier points to be reachable points, we can also use larger *minPts*. The probability of an outlier point to be taken as a core point with parameters *minPts* and *eps* is the same as 1-sensitivity of kNN outlier detection with parameters *k=minPts, radius=eps,* thus can be figured out from Figure 7. However, we should also exclude the cases where outliers are reachable points, thus we should decrease *eps* in order to increase outlier detection sensitivity. As a result, DBSCAN outlier removal in ReRCoP uses parameters of 0.2 for *minPts* (in the unit of total data points) and 1 for *eps* (in the unit of standard deviation of data points) by default.

3.4 Discussion

ReRCoP is a novel method for detecting and removing recombination from core genomes of large bacterial population samples for phylogenetic study. ReRCoP specifically aims to address the limitations of existing methods, and thus possesses the following four features that are distinct from other recombination detection methods: (1) ReRCoP can process whole genome sequences of a large number of bacterial isolates in a fast and computationally

efficient manner; (2) ReRCoP accepts both aligned genomic sequences, where sequences can be processed either gene by gene, or window by window, and unaligned genomics sequences, where core genomes would be identified, extracted, and processed gene by gene; (3) ReRCoP is robust to mutational hotspots and coldspots; and (4) ReRCoP can deal with both complete genomes and draft-quality assembled genomes.

Three recombination removal methods are implemented in ReRCoP: Grubbs' test, kNN, and DBSCAN. Grubbs' test is a statistical test, where a significance level is specified. The default value was set to 0.05 as usually used in statistical tests. When using default parameters, Grubbs' test has the lowest sensitivity and highest specificity. Though not as sensitive, Grubbs' test showed the best consistency and a balance between sensitivity and specificity when the sample size is small (10, for example), and is thus the best choice for studies of small sample sizes. For kNN and DBSCAN, simulations were conducted to assess the effect of parameters on detection sensitivity and specificity. Default parameters were set to balance the sensitivity and specificity, which can be adjusted based on the simulation results. Both kNN and DBSCAN have higher sensitivity and lower specificity than Grubbs' test and Gubbins. When the studied bacterial samples are closely related, DBSCAN has slightly higher sensitivity and comparable specificity when compared to kNN and is thus recommended to be used. However, when the studied bacterial samples are diverse, kNN performs better in detecting recombination that introduces a lower SNP density compared to the background level at a cost of slight decrease in the specificity, and is thus recommended for use.

ReRCoP adopts the strict criteria that genes present in all studied isolates are called core genes. For core genome identification, ReRCoP uses a simplified approach that core genomes are considered as composed by core genes without consideration of gene order or organization. More complex methods exist for core genome identification, which includes attempts to uncover the scaffolds of the genome, gene orders and gene adjacency [79, 80]. These are not as important for ReRCoP since it detects recombination gene by gene without using information of the surroundings.

Gene duplication is a common phenomenon in bacterial genomes [81] and is potentially problematic for recombination detection both using the core genome approach and the reference mapping approach. For ReRCoP, it is suggested to pre-process the input gene coding sequences by removing duplicated genes to avoid the likely overrepresentation of the duplicated genes in genomes containing single copies of the genes. When extracting the gene sequences, if more than one copy is identified, the one with the highest similarity would be chosen and extracted. Even by these measures, there is still no guarantee that the genes extracted are the same copy derived from a common ancestor. After all, it is hard to infer ancestry from duplicated genes.

One feature of ReRCoP is the capability of dealing with draft-quality genomes without a reference genome. In most cases, bacterial sequencing is conducted without purification to isolate the chromosome, making the sequencing reads a mixture of genetic sequences from chromosomes and various plasmids. As a result, many of the contigs have plasmid origins and should not be included in the phylogenetic analysis. This can however be resolved by first pre-processing the sequencing reads or assembled contigs to

exclude those belonging to plasmids. It is also possible to retain all sequencing reads for the analysis on the basis that genes on plasmids are neither conserved nor essential, and are thus unlikely to be shared by a diverse bacterial population and be featured as core genes. It is also probable that plasmids can be shared among outbreak isolates and bacteria are 'clonal' in the transmission, where the variations on the plasmids can bear useful information on the phylogenetic relationships as well.

As ReRCoP processes the sequences gene by gene, it is possible that ReRCoP fails to detect recombination events that either affect only a small fraction of a gene, or affect only several positions. Also, even when the entire gene in a sequence is the result of recombination, ReRCoP can fail to recognize a recombination event if the degree of variation between sequences at this gene is similar. ReRCoP fundamentally identifies recombinant genes that possess a significant degree of SNP density change.

In the simulations to assess the performance of ReRCoP and to compare with Gubbins, uniform mutation rates were used for the non-recombinant sequences without intentional introduction of mutational hotspots and coldspots. The fact that ReRCoP adopts a vertical comparison instead of a horizontal comparison as adopted by other methods like Gubbins makes ReRCoP more robust to uneven mutation rate, particularly in the presence of mutational hotspots and cold spots.

It can be inferred from the analysis of diverse *E. coli* chromosomes and ST131 *E. coli* isolates that removing recombination does not have a significant impact on the phylogeny of diverse strains, but can greatly influence the inferred relationships of closely related strains. This is consistent with the

results of the simulations that ReRCoP is less sensitive in detecting recombination in diverse bacterial strains but possesses much higher power in detecting recombination in closely related bacterial strains.

3.5 Conclusion

In this chapter, I introduced ReRCoP, a novel method for detecting recombination that is useful in bacterial genomes with the following features: (1) ReRCoP is able to efficiently process whole genome sequences of a large number of bacterial isolates; (2) ReRCoP is able to automatically identify and extract the core genomes; (3) ReRCoP is robust to mutational hotspots and coldspots; and (4) ReRCoP can deal with draft-quality assembled genomes. Simulations were conducted to show that ReRCoP is useful for detecting recombination caused by both HGT and homologous recombination. Comparison with Gubbins showed that ReRCoP is more time and memory efficient, more sensitive while less specific. ReRCoP was applied in analysis of both diverse and closely related bacterial strains, showing that recombination removal has a larger effect on closely related strains. ReRCoP would be a useful tool in bacterial phylogenetic study by eliminating the adverse effects of recombination.

Chapter 4

Local transmission and global dissemination of New Delhi metallo-beta-lactamase (*bla*$_{NDM}$): a whole genome analysis

The content of this chapter has been published as [82]. Reproduction of figures and tables is permitted by the publisher.

4.1 Background

The emergence of carbapenem-resistant *Enterobacteriaceae* (CRE) has become an important global health threat. CRE are primarily recognized in health care settings [83], with the prevalence in clinical samples increasing globally [84–88]. Outcomes of CRE infections are poor, where mortality associated with infections can reach over 40% [89, 90]. With the widespread dissemination of extended-spectrum beta-lactamases, carbapenems are the last class of safe and effective antimicrobials for treating multidrug-resistant Gram-negative bacterial infections, the effectiveness of which has been greatly undermined by CRE [91]. As a result, there is a pressing need to understand the transmission pathways of carbapenemases to inform infection control, which remains the main intervention to face the challenge of CRE.

New Delhi metallo-beta-lactamase (bla_{NDM}) was first detected in 2008 in a *K. pneumoniae* isolate from a Swedish traveler returning from the Indian subcontinent [92]. Since then, bla_{NDM} has been documented in all continents, with the earliest archived bla_{NDM}-positive sample from 2005 [93]. Two identical bla_{NDM}-positive plasmids (pTR3 and pTR4) have been reported in Singapore in unrelated *K. pneumoniae* isolates [94]. Compared with other carbapenemases, the spread of bla_{NDM} is characterized by alarming public health features: (1) broad Gram-negative bacterial host range, including highly virulent bacteria such as *Vibrio cholera* and *Shigella boydii* [95]; (2) frequent acquisition among *E. coli* and *K. pneumoniae*, which are Gram-negative species carried as gut flora and able to survive in inanimate environments; (3) widespread presence in the Indian subcontinent, Southeast and East Asia,

home to the largest human populations globally; and (4) co-carriage with other resistance genes on the *bla*$_{NDM}$-bearing plasmids [96].

Multiple seminal investigations have focused on determining the international and local transmission patterns of chromosome-mediated antimicrobial resistance [97–100]. However, there remained many unanswered questions about the spread of plasmid-borne antimicrobial resistant genes. While mass global travel and widespread antibiotic use have been widely recognized as population risk factors associated with the dispersal of *bla*$_{NDM}$ [96], investigation is still needed regarding the genomic factors associated with its rapid spread [101]. Antimicrobial resistance genes are often carried by mobile genetic elements like plasmids and transposons [102], which may also carry integrons or other gene mobilization elements [103, 104]. A key biological challenge in understanding plasmid-borne gene molecular epidemiology is the capability to exploit three tiers of gene spread: (1) inter-plasmid gene module transposition; (2) inter-bacteria plasmid conjugation; and (3) bacteria spread among humans, animals and the environment [96]. While SNP-based phylogenetic methods are proven to be successful in understanding transmission of chromosome-mediated antimicrobial resistance, these methods are ill-suited to determining the dynamics of multi-tiered gene flow of plasmid-mediated antimicrobial resistance due to the lack of conserved genomic regions in diverse plasmids.

By moving beyond conventional SNP-based phylogenetic study to a plasmid clustering approach based on distances measured by the degree of gene sharing and the similarity of shared genes between different plasmids, I analyzed a combined collection of all GenBank complete plasmid sequences

within Gram-negative bacterial hosts to date, thus having an unprecedented opportunity to profile the global dissemination of this important resistance gene. A total of 2,749 complete plasmid sequences from NCBI GenBank database were included in this study, of which 39 are *bla*$_{NDM}$-positive. This enabled an analysis of the largest collection of sequences to date, providing a comprehensive description on the distribution and genetic movement of *bla*$_{NDM}$. Moreover, in order to investigate the local transmission of *bla*$_{NDM}$ to compare with its global dissemination, 11 *bla*$_{NDM}$-positive CRE isolates in a local hospital were sequenced [105], from which the transmission pattern was inferred based on the identity of *bla*$_{NDM}$-positive plasmids and phylogenetic study of the chromosomes, in combination with the patients' records. In summary, this study suggested that *bla*$_{NDM}$-positive plasmid diversity is very low in a local transmission setting characterized by plasmid conjugation and bacteria spread, while the global *bla*$_{NDM}$-positive plasmids, due to the transposition of the *bla*$_{NDM}$ gene cassette into different plasmids, are highly variable, which can be clustered into 7 distinct clusters correlated with plasmid incompatibility group and geographical distribution. These findings advance understanding of plasmid-mediated antimicrobial resistance spread both locally and globally.

4.2 Methods

4.2.1 Clinical isolates

Tan Tock Seng Hospital (TTSH) is Singapore's second largest acute-care hospital with 36 clinical and allied health departments and more than 1400 beds. The first case of carbapenemase-producing *Enterobacteriaceae* (CPE) in

TTSH was detected in September 2010 (subject 16). From September 2010 to October 2011, a further 7 patients with CPE were detected, of which 2 were detected based on screening cultures. The infection control response to a new bla_{NDM}-positive patient detected in the course of routine testing included strict isolation of the patient, contact tracing within the same ward and in previously admitted wards, and screening of these contacts with rectal swabs for CPE carriage using draft guidelines issued by CDC [106]. Age, gender, travel history, history of ward locations and clinical diagnoses were collected by retrospective case-chart review.

4.2.2 Genome assembly

Sequencing reads have been submitted to the European Nucleotide Archive (ENA) under accession PRJEB13304. *De novo* assembly was performed using Velvet [27], parameters of which were optimized by VelvetOptimiser with k-mer lengths ranging from 55 to 63. For all the 11 isolates, VelvetOptimizer achieved the best assembly at the k-mer length of 63.

The bacterial species were identified by searching the assembled contigs in the NCBI 'nt' database. If the top five hits for a contig are all chromosomal DNA, this contig is assigned to the chromosome and the hits are taken as candidate chromosomes. For each isolate, candidate chromosomes of at least one contig would each be used as the reference sequence, against which all the contigs would be aligned. The genome coverage by the contigs would then be calculated, where the candidate chromosome with the highest genome coverage would be taken as the most similar bacterial strain and its species would be identified as the bacterial species of the isolate. MLST of *E. coli* and

K. pneumoniae isolates was inferred using MLST 1.8 provided by the CGE server [107].

4.2.3 Molecular epidemiology

Sequencing reads were aligned to the reference genome (JJ1886 [GenBank:CP006784.1] for *E. coli*, and HS11286 [GenBank:CP003200.1] for *K. pneumoniae*) using BWA-MEM [31]. Single-nucleotide variants were called using SAMtools [33]. Positions with less than 10 reads or with a minor allele frequency between 0.25 and 0.75 would be marked as 'unknown' data. Variants would then be called if the alternate allele frequency is above 0.75. Maximum likelihood phylogenetic trees were constructed using RAxML [37], where a substitution model of GTRGAMMA was used and rapid bootstrap analysis was conducted on 500 runs.

4.2.4 *bla*$_{NDM}$-positive plasmid identification

For each isolate, the contig with the *bla*$_{NDM}$ gene was first identified and extracted, after which the contig sequence was searched in the NCBI 'nt' database for complete plasmid sequences with more than 2000 bp identity. The similar complete plasmid sequences were then each used as the reference sequence, against which all the contigs were aligned to calculate the sequence coverage by the contigs. Complete sequences with the highest sequence coverage would then be taken as the most similar plasmids.

4.2.5 Plasmid mapping, genome coverage calculation and variant calling

Novoalign was used for read mapping against a reference plasmid sequence, after which realignment was conducted with GATK IndelRealigner [34], and the coverage was calculated with GATK DepthOfCoverage. Variants were called with UnifiedGenotyper in GATK, with filtering criteria: "MQ < 40.0, QD < 2.0, FS > 60.0, HaplotypeScore > 13.0".

4.2.6 Complete plasmid sequences

All the 2,749 available complete plasmid sequences within Gram-negative bacterial host in the NCBI plasmid database (April 2014) were downloaded for analysis, of which 39 are bla_{NDM}-positive. Information on sampling location and date, sample source, subject's travel history, host bacterial species and bacterial antimicrobial resistance phenotypes were obtained from GenBank entries or accompanying references.

4.2.7 Plasmid clustering

Plasmid clustering was conducted based on the virtual hybridization method as described by Zhou *et. al.* [71] to investigate the similarity of the diverse complete plasmid sequences.

For each plasmid, all coding sequences, as determined by their original investigators, were downloaded from NCBI. Duplicate genes on the same plasmid, defined as coding DNA sequences having similarity value (length of matching sequences * BLAST identity / length of reference sequence) above 0.45, were removed. This resulted in a set of 234,450 genes. Additionally, insertion sequences within each plasmid were detected using IS Finder

(https://www-is.biotoul.fr/) with default parameters at a cut-off e-value of $1e^{-20}$, which identified 1,496 unique insertion sequences.

For genetic sequence comparison, a similarity score is calculated as 2*(length of matching sequences)*(BLAST identity)/ (length of reference sequence + length of matching sequences). The 2,749 complete plasmid sequences were then compared using nucleotide BLAST algorithm against each of the 234,450 genes and 1,496 insertion sequences to calculate a similarity score, which resulted in a 2,749 by 235,946 matrix of similarity scores. A hypothetical plasmid sequence with all similarity scores set to zero was used as outgroup.

To achieve computational tractability, 1,000 random matrices were generated, each of which was composed of 20% of the similarity score matrix's columns that were randomly selected without replacement, showing the similarity scores represented by 20% randomly selected genes. For each matrix of similarity scores, pair-wise Euclidean distances between plasmid sequences were calculated and formulated into a distance matrix, after which a Neighbor-Joining tree was constructed with the 'neighbor' program in PHYLIP [40]. A consensus tree was constructed using the 'consense' program in PHYLIP with the majority rule as the consensus type.

Clusters of *bla*$_{\text{NDM}}$-positive plasmid based on the consensus tree were defined using a stringent criterion of having at least 2 unique *bla*$_{\text{NDM}}$-positive plasmids, with all internal nodes having ≥99% support at 1000 bootstraps.

4.2.8 Phylogenetic tree for cluster refinement

Cluster refinement was conducted for each cluster respectively. For each cluster, coding DNA sequences present in all plasmid sequences with a nucleotide BLAST e-value less than $1e^{-5}$ and an identity above 80% were extracted, aligned, and concatenated. Maximum likelihood phylogenetic trees were constructed using RAxML [37], where a substitution model of GTRGAMMA was used and rapid bootstrap analysis was conducted on 500 runs.

4.2.9 Incompatibility groups of plasmids

To determine the incompatibility (Inc) groups of plasmids, nucleotide BLAST was used to find sequences for specific Inc groups that would produce theoretical PCR amplicons for known Inc group sequences [108].

4.2.10 Comparative genomics

Plasmid sequences were compared and visualized with the Artemis comparison tool ACT [109].

4.3 Results

4.3.1 Local *bla*$_{NDM}$-positive plasmid diversity in a single hospital

The first 11 CPE isolates from 8 patients in a single Singapore hospital were isolated, of which the patient demographics and sample features were summarized in Table 5 and Figure 8. The median duration of hospitalization to positive CPE culture was 3 days (range: 1 to 153 days). Six patients (subjects 16, 11, 1, 41, 51 and 53) had *bla*$_{NDM}$ detected on clinical cultures. One patient

(subject 21) was co-infected with 4 CPE isolates, where 2 different strains of *Enterobacteriaceae* were isolated from the patient's stool and urine samples, respectively. Of the 8 patients, only two had travelled out of Singapore in the past 2 years, including subject 21, who had travelled to Australia and subject 41, who had travelled to Malaysia. Whole genome sequencing was conducted on Illumina MiSeq, with the sequencing statistics summarized in Table 6.

Table 5. Patient demographics and sample features.

Subject ID	Travel History	Clinical Diagnosis	Sample ID
16	NA	Colonization	EN-M80M-U-060910
11	NA	Disease	KP-F78C-U-090910
1	NA	Colonization	EC-M94C-U-220910
21	Australia	Colonization	KP-F86E-U-141010
		Colonization	KP-F86E-R-141010
		Colonization	EC-F86E-U-141010
		Colonization	EC-F86E-R-141010
41	Malaysia	Colonization	EC-M59C-U-101210
46	NA	Colonization	EC-M28M-R-141210
51	NA	Disease	EC-F76C-B-220911
53	NA	Disease	EC-F60C-U-191011

Subject ID	MLST	Identity of bla_{NDM}-encoding plasmid	Rationale for sample
16	NA	pTR3	Clinical Sample
11	437	pNDM-KN*	Clinical Sample
1	410	pTR3	Clinical Sample
21	48	pTR3	Clinical Sample
	48	pTR3	Clinical Sample
	69	NA	Clinical Sample
	69	pTR3	Clinical Sample
41	131	pTR3	Clinical Sample
46	131	pTR3	Contact Screening for Index Subject 41
51	205	pNDM_MGR194*	Clinical Sample
53	131	pTR3	Clinical Sample

Sample ID format: Organism-Gender/Age/Race-Specimen site-Date of Isolation (DD/MM/YY)
Organism: EC = *Escherichia coli*, KP = *Klebsiella pneumoniae*, EN = *Enterobacter cloacae*.
Gender: F = Female, M = Male.
Race: C = Chinese, E = Eurasian, M = Malay.
Specimen site: U = Urine, R = Rectal swab, B = Bile.

* Closest reference plasmid identified based on minimum 75% reference sequence coverage.

Table 6. Summary of Illumina sequencing and *de novo* assembly statistics.

Sample ID	Illumina sequencing statistics			
	# Reads	Reads per pair	# Bases	Estimated coverage*
EC-M94C-U-220910	4,638,924	2,319,462	1,159,731,000	~230X
KP-F78C-U-090910	4,993,178	2,496,589	1,248,294,500	~250X
EN-M80M-U-060910	2,551,658	1,275,829	637,914,500	~125X
KP-F86E-U-141010	5,481,114	2,740,557	1,370,278,500	~275X
KP-F86E-R-141010	5,971,648	2,985,824	1,492,912,000	~300X
EC-F86E-U-141010	4,020,020	2,010,010	1,005,005,000	~200X
EC-F86E-R-141010	4,866,162	2,433,081	1,216,540,500	~245X
EC-M59C-U-101210	3,610,924	1,805,462	902,731,000	~180X
EC-M28M-R-141210	3,531,240	1,765,620	882,810,000	~175X
EC-F76C-B-220911	3,694,724	1,847,362	923,681,000	~185X
EC-F60C-U-191011	5,358,750	2,679,375	1,339,687,500	~270X

* Coverage is estimated by Total number of bases (bp)/5,000,000 (bp/genome)

Sample ID	*De novo* assembly statistics				
	# Contigs	Total length (bp)	Maximum length (bp)	N50	N90
EC-M94C-U-220910	283	4,924,755	311,367	119,021	28,367
KP-F78C-U-090910	153	5,517,983	679,086	269,381	86,172
EN-M80M-U-060910	250	5,360,533	262,681	143,510	35,933
KP-F86E-U-141010	145	5,628,326	718,541	370,983	84,568
KP-F86E-R-141010	360	5,847,032	540,865	221,458	27,866
EC-F86E-U-141010	281	5,471,732	428,602	161,290	30,311
EC-F86E-R-141010	301	5,515,296	381,553	131,304	26,186
EC-M59C-U-101210	248	5,278,528	529,288	172,834	31,299
EC-M28M-R-141210	171	5,267,509	452,523	173,849	41,299
EC-F76C-B-220911	530	5,238,278	498,227	116,800	22,645
EC-F60C-U-191011	236	5,314,797	411,061	177,269	25,869

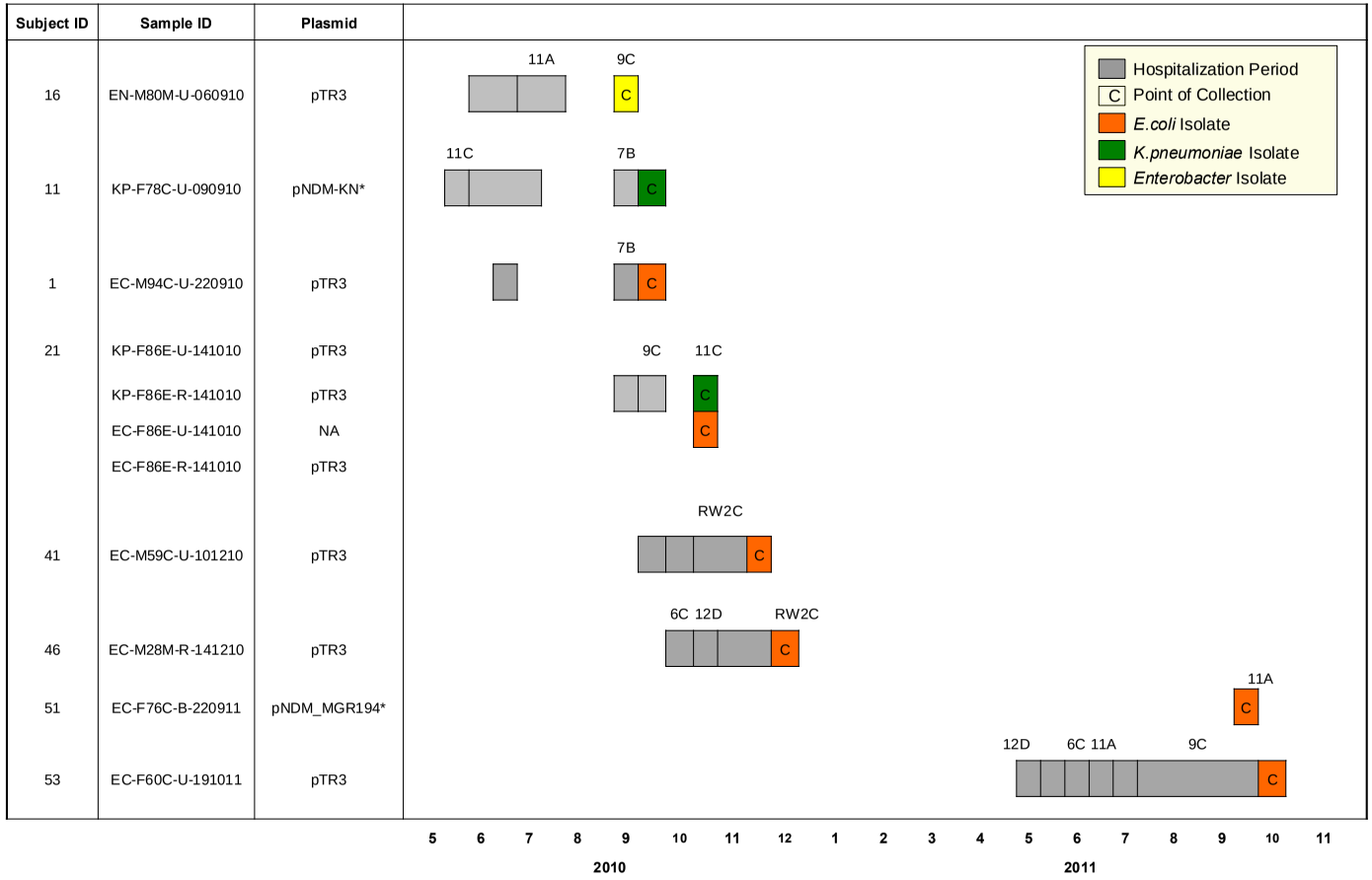


Figure 8. Patient transmission dynamics of local bacterial samples. The *bla*$_{NDM}$ cases were identified in a local hospital from 2010 to 2011 as represented in the timeline. Each patient is represented by a horizontal track. Subject ID, sample ID and *bla*$_{NDM}$-positive plasmid found in the isolate are indicated in the first 3 columns. Patient's stay in the same ward is denoted by gray box. Only wards with ≥2 reported *bla*$_{NDM}$ cases are indicated in the diagram. *: closest reference plasmid identified based on minimum 75% sequence coverage.

Plasmid identification was conducted with *de novo* assembly in combination with candidate plasmid identification, plasmid mapping and genome coverage calculation as elaborated in the Methods. The *de novo* assembly statistics was summarized in Table 6. Among the 11 samples, 10 bla_{NDM}-positive plasmids were identified, of which 8 were identified as pTR3 [GenBank:JQ349086.2], 1 was identified as pNDM-KN [GenBank:NC_019153.1] with the last being identified as pNDM_MGR194 [GenBank:NC_022740.1] (Table 5). Plasmid identification was most confident for the 41,187 bp plasmid pTR3 (100% genome coverage in all the 8 identified samples at very high read depths) and the 46,253 bp plasmid pNDM_MGR194 (100% genome coverage in sample EC-F76C-B-220911 at reasonable read depths). The 162,746 bp plasmid pNDM-KN was identified in sample KP-F78C-U-090910 with 76.3% genome coverage at very high read depths. No bla_{NDM}-positive plasmid was detected in sample EC-F86E-U-141010. The genome coverage and read depths were summarized in Figure 9.

Variant calling was performed for the 8 samples containing pTR3, the most prevalent bla_{NDM}-positive plasmid, to compare the pTR3 plasmid sequences in respective samples with the reference pTR3 sequence [GenBank:JQ349086.2]. Inspection of the variants revealed that 7 pTR3 plasmid sequences were identical to the reference pTR3 sequence, while one pTR3 plasmid sequence had only one SNP compared to the reference pTR3 sequence. In EN-M80M-U-060910 (isolated from subject 16), the pTR3 sequence had one synonymous mutation at the coding region of a putative transposase (position 22107), resulting in a codon change of GCC→GCT.

These results showed that local *bla*$_{NDM}$-positive plasmids had limited diversity with the majority of the plasmids being identical copies of pTR3, which is a strong indication of clonal plasmid spread. The other two *bla*$_{NDM}$-positive plasmids had identities of pNDM-KN and pNDM_MGR194. The major differences between the three plasmids (pTR3, pNDM-KN, and pNDM_MGR194) strongly indicated independent plasmid introductions into the hospital ecology.

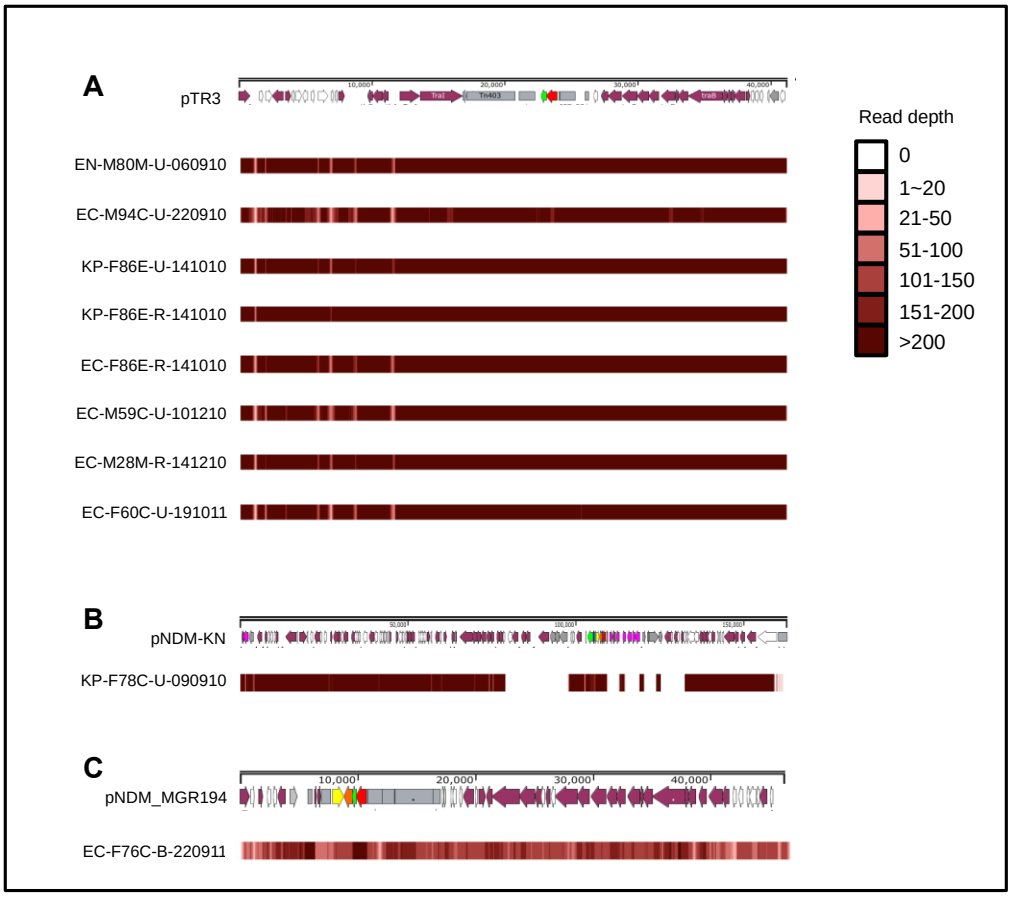


Figure 9. Read depths along the reference plasmid sequences based on Illumina MiSeq sequencing reads mapping. Sequencing reads were mapped to the plasmid sequences to calculate the read depths along the reference sequences of: pTR3 (A), pNDM-KN (B), and pNDM_MGR194 (C). In A, the read depths are reasonable for all samples along the complete pTR3 sequence, which strongly supports the presence of pTR3 in the samples. In B, 76% of pNDM-KN has been covered by the sample at reasonable read depths with major absences of genomic sequences. In C, the full length of pNDM_MGR194 is covered at reasonable read depths, strongly suggests its presence in the sample.

4.3.2 Bacterial host range at the local level

The bacterial species harboring *bla*$_{NDM}$-positive plasmids were: *E. coli* (7/11), *K. pneumoniae* (3/11) and *Enterobacter cloacae* (*E. cloacae*, 1/11) (Table 5). Of the 7 *E. coli* isolates, 3 were most similar to ST131 *E. coli* strain NA114 [GenBank:NC_017644.2], while the remaining isolates were most similar to ST23 *E. coli* strain APEC O78 [GenBank:NC_020163.1], ST597 strain UMN026 [GenBank:NC_011751.1] and ST1128 strain IAI1 [GenBank:NC_011741.1]. For the *K. pneumoniae* isolates, three *K. pneumoniae* strains was identified to be similar, including: ST11strain HS11286 [GenBank:NC_016845.1], ST23 strain NTUH-K2044 [GenBank:NC_012731.1], and ST23 strain 1084 [GenBank:NC_018522.1]. Consistent with previous report [110], there appeared to be no evidence of association between *Enterobacteriaceae* host species and specific plasmid identities.

Maximum likelihood phylogenetic trees were constructed for the bacterial chromosomes respectively for *E. coli* (Figure 10A) and *K. pneumoniae* (Figure 10B), both of which showed great diversity. The diversity of bacterial strains harboring pTR3 highlighted the propensity of *bla*$_{NDM}$-positive plasmids to spread via inter-bacteria plasmid conjugation, and would explain a key challenge in relying upon phylogenetic analysis alone to understand *bla*$_{NDM}$ dissemination.

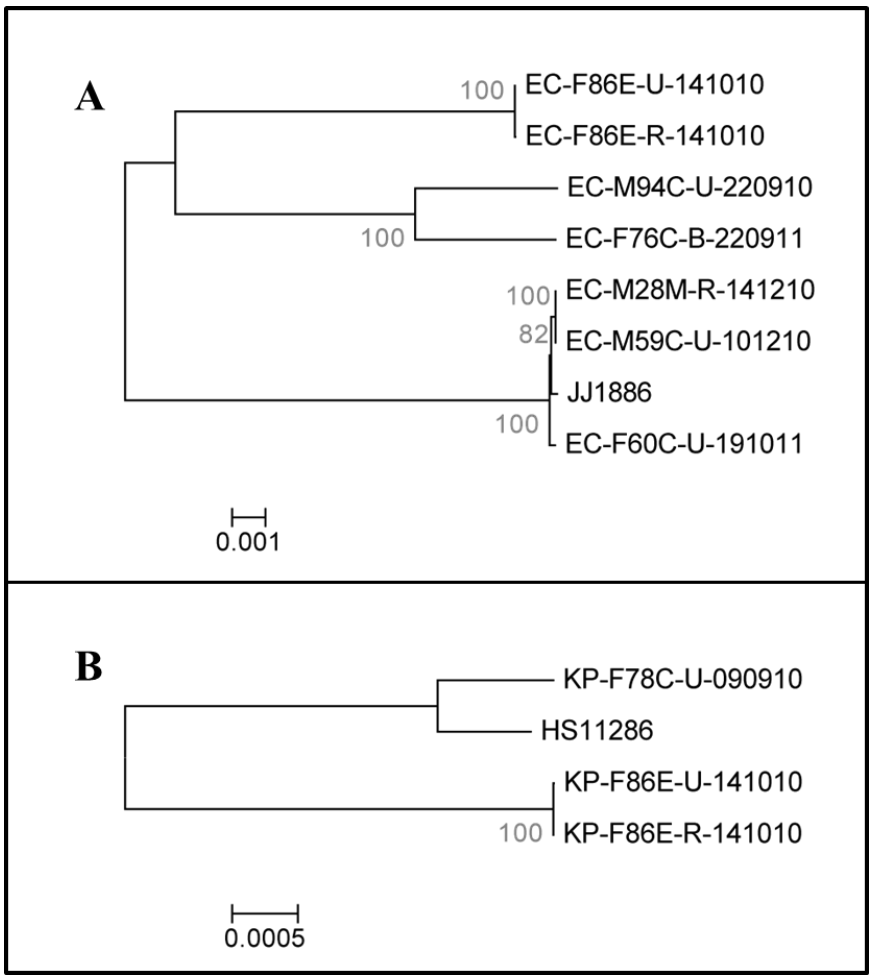


Figure 10. Whole-genome phylogenetic tree of local *bla*$_{NDM}$-positive bacteria. Maximum likelihood trees were constructed based on sequence alignments of *E. coli* (A) and *K. pneumoniae* (B). JJ1886 and HS11286 are the reference genomes for *E. coli* and *K. pneumoniae*, respectively. The branch lengths were calculated by RAxML and reflect the number of expected mutations per site. Bootstrap values are in a scale of 0 to 100, and are shown at each node in grey.

4.3.3 Inter- and intra- patient bacteria spread at the local level

Phylogenetic trees of the bacterial chromosomes in Figure 10 suggested clonal bacteria spread in 3 instances. The first instance involved ST131 *E. coli* detected in 2 patients – subjects 41 and 46, which clustered tightly as EC-M59C-U-101210 and EC-M28M-R-141210 in Figure 10A and differs by only 4 SNPs. The limited number of SNPs thereby suggested potential inter-patient bacteria spread between subject 41 and subject 46.

The other two instances involved bacteria with identical sequence types isolated from different body sites in the same patient (subject 21). KP-F86E-U-141010 (isolated from urine) and KP-F86E-R-141010 (isolated from rectal swab) are both ST48 *K. pneumoniae* that harbored the pTR3 plasmid, which clustered tightly in Figure 10B with 25 SNPs. EC-F86E-U-141010 (isolated from urine) and EC-F86E-R-141010 (isolated from rectal swab) are both ST69 *E. coli* that clustered tightly in Figure 10A with 58 SNPs. Sample EC-F86E-U-141010 was bla_{NDM}-negative and positive for $bla_{\text{IMP-1}}$, a class B carbapenemase. Subject 21 here represents a possible case of intra-host conjugation.

As discussed, the pTR3 plasmids remained 100% identical in all but 1 isolate at the nucleotide level in scenarios of inter- and intra-patient bacteria transfer, and inter-bacteria plasmid conjugation within the same host. These results suggested early spread of endemic plasmids at the local level was predominantly clonal.

4.3.4 Clustering of global plasmids from Gram-negative bacterial host

Complete genomic sequences of 2,749 plasmids within Gram-negative bacterial hosts were downloaded from the NCBI database. The median plasmid sequence length is 30,949 bp (range: 744 to 2,580,084), with the median number of genes annotated per plasmid being 36 (range: 1 to 2,235). Out of the 2,749 plasmids, the majority belong to the *Enterobacteriaceae* family (n=877, 31.9%), followed by *Spirochaetaceae* (n=405, 14.7%), *Rhodobacteraceae* (n=85, 3.1%), *Moraxellaceae* (n=81, 2.9%), and others (n=1301, 47.3%). Amongst, 39 plasmid sequences are bla_{NDM}-positive (Table

7). These plasmids were sampled from all continents except Antarctica over an 8 year period (2005 – 2013). Thirty-eight of the 39 *bla*$_{NDM}$-positive plasmid samples have a human origin, while one sample has an animal origin (pig). The median plasmid sequence length for *bla*$_{NDM}$-positive plasmids is 73,209 bp (range: 35,947 to 288,920), with the median number of genes annotated per plasmid being 89 (range: 31 to 372).

While construction of a SNP-based phylogenetic tree is the most common method to investigate evolutionary relationships among groups of organisms or strains, it is not applicable to plasmid phylogenetic study as there is no common genomic region shared among all the 2,749 complete plasmid sequences. An alternative approach based on the relative distances measured by the degree of gene sharing and the similarity of shared genes was applied to cluster the plasmids. The pair-wise distances based on a total of 234,450 genes and 1,496 insertion sequences were calculated as elaborated in the Methods, resulting in a Euclidean-distance derived distance matrix. A Neighbor-Joining tree was constructed with the distance matrix, upon which clustering analysis was based (Figure 11). The clustering of global plasmid showed high global plasmid diversity with *bla*$_{NDM}$-positive plasmids located in different clusters.

Table 7. Names and accession numbers of *bla*$_{NDM}$-positive plasmids.

Name	Accession
p271A	JF785549.1
pAB_D499	NZ_AGFH01000030.1
pAbNDM-1	JN377410.2
pGUE-NDM	JQ364967.1
pKOX_NDM1	JQ314407.1
pKp11-42	KF295829.1
pKPN5047	KC311431.1
pKPX-1	AP012055.1
pM131_NDM1	JX072963.1
pMC-NDM	HG003695.1
pMR0211	JN687470.1
pNDM102337	JF714412.2
pNDM10469	JN861072.1
pNDM10505	JF503991.1
pNDM-1_Dok01	AP012208.1
pNDM-1saitama01	AB759690.1
pNDM-AB	KC503911.1
pNDM-BJ01_1	JQ001791.1
pNDM-BJ01	KF702385.1
pNDM-BJ02	JQ060896.1
pNDM-BTR	KF534788.1
pNDMCFuy	HG428757.1
pNDM-CIT	JX182975.1
pNDM-HK	HQ451074.1
pNDM-HN380	JX104760.1
pNDM-KN	JN157804.1
pNDM-MAR	JN420336.1
pNDM-OM	JX988621.1
pNDM-US	CP006661.1
pPrY2001	KF295828.1
pRJA274	KF877335.1
pRJF866	KF732966.1
pTR3	JQ349086.2
pYE315203	JX254913.2
pABCA95	NC_019322.1
pEcNDM	NC_023909.1
pKpNDM1	NC_023911.1
pNDM-HF727	NC_023914.1
pNDM_MGR194	NC_022740.1

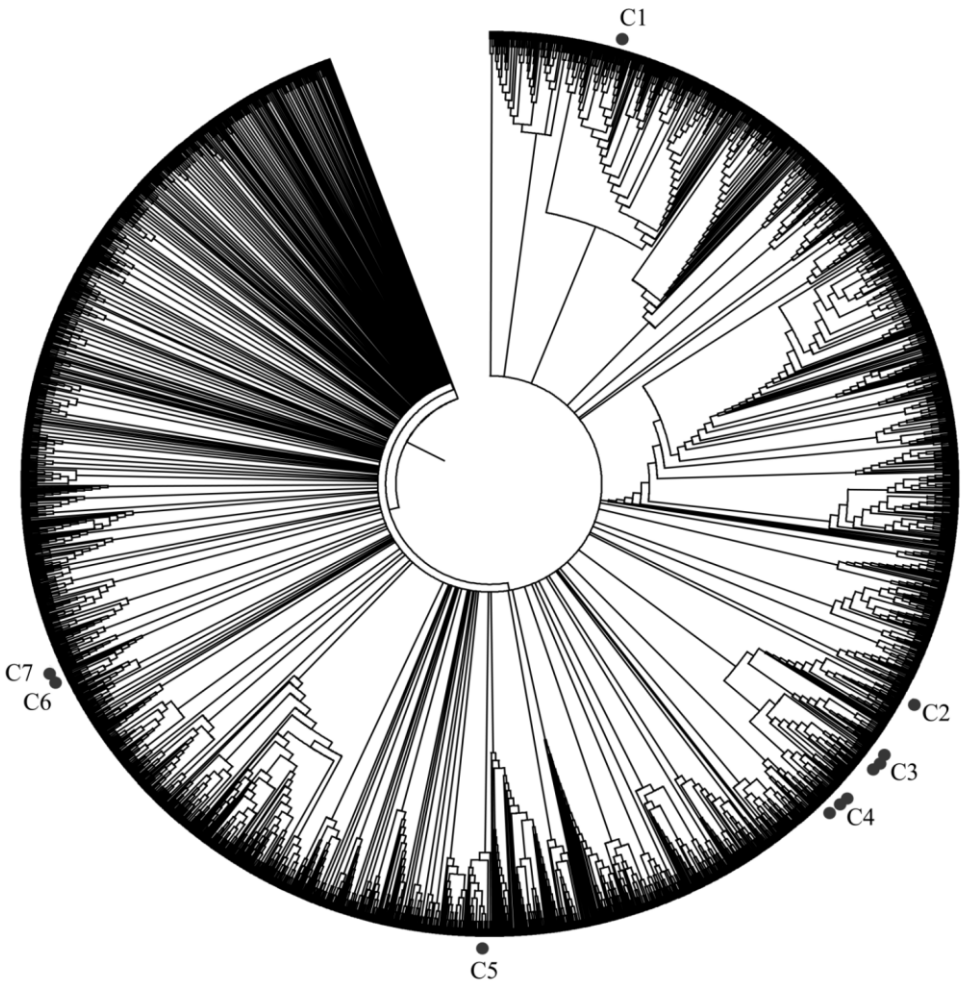


Figure 11. Clustering of global plasmids in Gram-negative bacteria hosts. The Neighbor-Joining tree consisting of 2,749 Gram-negative plasmid genomes was constructed to reflect the gene composition similarity of the plasmids. Seven bla_{NDM}-positive plasmid phylogenetic clusters were identified using stringent criteria (all internal nodes ≥99% bootstrap support, minimum 2 unique bla_{NDM}-positive plasmids). Clusters with bla_{NDM}-positive plasmids are indicated with dots and labeled C1-C7.

4.3.5 Clustering and phylogenetic study of *bla*$_{NDM}$-positive plasmids

Seven distinct clusters (represented by red dots in Figure 11) were identified to contain *bla*$_{NDM}$-positive plasmids, which range in size from 2 to 10 plasmids. For better clarity, the plasmids within the seven clusters were extracted and a new Neighbor-Joining tree was constructed, which is presented as Figure 12 with the plasmids' information.

The number of shared genes increased markedly for plasmids within the same cluster, allowing for the construction of a phylogenetic tree based on nucleotide sequence alignment within the shared regions. For clusters with more than three sequences, a concatenated alignment of the homologous genes was generated, after which a phylogenetic tree would be constructed to study the phylogenetic relationship (Figure 13). The concatenated sequences within each cluster showed great similarity to each other, as can be identified by the short branch lengths.

While the distance-based clustering method provided a tree based on the gene composition similarity, the cluster refinement phylogenetic tree used SNPs to investigate the evolutionary relationship within each cluster, which were similar in topology with the clustering method.

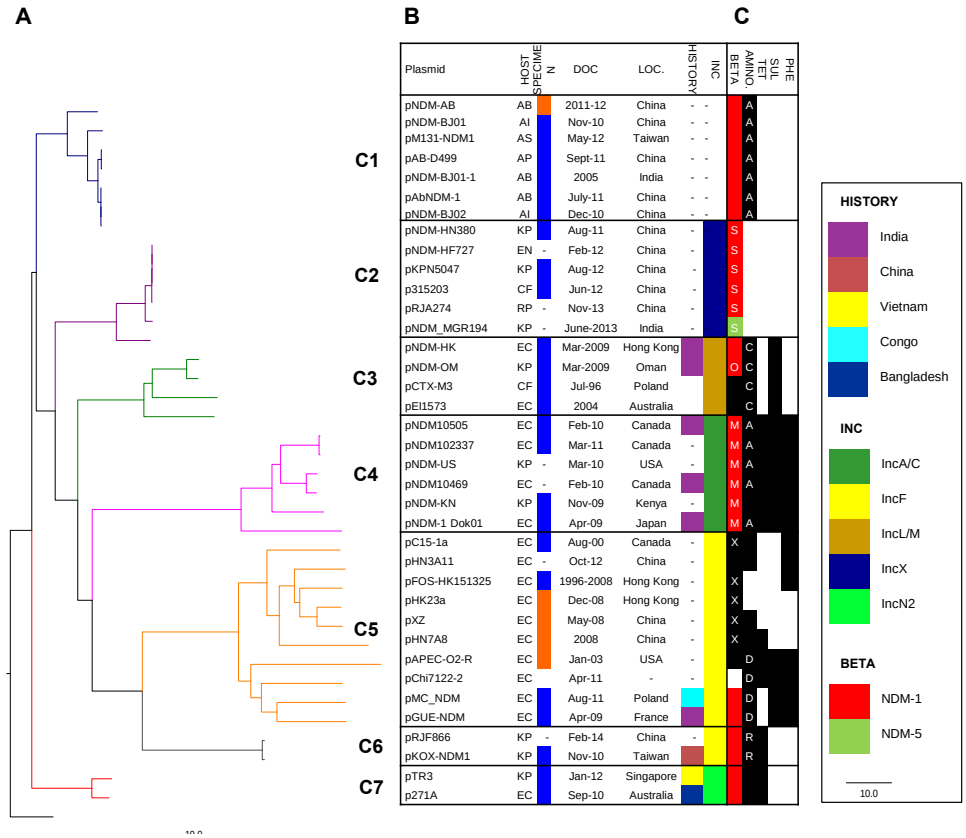


Figure 12. Clustering of *bla*$_{NDM}$-positive plasmids. (A) Neighbor-Joining tree of plasmids in the 7 *bla*$_{NDM}$ clusters. Branches of each cluster are colored distinctively with blue (C1), purple (C2), green (C3), magenta (C4), orange (C5), grey (C6), and red (C7). The tree is rooted using an outgroup in black. Branch lengths were Euclidean distances calculated from similarity scores and are reflective of the similarity of plasmid gene composition and the similarity of shared genes. (B) Table showing the identity (PLASMID), bacterial host (HOST), specimen type (SPECIMEN), date of collection (DOC), geographical sampling location (LOC), travel history (HISTORY) and incompatibility group (INC) for each plasmid. Abbreviations: AB, *Acinetobacter baumannii*; AI, *Acinetobacter iwoffii*; AP, *Acinetobacter pittii*; AS, *Acinetobacter soli*; CF, *Citrobacter freundii*; EN, *Enterobacter cloacae*; EC, *Escherichia coli*; KP, *Klebsiella pneumoniae*; and RP, *Roultella planticola*. (C) The matrix displays the resistance genetic determinants identified in the corresponding plasmid genome. A black-shaded box indicates a positive genotypic trait conferring resistances, the antibiotic classes of which are indicated by the text at the top of the column. Resistance determinants against the following antibiotics were identified: beta-lactam, BETA; aminoglycoside, AMINO; tetracycline, TET; sulphonamide, SUL; and phenicol, PHE. Abbreviations: A, APH; C, AAC; D, AAD; K, KPC; M, CMY; O, OXA; S, SHV; R, RMT; and X, CTX. Presence of *bla*$_{NDM-1}$ was shaded red and *bla*$_{NDM-5}$ shaded green.

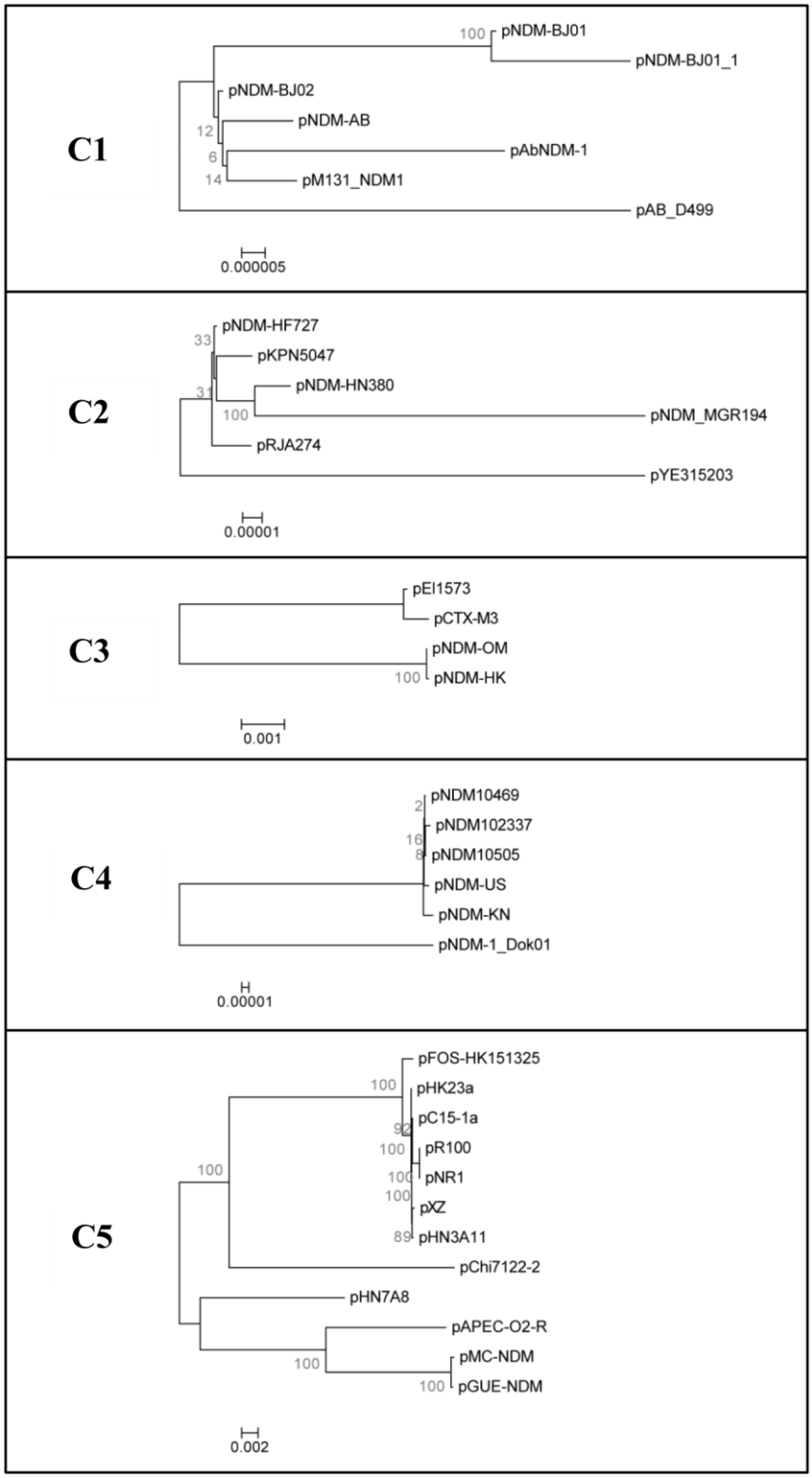


Figure 13. SNP-based refinement maximum likelihood trees of *bla*$_{NDM}$ plasmid clusters. For each cluster, sequences of all plasmids within this cluster were extracted, whose shared regions were aligned and concatenated for the construction of the maximum likelihood trees shown above. The results for C6 and C7 were not shown as the clusters only consist of 2 isolates each. The branch lengths were calculated by RAxML and reflect the number of expected mutations per site. Bootstrap values are in a scale of 0 to 100, and are shown at each node in grey.

4.3.6 Global *bla*$_{NDM}$-positive plasmid diversity: gene transposition

At least 6 events in the 7 clusters (C1 to C7) of *bla*$_{NDM}$-positive plasmids have been observed to indicate independent recombination events introducing *bla*$_{NDM}$ into different plasmid backbones of *bla*$_{NDM}$-negative plasmids (Figure 14).

In the process of adaptive evolution, diversity of microbial genomes is primarily driven by recombination or point mutation [111, 112]. As the clustering approach makes use of plasmid gene composition diversity arising through recombination rather than point mutations, these findings suggested the *bla*$_{NDM}$-positive plasmids have undergone extensive mobile genetic element transposition to adapt to varying environmental niches. As mentioned earlier, there was minimal intra-cluster SNP difference, suggesting that polymorphisms due to point mutation play minimal role to account for the diversity of the plasmids.

Transpositions facilitated by transposons (Tn), insertion sequences (IS) elements and IS common region (ISCR) are detected frequently in plasmids that involve antimicrobial genes, non-antimicrobial genes and transposable genetic elements. With respect to the *bla*$_{NDM}$ gene, transposition mechanism involving *bla*$_{NDM}$ was discernible by comparative genomics in 4 instances: pNDM_HN380 [GenBank: JX104760.1] (C2, IS*Aba125*-mediated transposition, Figure 14A), pNDM-OM [GenBank: JX988621.1] (C3, recombination into Tn1548-borne class I integrin, Figure 14B), pEcNDM [GenBank: NC_023909.1] (unclustered, IS*CR1*-mediated transposition, Figure 14C), and pNDM-BTR [GenBank: KF534788.1] (unclustered, *fip*A gene hotspot recombination, Figure 14D). The Tn125 composite transposon

platform has been theorized to be the original vehicle to mobilize *bla*$_{NDM}$ among *Acinetobacter* species. The results reveal that *bla*$_{NDM}$ introductions also occurred in the context of IS*CR1*-mediated transposition, *fip*A gene hotspot recombination and Tn1548-borne class I integron recombination. Larger datasets of genomic sequences involving *bla*$_{NDM}$-positive plasmids and nearest neighbors will enhance the understanding of *bla*$_{NDM}$ transposition globally.

4.3.7 Global *bla*$_{NDM}$-positive plasmid diversity: incompatibility group and geographical distribution

The plasmid clustering based on gene composition diversity tends to cluster the plasmids with the same backbone together, thus showing a clear clustering of the plasmid Inc groups for *Enterobacteriaceae* plasmids: plasmids in C2 are all Inc X plasmids, plasmids in C3 are Inc L/M, plasmids in C4 are Inc A/C, plasmids in C5 and C6 are Inc F, while plasmids in C7 are Inc NII (Figure 12).

The plasmid clusters also showed some association with geographical distributions. Some clusters were spreading mainly via regional transmission to date: (1) C1, a cluster of plasmids *Acinetobacter* sp. host, is limited to South Asia and East Asia; (2) C2 and C6 are limited to South and East Asia; and (3) C7 was found in Southeast Asia and Oceania. Other clusters (C3, C4, and C5) had wider geographic dispersion involving South Asia, East Asia, Middle East, North America, Africa and Europe.

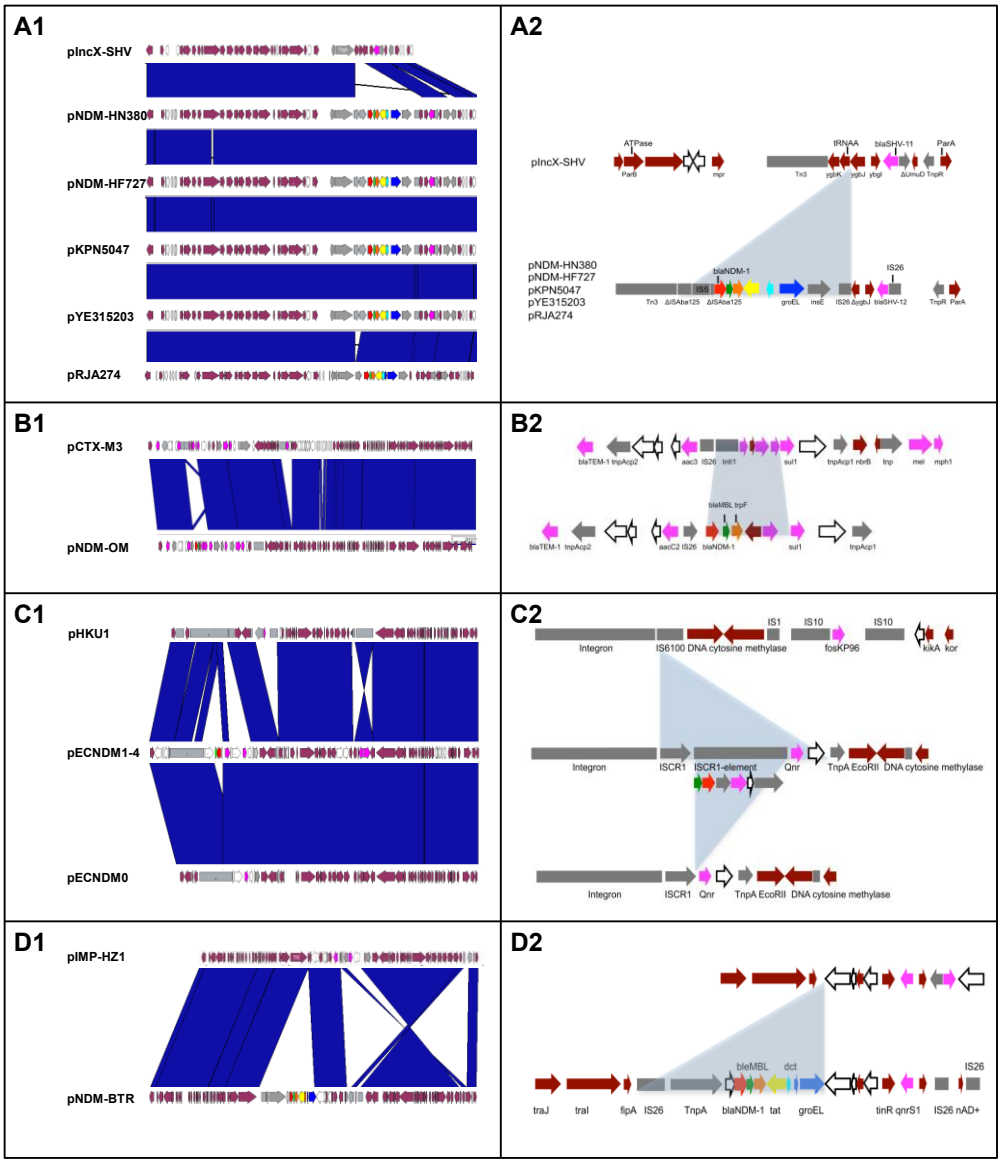


Figure 14. Acquisition of *bla*$_{NDM}$ cassettes. A1, B1, C1, and D1: A comparison of the *bla*$_{NDM}$-positive plasmid genomes with their putative backbone plasmids as identified in the plasmid clustering. The corresponding backbone plasmids are placed at the top of each column. Blue bands between panels indicate nucleotide BLAST matches with more than 99% sequence similarity. A2, B2, C2, and D2: Schematic representations of insertions in the *bla*$_{NDM}$-positive plasmids (shaded in light blue) corresponding to A1, B1, C1, and D1. Annotated genes in these regions are colour coded. Arrows indicate predicted open-read frames, genes with known functions (maroon), antimicrobial resistance genes (magenta), transpositional genetic elements (grey) and hypothetical proteins (white). Genes from the *bla*$_{NDM}$ cassette are indicated by arrows coloured as follows: red, *bla*$_{NDM}$; green, *ble*$_{MBL}$; orange, *trpF*; yellow, *tat*; light blue, *dct*; and dark blue, the *groES- groEL* cluster. Plasmid pECNDM0 represents an *bla*$_{NDM}$-negative laboratory-derived plasmid, where the *bla*$_{NDM}$ cassette was mobilized from pECNDM1-4 as a free form.

4.3.8 Local *bla*$_{NDM}$-positive plasmid in the global context

As detailed in the global analysis, pTR3 clustered tightly with p271A [GenBank: JF785549.1], a plasmid described in Australia (Figure 12, C7). The other two plasmids were located in different plasmid clusters: pNDM-KN in C4 and pNDM_MGR194 in C2. In contrast to global plasmid diversity, the presence of near identical pTR3 plasmids in 8 out of 11 local samples suggested the *bla*$_{NDM}$-positive plasmid diversity at the local level to be very low. On the other hand, the 2 non-pTR3 plasmids, which were related to different plasmid clusters in the global plasmid phylogeny, were detected each in only one patient, which suggested independent plasmid introductions into the hospital ecology.

4.4 Discussion

By analyzing whole genome sequences of 11 *bla*$_{NDM}$-positive CPE isolated in a local hospital and 2,749 complete plasmid sequences (including 39 *bla*$_{NDM}$-positive plasmids) in the NCBI database, I investigated the local transmission and global dissemination of the *bla*$_{NDM}$ gene. This analysis has highlighted the complex genetic pathways of *bla*$_{NDM}$ spread. Globally, *bla*$_{NDM}$ spread involved marked plasmid diversity with no predominant bacterial clone. The *bla*$_{NDM}$-positive plasmids were carried by multiple species of *Acinetobacter* and *Enterobacteriaceae*, thereby highlighting the propensity for conjugation of *bla*$_{NDM}$-positive plasmids among different bacterial species. The *bla*$_{NDM}$ gene module mobilized between different plasmid backbones on at least 6 independent occasions. In contrast to the global plasmid diversity, early local spread of *bla*$_{NDM}$-positive plasmids in a single Singapore hospital was

characterized by clonal spread of a predominant plasmid pTR3 with 2 sporadic instances of plasmid introduction (pNDM-KN and pNDM_MGR194).

The plasmid clustering approach is crucial to the current analysis as it allows quantitative analyses of plasmid molecular epidemiology involving a large number of diverse plasmids as a tool in analyzing global spread of plasmid-borne genes. Prior genomic investigations of *bla*$_{NDM}$ spread have been mainly restricted to comparisons of less than 10 closely related plasmids due to the lack of phylogenetic congruence, and hence have not been able to discern the patterns of *bla*$_{NDM}$-positive plasmid clustering at a global level. Establishment of nearest-neighbor relationships facilitated the determination of transposition events involving genomic regions (genes and insertion sequences). Determination of cluster relationships subsequently opened the ability to correlate clusters with specific properties (for example, extent of global spread or plasmid Inc groups).

Whole genome studies of successful bacterial clones have been used to understand transmission of chromosomally-mediated antimicrobial resistant bacteria, MRSA for example. However, prior studies relying upon bacterial chromosomes to understand *bla*$_{NDM}$ transmission have been hindered by the diversity of bacterial species and strains harboring *bla*$_{NDM}$, even in a single geographic locale [113]. The current study highlighted three vital evolutionary mechanisms underlying *bla*$_{NDM}$-positive bacteria diversity: (1) *bla*$_{NDM}$-gene module transposition, (2) *bla*$_{NDM}$-positive plasmid conjugation, and (3) *bla*$_{NDM}$-positive bacteria spread. Future studies of *bla*$_{NDM}$ transmission would have to take into account these three levels of gene spread.

Gene module transposition was a vital factor in the successful spread of *bla*$_{NDM}$ for at least three reasons: (1) mobilization of *bla*$_{NDM}$ from *Acinetobacter sp.* plasmids to *Enterobacteriaceae* plasmids as has been recognized before; (2) mobilization of *bla*$_{NDM}$ among *Enterobacteriaceae* plasmids of differing Inc groups; and (3) non-*bla*$_{NDM}$ gene movement facilitating adaptation of plasmids to differing selection pressures.

Local *bla*$_{NDM}$ spread in a single Singapore hospital context was characterized predominantly by conjugation of a clonal plasmid (pTR3) between *Enterobacteriaceae* (inter-bacteria plasmid conjugation), and inter-human host *bla*$_{NDM}$-positive bacteria transmission (bacteria spread). The finding of the pTR3 plasmid in 2 distinct *K. pneumoniae* strains in another Singapore hospital further supported a significant role of inter-human host transmission and clonal plasmid conjugation in local spread. Three recent publications using whole genome sequencing also reported the predominant role of inter-human host transmission (via the inanimate environment in some cases) and HGT in local hospital spread of carbapenemases [113–115].

One potential reason for the difference in the local and the global plasmid diversity is the sampling and the time period. While the 39 global complete *bla*$_{NDM}$-positive plasmid sequences has a long time range of eight years, the 11 local isolates were isolated within a one-year period.

The current analysis offers a glimpse of the genetic armamentarium available to *bla*$_{NDM}$ for dissemination to multiple environments. The limited data available for understanding transmission of this important resistance gene is highlighted by availability of only approximately 39 *bla*$_{NDM}$-positive and 2,749 Gram-negative whole plasmid sequences globally. Whole genome

sequencing of *bla*$_{NDM}$-positive isolates from diverse geographies on a much larger scale will increase the understanding of *bla*$_{NDM}$ evolution and spread, and may prove crucial to long-term control of *bla*$_{NDM}$.

Since this study was conducted (April 2014), more bacterial isolates have been sequenced and more plasmid sequences have been archived in the NCBI database. Till the time of this thesis (March 2016), the number of *bla*$_{NDM}$-positive complete plasmid sequences has increased to 98. Though the number is still limited, including more plasmid sequences in the analysis could potentially provide more insights into the transmission pattern of the *bla*$_{NDM}$ gene and the control of its spread.

Also, a 41,190 bp plasmid pNDM-ECS01 [GenBank:KJ413946.1] in ST131 *E. coli* was later reported in Thailand as a *bla*$_{NDM}$-positive plasmid highly similar to pTR3, differing only by three nucleotide insertions [116]. However, the isolate was reported to be sequenced by Illumina MiSeq, the mere use of which can hardly generate complete plasmid sequences. Thus no inferences about the spread of pTR3-like plasmids were made based on this plasmid.

Assembly error is a common problem for *de novo* assembly, which may result in relocations, translocations, inversions and local errors of misjoins [117]. Assemblies of Velvet has also been reported to contain these errors [117]. Thus, in order to avoid false inference on the global structure, downstream analysis using assembled contigs mainly made use of local sequences, whether by means of using BLAST for local sequence alignment, MuMMer [118] to get local hits, or reference-based mapping and calling to determine variants.

4.5 Conclusions

The analysis has revealed the complex genetic pathways of bla_{NDM} spread, where the global dissemination is mainly characterized by transposition of the bla_{NDM} gene cassette into different plasmids while early local transmission is mainly a result of plasmid conjugation and bacteria spread. These findings advance understanding of plasmid-mediated antimicrobial resistance spread both locally and globally.

Chapter 5

Gene evolution by duplication: innovation, amplification, innovation and divergence

5.1 Background

Gene duplication is regarded as a major force for genome evolution [119] and is prevalent in genomes of all three domains of life [81]. While the generation of gene duplication can be attributed to unequal crossing over, retroposition, or chromosomal duplication in Eukaryotes [81], in bacteria, however, two important forces are causing gene duplication. One is HGT that copies a gene into another genome. The other is homologous recombination between identical sequences that can cause gene duplication by generating tandem repeats.

Originating from an individual, a duplicated gene would either get removed for the extra burden and functional redundancy it costs to the genome or it get fixed in the population. The fate of duplicated genes raises the Ohno's dilemma [120], which states that the duplicated gene should be allowed sufficient time to accumulate mutations for new functionalities to arise, and that selection as a most probable force for the maintenance of the new copy would actually limit the loss of old functions and the generation of new functions.

Attempts have been made to account for the mechanism for the maintenance of gene duplicates in the genome, and can be summarized into the following models: (1) Neofunctionalization. This model states that one of the copies is maintained by purifying selection, thus retaining the original function, while the other can evolve freely to acquire mutations for new gene functions [119, 121, 122] . One of the predictions of this model is that since purification selection exerts different pressure on the copies, they have different mutation rates. Once the accumulated mutations lead to new

functions, they are enhanced by positive selection [123]. (2) Subfunctionalization, also known as the complementation-degeneration model. This model proposes that each of the copies adopts different aspects of the original functions of the gene, which predicts symmetry in evolutionary rates between the two copies due to the same mechanism of mutation accumulation [122–125]. One form of subfunctionalization is differential expression, which can either be different expression in different organs [122] or different expression in adaptation to environmental changes [126]. (3) Increased-dosage advantage. In this model, the mere increase in the amount of gene product is an advantage, fixing the duplicates rapidly and maintained thus. However, this is more often than not a reversible process that once the selection pressure relieves, the augmented gene would be removed for its obvious fitness cost [127]. (4) IAD model. According to this model, a side functional trait arises by innovation before dene duplication, after which environmental changes value the new trait and select for its increase in level via amplification. The increase in copy number enables more beneficial mutations and compensates for the potential negative effects of a new mutation. Then selection further favors the mutations, thus facilitating their divergence [120, 128].

Microevolution is referred to as the changes in one or a few loci within a clonal population [129], which is regarded as a major evolution method for clonal populations. It has been used to explain biological phenomenon such as the immune escape during clonal spread of *Neisseria meningitides* and host specificity in *Campylobacter jejuni* [130]. Bacterial populations are shaped strongly by microevolution, and thus are stably polymorphic in certain sites.

After long time culturing, the genome is polymorphic for duplications, thus enabling the rapid adaptation and divergence under selection pressures [131].

Porins are bacterial pores located on the outer membrane of Gram-negative bacteria. Maltoporins, also known as the *LamB* porins because they are coded by the *LamB* gene, are a family of outer membrane proteins that specifically transport maltose and maltodextrins. Maltoporin is also a lambda phage receptor. Active maltoporin is a trimer [132]. Each monomer contains an independent channel, but all three monomers of a trimer are needed for phage adsorption. While the phage receptor site is exposed on the surface, the sugar binding site potentially resides within the channel [132]. Porins, as channels for molecules to diffuse, are always produced in large amounts.

5.2 Methods

5.2.1 Haplotype reconstruction with QuasQ

QuasQ is a software for reconstructing haplotypes from fragmented next-generation sequencing reads, which is written in Perl and is freely available at: http://www.statgen.nus.edu.sg/~software/quasq.html. This software is published on BMC Bioinformatics with the title "Viral quasispecies inference from 454 pyrosequencing" [133], where a detailed description of the algorithm and evaluation of the performance can be found.

Initially designed for 454 sequencers, QuasQ is capable of handling sequencing reads having an average length of several hundred base pairs and a quality score for each sequenced base, which cat be translated to the probability that the base call is correct. QuasQ consists of four parts: (1) mapping the reliable sequencing reads to a reference sequence after pre-

processing and quality filtering; (2) local error correction; (3) haplotype reconstruction and collapsing; and (4) frequency estimation.

5.2.1.1 Pre-processing

Low quality reads with sequencing errors would affect haplotype reconstruction by inflating the estimated number of haplotypes and affect the population size estimation, and thus should be eliminated. Two kinds of 454 reads are supposed to harbor more errors than others: reads with at least one 'N' call and reads of extreme lengths [134]. In the pre-processing step, reads having at least one 'N' call or reads of extreme lengths (defined as reads with lengths beyond the 1% extremes on either side of the read length distribution) would be removed.

5.2.1.2 Mapping

Reads that passed the quality filtering would be aligned against a user-specified reference sequence with Bowtie2 [30]. Reads uniquely aligned with alignment length and identity both above 80% are retained for downstream processing. The homopolymer problem, which is a misrepresentation of the number of bases when faced with a stretch of identical bases, is well addressed by Bowtie2.

5.2.1.3 Local error correction

An issue with haplotype reconstruction is that point mutations in a sequencing read can either be real variants harbored by a haplotype or a sequencing error. To reduce the possible inflation of the haplotype number caused by

sequencing error, local error correction is conducted in a sliding-window manner. Within each window, all allele combinations whose frequencies are below 0.5% are corrected to be the combination with the shortest hamming distance. By doing this, sequencing errors are being corrected at the cost of sacrificing the haplotypes whose frequencies are below 0.5%.

5.2.1.4 Haplotype reconstruction

The method for haplotype reconstruction is shown below in Figure 15. Polymorphic sites refer to sites with more than one allele supported by sequencing reads. QuasQ first identifies the polymorphic sites (Figure 15A), which are used for haplotype reconstruction. After reducing sequencing reads to only polymorphic sites (Figure 15B), the reads are grouped into sets based on the starting position (Figure 15C). Within each set, reads that are subsets of other reads in the same set are filtered out (Figure 15C). A read graph method is used with each graph node to be the combination of alleles at the polymorphic sites within each corrected read, and each directed edge connecting two nodes if the postfix of the first node is a prefix of the second node (Figure 15D). To rid the possibility of the overlapping polymorphic sites being *in vitro* artifacts, at least one sequencing read that spans the polymorphic site as well as the immediate neighboring polymorphic sites is needed to support the join (Figure 15E). A gap is left when such supporting reads cannot be found. Parts before and after a gap would be assembled separately and joined in all possible ways.

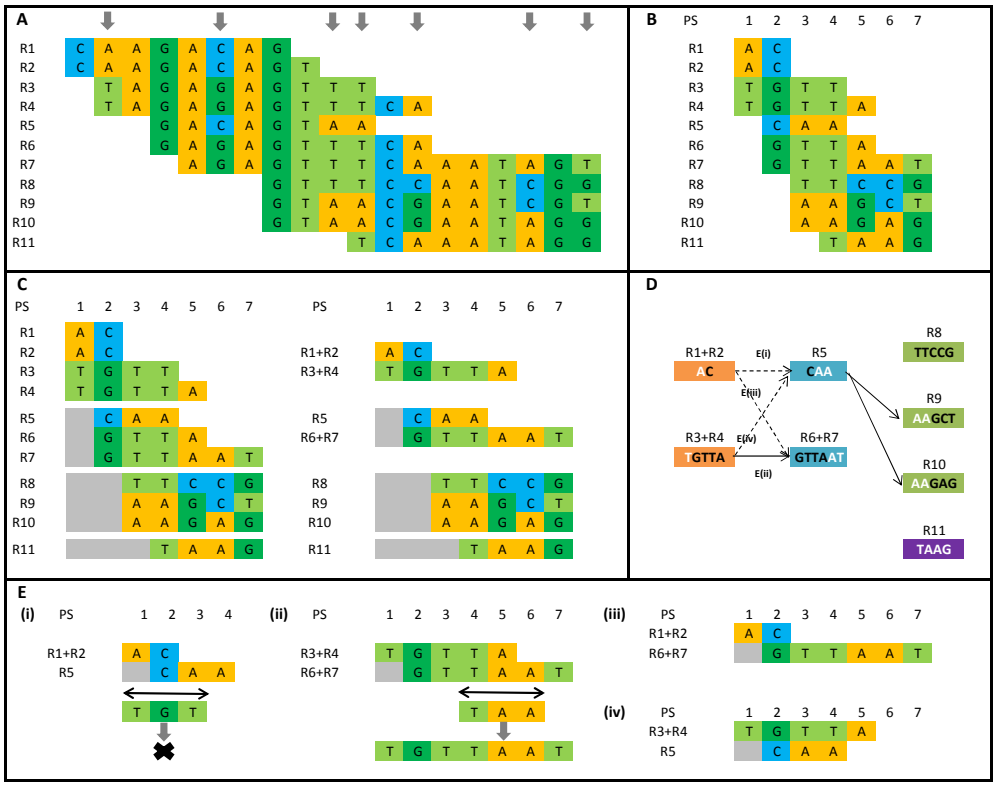


Figure 15. A schematic representation of the QuasQ haplotype reconstruction workflow. Sequencing reads that have passed the quality filtering at the pre-processing step with sequencing errors corrected locally are (A) piled up to identify all polymorphic sites (PS), defined as sites with two or more alleles. (B) Non-polymorphic sites are removed from the sequencing reads, with only PS for downstream reconstruction. (C) The processed reads are then grouped into sets based on their starting positions. Within each set, reads that are subsets of other reads in the same set are removed (R3 removed as a subset of R4). (D) Haplotypes are constructed with a read-graph method. Nodes are read sequences, with directed edges connecting two nodes if a postfix of the first node is a prefix of the second one. Solid arrows such as E(ii) represent possible directed edges, while non-probable edges such as E(iii) and E(iv), shown with dotted arrows, are due to non-identical node sequences overlap. In E(i), nodes with identical overlap like R1 + R2 and R5 are considered probable only if there are sequencing reads spanning the nodes and the immediate neighboring PS. For pairs of reads with identical overlap E(i), if the allele combination in the defined region is not the same as that of any other read in that region, the two nodes are not joined. In this figure, the only constructed haplotype is thus R3 + R4 + R6 + R7. (This figure is modified from Figure 7 in the original article [133].)

5.2.1.5 Sequence collapsing

The constructed haplotypes with an identity over 90% are collapse to a single sequence as a representative of the highly similar sequences.

5.2.1.6 Frequency estimation

Frequency for the constructed haplotypes is estimated with the freqEst program [135] implemented within the ShoRAH [136] package, which is based on an EM algorithm.

5.2.2 Identification of *LamB* gene sequences

LamB gene sequence (corresponding protein ID: AFQ63346.1) was extracted from *K. pneumoniae* 1084 genome [GenBank:CP003785.1] and was used as a query sequence to search for similar sequences in the NCBI 'nt' database using nucleotide BLAST. A similarity score is calculated for each of the hit as: length of matching sequence * BLAST identity / length of the reference sequence. An similarity score cut-off is set at 0.45 [71] to define the gene as 'present' in the genome.

5.2.3 Construction of Neighbor-Joining SNP tree

Genetic sequences were aligned with ClustalW [137], after which the evolutionary history was inferred with MEGA6 [138] using the Neighbor-Joining method [139] with a bootstrap test of 1000 replicates. The distances are calculated as the number of differences with all ambiguous positions removed for each sequence pair and are in the unit of number of base differences per sequence.

5.2.4 Haplotype reconstruction and minimum spanning tree construction

QuasQ v1.2 was used for haplotype reconstruction using *LamB* gene sequence extracted from *K. pneumoniae* 1084 genome [GenBank:CP003785.1] as the

reference sequence at similarity level of 0.95 and the rest of the parameters were set to default. The resulting base counts for each position were used to calculate major allele frequency. The reconstructed haplotypes were used for minimum spanning tree construction and phylogenetic study. Minimum spanning trees were constructed with the function 'spantree' implemented in the R package 'vegan'.

5.2.5 Variant calling for heterogeneity from sequencing reads

Sequencing reads were aligned using Novoalign with default parameters, taking *K. pneumoniae* 1084 genome [GenBank:CP003785.1] as the reference genome. After indel realignment with GATK IndelRealigner [140] and duplicate removal with Picard Tools 1.100, heterogeneous variants were called with LoFreq [141] with default parameters. Variant sites were extracted with the respective allele frequencies. Shannon entropy was calculated as:

$$H = -\sum_i p_i \log(p_i) \quad \text{for i in A, T, C, and G}$$

5.2.6 Core genome tree of chromosomes of *K. pneumoniae* and related species

Annotated coding sequences of *K. pneumoniae* 1084 [GenBank:CP003785.1] were taken from NCBI. Sequences containing any of the following features: (1) phage sequences; (2) CRISPR region; and (3) tandem repeats were removed, resulting in a total of 4,919 coding sequences as candidate sequences. Those candidate sequences present in all the chromosomes were taken as the core genes for those chromosomes, which contains 2,945 gene sequences. After extracting these gene sequences in each chromosome and aligning properly,

they were concatenated into core genomes for building Neighbor-Joining SNP tree.

5.2.7 Protein structure prediction

Protein structures were predicted with I-TASSER server v4.2 [142–144] with default parameters.

5.3 Results

5.3.1 IAID model for gene evolution by duplication

In the IAID (Innovation-Amplification-Innovation-Divergence) model for gene evolution by duplication can be divided into four stages (Figure 16). Firstly, the gene is present in the form of a cloud of similar sequences in the population, generated by microevolution. Mutations can be beneficial, neutral or deleterious and some are preserved in the population with secondary activities. This stage, characterized by microevolution, is called innovation. Secondly, amplification takes place. In Eukaryotes, this can be achieved by unequal crossing over, retroposition, or chromosomal duplication. In bacteria, this can be attributed to tandem duplication or HGT. Thirdly, after the amplification, both of the amplified genes are still existent as sequence clouds in the population, experiencing the same innovation process as in the first stage. The evolution rates may differ given different selection forces. Lastly, advantaged sequences for each copy would then prevail under selection pressure, facilitating the divergence of the gene copies.

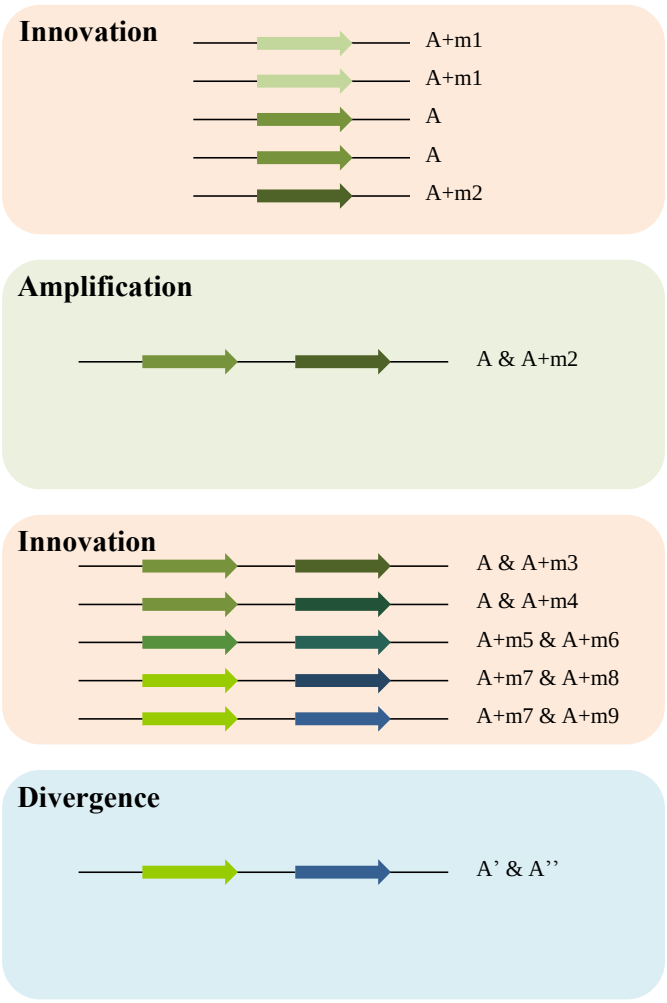


Figure 16. A schematic representation of the IAID model of gene evolution by duplication. First, the gene (A) is present in the population as a cloud of similar sequences, some of which have minor functional changes (m1, m2) generated by microevolution. This is a step called innovation. Then, there is an amplification of the gene. In bacteria, for example, the amplification can be generated by tandem duplication or imported via horizontal gene transfer. After the amplification, both of the amplified gene copies are still existent as sequence clouds in the population, produced by microevolution while selected by similar or different pressures. Advantaged sequences would then prevail under selection pressure, facilitating the divergence of the copies into A' and A'' with functional improvements or new functions.

The IAID model is a derivative of the IAD model. It differs from the IAD model in two aspects: (1) In the IAID model, point mutation is an important source of mutation for the divergence of the genes both before and after duplication. Considering the population instead of focusing on individuals,

point mutations, even at low mutation rate, can accumulate to a big pool in the population. (2) For bacteria specifically, HGT is regarded as a means by which amplification can take, as a stage of the gene evolution, rather than an independent way of gene revolution by duplication. Thus, HGT is within the range of the IAID model.

5.3.2 *LamB* gene is duplicated in *K. pneumoniae* and other related species

One copy of the *LamB* gene sequence was taken from *K.pneumonia* 1084 genome [GenBank:CP003785.1], with its translated protein ID being AFQ63346.1, and was queried in the NCBI 'nt' database for genomes harboring similar sequences. Altogether 83 hits were identified with a similarity score above 0.45. Interestingly, in all the bacterial chromosomes picked up by BLAST as having similar genes, all the *K. pneumoniae*, *Klebsiella variicola* (*K. variicloa*), *Enterobacter aerogenes* (*E. aerogenes*), *Klebsiella oxytoca* (*K. oxytoca*) *and Raoultella ornithinolytica* (*R. ornithinolytica*) chromosomes have 2 hits as summarized in Table 8. This illustrates that this copy of the *LamB* gene is widely duplicated in *K. pneumoniae* and other related strains (core genome SNP tree of the chromosomes is presented in Figure 17).

A Neighbor-Joining SNP tree was constructed to uncover the phylogenetic structure of these duplicate genes, in which six distinct clusters were defined (Figure 18). Cluster1 and Cluster2 contain sequences only from *K. pneumoniae*. Cluster3 has one *K. variicola* strain and two *K. pneumoniae* strains isolated from plants. Cluster4, Cluster5 and Cluster6 correspond respectively to *E. aerogenes, K. oxytoca and R. ornithinolytica*. The gene

sequences cluster primarily based on their species, probably as a reflection of their diverse environmental niches, life style, as well as selection pressures. Within each species, the two copies from the same chromosome fall into different branches, leading to a bifurcating topology within each species branch. This clearly shows that in all the chromosomes, there are two copies of *LamB* that are similar yet stably maintaining their differences.

Table 8. Summary of complete bacterial genomes harboring two copies of *LamB* gene and the plasmid harboring *LamB* gene.

Accession	Name	Length
CP006923.1	*K. pneumoniae* 30660/NJST258_1	5,263,229
CP006918.1	*K. pneumoniae* 30684/NJST258_2	5,293,301
CP000964.1	*K. pneumoniae* 342	5,641,239
CP006659.1	*K. pneumoniae* ATCC BAA-2146	5,435,369
CP006648.1	*K. pneumoniae* CG43	5,166,857
CP006656.1	*K. pneumoniae* JM45	5,273,813
CP002910.1	*K. pneumoniae* KCTC 2242	5,259,571
FO834906.1	*K. pneumoniae* str. Kp52.145	5,438,894
CP009114.1	*K. pneumoniae* strain blaNDM-1	5,297,511
CP008929.1	*K. pneumoniae* strain PMK1	5,317,001
CP003785.1	*K. pneumoniae* subsp. pneumoniae 1084	5,386,705
CP003200.1	*K. pneumoniae* subsp. pneumoniae HS11286	5,333,942
CP003999.1	*K. pneumoniae* subsp. pneumoniae Kp13	5,307,003
CP008700.1	*K. pneumoniae* subsp. pneumoniae KP5-1	5,365,144
CP008827.1	*K. pneumoniae* subsp. pneumoniae KPNIH1	5,394,056
CP007727.1	*K. pneumoniae* subsp. pneumoniae KPNIH10	5,395,263
CP008797.1	*K. pneumoniae* subsp. pneumoniae KPNIH24	5,396,164
CP007731.1	*K. pneumoniae* subsp. pneumoniae KPNIH27	5,241,638
CP008831.1	*K. pneumoniae* subsp. pneumoniae KPR0928	5,309,305
CP000647.1	*K. pneumoniae* subsp. pneumoniae MGH 78578	5,315,120
AP006725.1	*K. pneumoniae* subsp. pneumoniae NTUH-K2044	5,248,520
CP006798.1	*K. pneumoniae* subsp. pneumoniae PittNDM01	5,348,284
CP009208.1	*K. pneumoniae* ATCC 43816 KPPR1	5,374,834
FO203501.1	*K. pneumoniae* subsp. rhinoscleromatis strain SB3432	5,270,770
CP001891.1	*K. variicola* At-22	5,458,505
CP004142.1	*R. ornithinolytica* B6	5,398,151
FO203355.1	*E. aerogenes* EA1509E	5,419,609
CP002824.1	*E. aerogenes* KCTC 2190	5,280,350
CP003683.1	*K. oxytoca* E718	6,097,032

CP004887.1	*K. oxytoca* HKOPL1	5,914,407
CP003218.1	*K. oxytoca* KCTC 1686	5,974,109
CP008788.1	*K. oxytoca* KONIH1	6,152,190
CP008841.1	*K .oxytoca* strain M1	5,865,090
CP007734.1	*K. pneumoniae* KPNIH27 plasmid pKPN-262*	338,850

* This plasmid has only one copy of the *LamB* gene.

Accession	Copy 1 start	Copy 1 end	Copy 2 start	Copy 2 end	Distance (bp) **
CP006923.1	50,239	51,618	4,680,991	4,682,373	631,094
CP006918.1	50,239	51,618	4,694,482	4,695,864	647,675
CP000964.1	37,297	38,676	4,966,471	4,967,853	710,682
CP006659.1	5,346,148	5,347,527	627,268	628,650	715,109
CP006648.1	4,317,374	4,318,753	5,020,657	5,022,039	701,905
CP006656.1	86,356	87,735	4,668,822	4,670,204	689,964
CP002910.1	4,911,094	4,912,473	376,166	377,548	723,263
FO834906.1	50,383	51,762	5,024,511	5,025,893	463,383
CP009114.1	3,422,696	3,424,075	4,140,502	4,141,884	716,428
CP008929.1	1,670,727	1,672,106	2,345,217	2,346,599	673,112
CP003785.1	50,159	51,538	4,693,016	4,694,398	742,465
CP003200.1	5,248,017	5,249,396	605,527	606,909	690,072
CP003999.1	50,015	51,394	4,653,313	4,654,695	702,322
CP008700.1	2,157,244	2,158,623	2,820,988	2,822,370	662,366
CP008827.1	5,308,012	5,309,391	608,652	610,034	693,316
CP007727.1	5,309,219	5,310,598	608,652	610,034	693,316
CP008797.1	5,310,120	5,311,499	608,652	610,034	693,316
CP007731.1	5,154,246	5,155,625	594,963	596,345	680,975
CP008831.1	5,223,261	5,224,640	608,653	610,035	693,317
CP000647.1	4,445,232	4,446,611	5,176,653	5,178,035	730,043
AP006725.1	5,162,537	5,163,916	659,151	660,533	743,754
CP006798.1	3,647,566	3,648,945	2,944,639	2,946,021	701,544
CP009208.1	3,500,129	3,501,508	2,771,545	2,772,927	727,201
FO203501.1	49,674	51,053	822,558	823,940	771,504
CP001891.1	40,331	41,710	4,829,455	4,830,837	667,998
CP004142.1	4,199,789	4,201,174	3,510,203	3,511,591	688,199
FO203355.1	985,271	986,639	4,001,763	4,003,152	2,401,727
CP002824.1	1,411,129	1,412,508	4,450,454	4,451,832	2,239,646
CP003683.1	5,983,739	5,985,116	2,810,230	2,811,610	2,922,145
CP004887.1	755,685	757,062	3,933,001	3,934,379	2,735,712
CP003218.1	1,321,363	1,322,741	4,085,470	4,086,850	2,762,730
CP008788.1	6,037,418	6,038,795	2,731,493	2,732,873	2,844,887
CP008841.1	996,088	997,465	3,677,470	3,678,850	2,680,006
CP007734.1	187,727	189,099	NA	NA	NA

** The distances are between the CP1 3'-end and the CP2 5'-end except for FO203501.1, FO203355.1 and CP002824.1 whose gene copies are in different directions, where the distance is the shorter distance between the copies considering that the genome is circular.

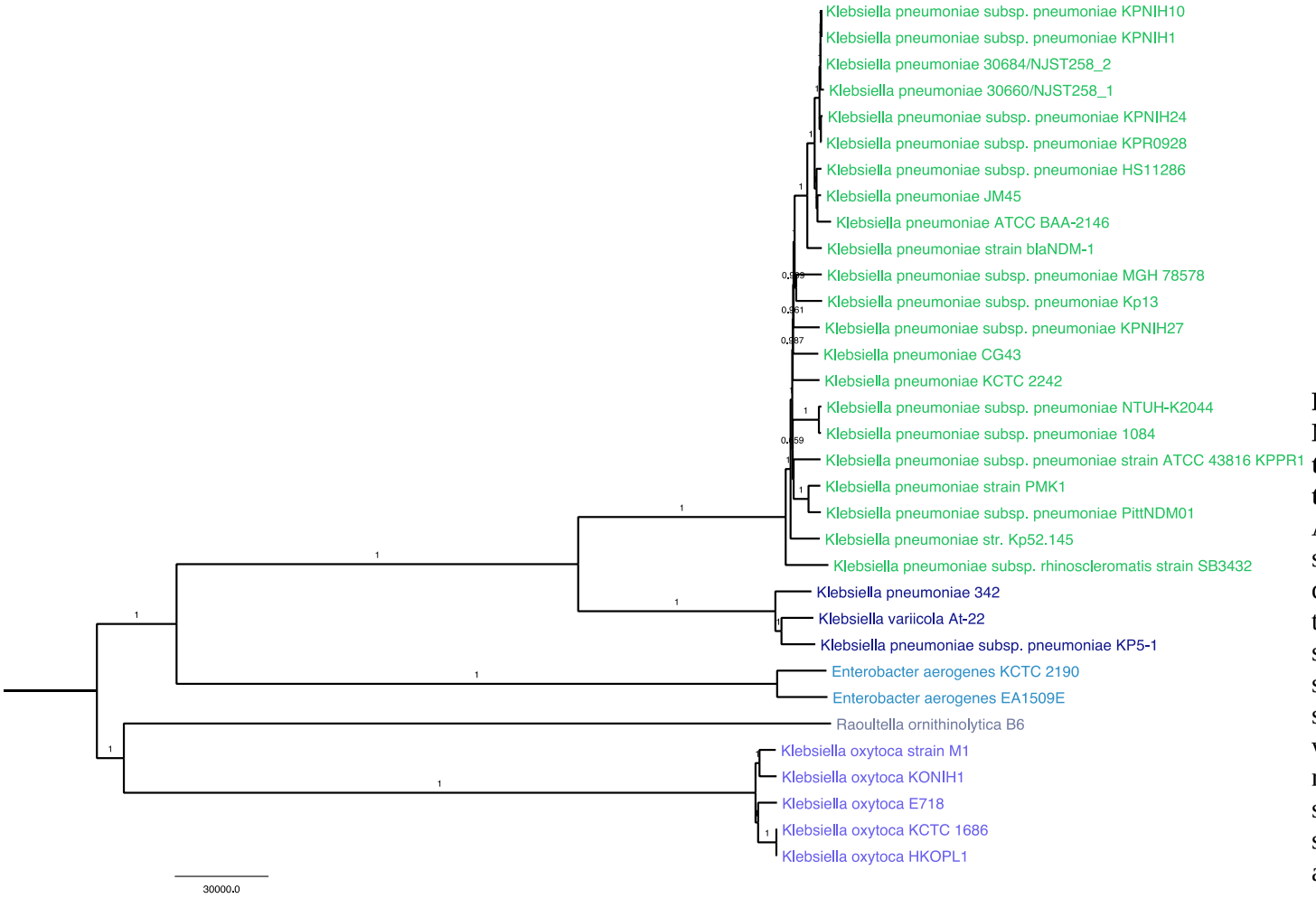


Figure 17. Core genome Neighbor-Joining SNP tree of the chromosomes harboring two copies of the *LamB* gene. According to the tree, different species show distinctly different core genomes except that *K. variicola* shares a great similarity with *K. pneumoniae* strains isolated from plant sources. The branch lengths were calculated by MEGA and reflect the number of differing sites. Bootstrap values are in a scale of 0 to 1, and are shown at each node.

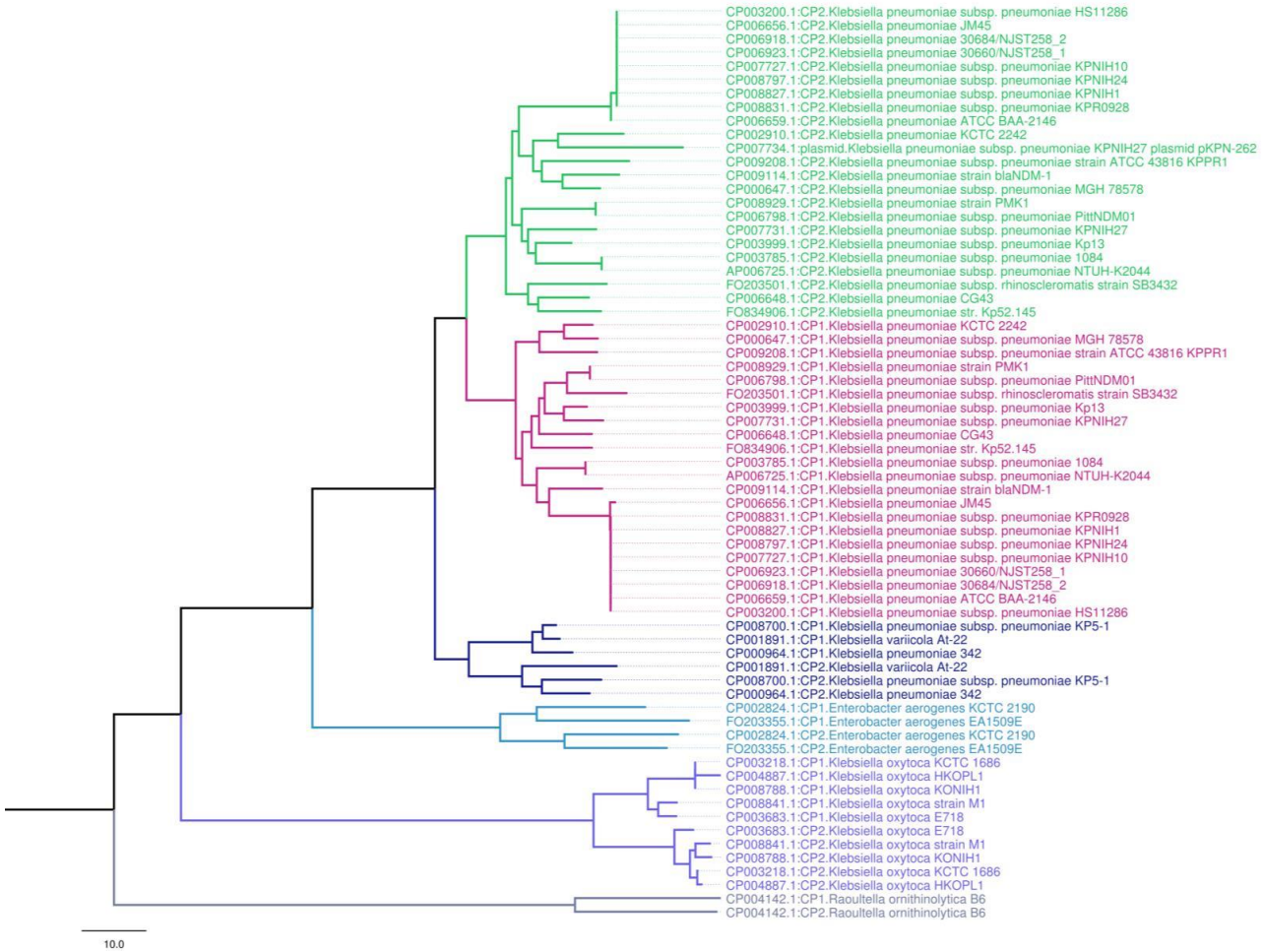


Figure 18. Neighbor-Joining SNP tree of *LamB* gene sequences summarized in Table 8. According to the tree, *LamB* gene sequences can be divided into six distinct clusters. The sequences in Cluster1 and Cluster2 are all from *K. pneumoniae* isolated form humans or other animal sources. Interestingly, each copy of a pair in the same chromosome falls into different clusters, either into Cluster1 or into Cluster2. Cluster3 contains one *K. variicola* strain and two *K. pneumoniae* strains isolated from plants. Cluster4, Cluster5 and Cluster6 correspond respectively to *E. aerogenes*, *K. oxytoca* and *R. ornithinolytica.* The branch lengths were calculated by MEGA and reflect the number of differing sites.

5.3.3 Amplification of *LamB* gene by tandem duplication

A walkthrough of the genetic distances between the gene copies in all the chromosomes show that there are differences across species while the within-species difference is much smaller (Figure 19A). While *K. pneumoniae, R. ornithinolytica* and *K. variicola* show similar between-copy distances, *E. aerogenes* and *K. oxytoca* chromosomes have much larger distances.

In Figure 19A, the large diamond on the left represents the *K. pneumoniae* str. Kp52.145 chromosome [GenBank:FO834906.1], which was isolated before 1935 in Indonesia, Java from a human host [145]. Compared to the more recent *K. pneumoniae* isolates, it has a much shorter distance. With the genome size stable, this increase in the between-copy distance is an implication that the initial gene was amplified by tandem duplication and the distance increases as there are introductions of new genes and genomic islands.

An inspection of the surrounding regions of the gene duplicates compared to the *K. pneumoniae* 1084 genome (Figure 20A) shows that the surrounding regions share a great sequence similarity across all species (the plasmid excluded) with the implication that the duplications may be traced to the same amplification event and passed to the rest of the genomes. Apart from the sequence similarity in the gene surrounding regions, the region between the copies were also examined for similarity. The region between *LamB* gene copies for the *K. pneumoniae* 1084 genome is similar to that of *K. variicola* At-22 (Figure 20C) and *R. ornithinolytica* B6 (Figure 20D), and is similar to *K. pneumoniae* Kp52.145 with more insertions in beween (Figure 20B). Similarly, *K. oxytoca* genomes share sequence content with *K. pneumoniae* 1084, but with major insertions taken place (Figure 19B). Apart from the similarities, *E.*

aerogenes chromosomes have different sequence content from the rest of the genomes. Given the similarity of the regions adjacent to the genes, it is supposed that *E. aerogenes* got introduced the *LamB* gene pairs at an early stage without the region between the gene pairs stably established.

Amplification via HGT, an important driving force for gene duplication in bacteria, still preserves its possibility here since a plasmid [GenBank:CP007734.1] was identified as harboring *LamB* gene, though the transfer of the gene copy by this plasmid is not firmly corroborated with the experimental data.

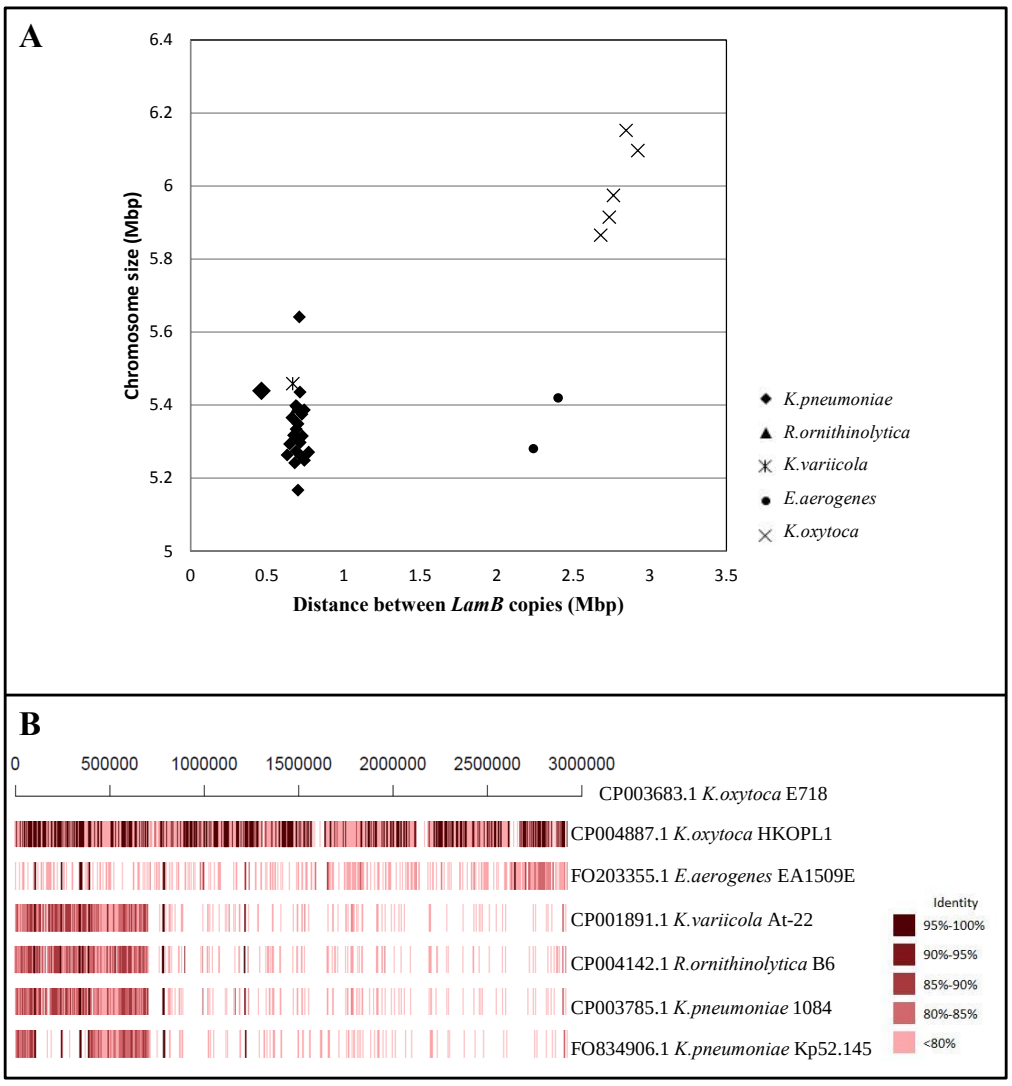


Figure 19. Characterization of the regions between *LamB* gene copies within chromosomes: distance (A) and sequence similarity (B). In A, for each chromosome, the distance between the *LamB* copies was plotted against the chromosome size and was labeled according to its species. While *K. pneumoniae*, *R. ornithinolytica* and *K. variicola* share similar between-copy distance except for the historical sample *K. pneumoniae* Kp52.145, *E. aerogenes* and *K. oxytoca* samples have different between-copy distances from other species. In B, the sequence similarities were examined for each chromosome compared to the between-pair sequence of *K. oxytoca* E718. *K. oxytoca* HKOPL1, shown here as a representative of other *K. oxytoca*, has a great sequence similarity with the reference. *K. variicola* At-22, *R. ornithinolytica* B6 *and K. pneumoniae* 1084 (a representative of the majority of *K. pneumoniae* samples) have shorter distances similar to a part of the reference. *K. pneumoniae* Kp52.145, with an even shorter distance, is also similar to the part of the reference. This reflects multiple insertions of gene elements between the gene pairs. Apart from the similarities, *E. aerogenes*, represented by *E. aerogenes* EA1509E, has quite different sequence content from all the other species.

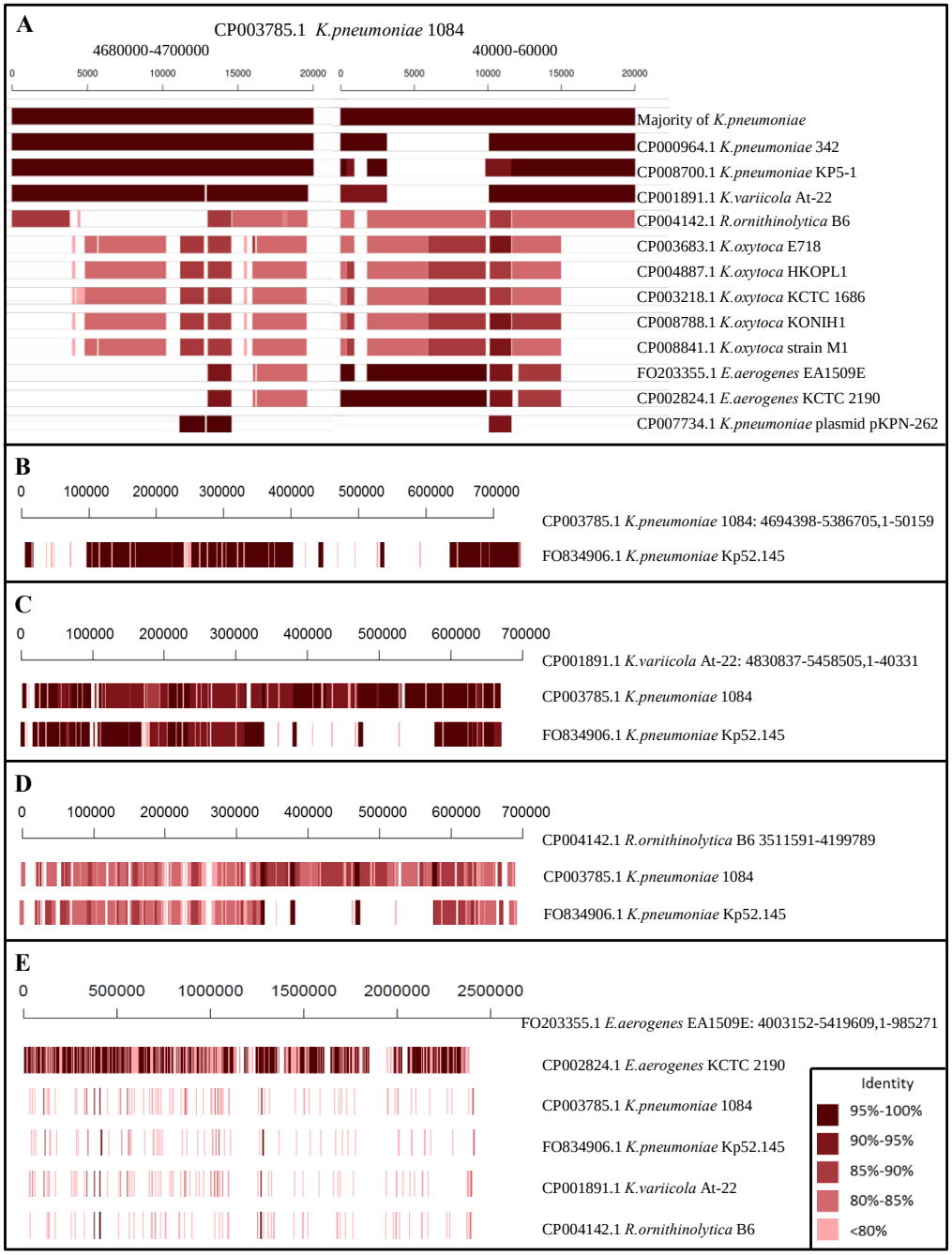


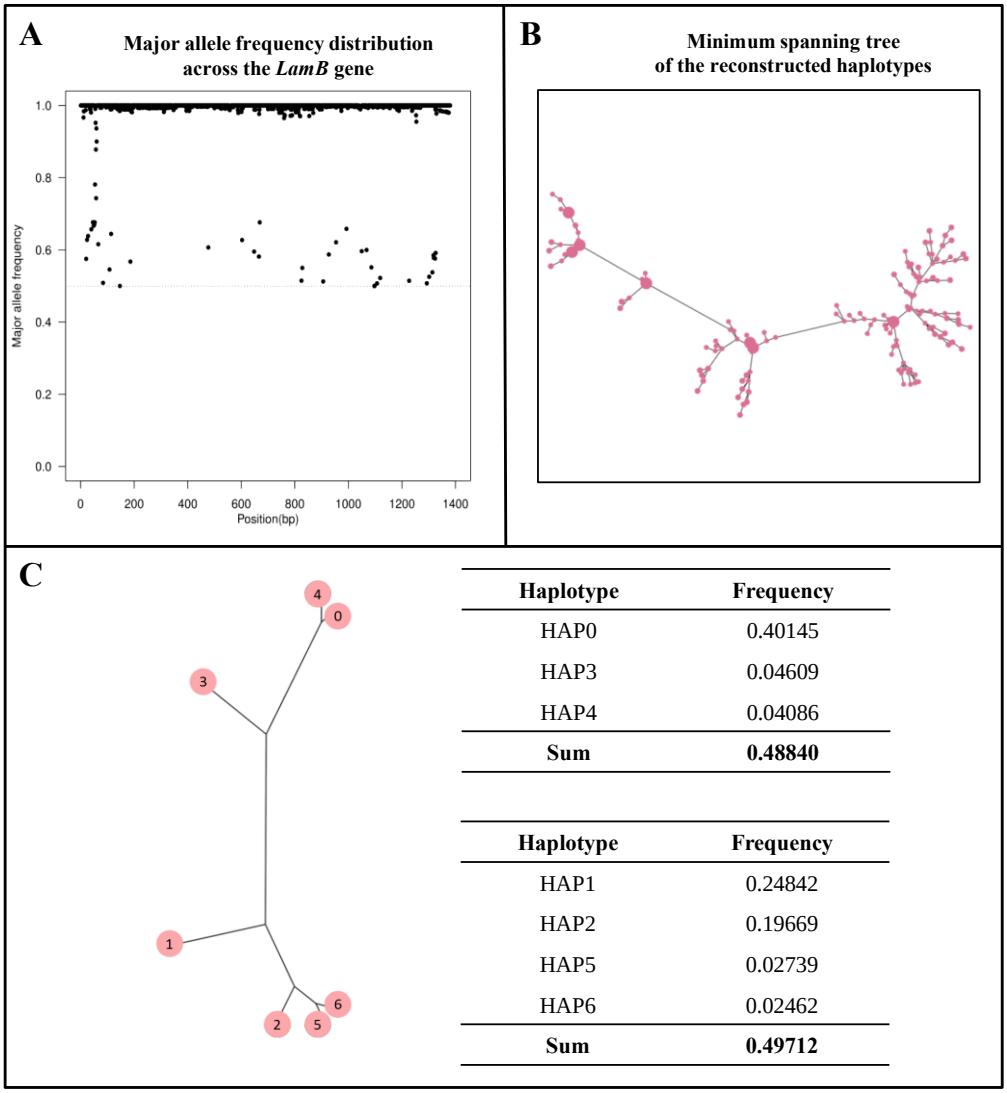
Figure 20. Similarity of gene surrounding regions (A) and between-gene regions (B, C, D, E). *LamB* gene sequences on the *K. pneumoniae* 1084 genome were extracted with their surrounding regions and searched for similarities in other chromosomes (A). Similarities were shared in all the chromosomes but not the *LamB*-bearing plasmid, suggesting the possibility that the duplications originate from a single amplification event and passed to other chromosomes. Examination of the similarity of the between-gene sequences shows that recent *K. pneumoniae* chromosomes differ from historical sample by a number of insertions (B), that *K. variicola* and *R. ornithinolytica* share a great similarity with recent *K. pneumoniae* chromosomes (C, D), and that *E. aerogenes* has few in common with the rest of the chromosomes (E).

5.3.4 *LamB* gene innovation via microevolution

Colonies of cultured clinical isolates were combined for whole genome sequencing with Ion Proton Sequencer and independent experiments with different isolates were conducted with Illumina HiSeq Sequencer for the purpose of verification, giving sequencing statistics summarized in Table 9. Shannon entropy distribution across the complete genome of called variant sites with LoFreq [141] using *K. pneumoniae* 1084 as the reference genome demonstrated a great degree of polymorphism, as a result either of repeat regions or real polymorphism shaped by microevolution. To uncover the microevolution of the *LamB* gene, haplotypes were reconstructed with QuasQ, using the gene sequence coding the protein AFQ63346.1 as the reference for isolates sequenced with Ion Proton. The isolates sequenced with Illumina were not included in haplotype reconstruction since read length is not long enough. Summary of the read depth proves it reasonable for haplotype reconstruction. Major allele frequencies were calculated based on the reconstruction results, showing multiple polymorphic sites along the gene sequences (Figure 21A as a representative). Haplotypes for each isolate were taken to build minimum spanning tree (Figure 21B as a representative). According to the minimum spanning tree in Figure 21B, *LamB* gene sequence evolves like a cloud of sequences similar to each other. A Neighbor-Joining SNP tree constructed with high-frequency haplotypes (haplotypes with a frequency larger than 1%) (Figure 21C as a representative) splits into two distinct clusters, each of which may represent one copy of the gene, with the frequency summed up to next to 50%. It can be seen from the tree that both of the gene copies are evolving by forming a cloud of closely related sequences, which is a result of

microevolution. Other isolates show similar figures as in Figure 21 and are thus not included here for brevity.

In bacterial population, *LamB* gene copies differ from one another by some point mutations, due to which the innovation and generation of new side functions of the gene is possible. This is a constant process, providing a large gene pool from which to acquire new functions or on which a bacterial population can bank to survive new selection pressures. Unlike in the IAD model where constant duplication of genes is regarded as the major force for innovation, it is posed here that point mutation is a driving force for gene innovation before or after gene duplication.



Haplotype	Frequency
HAP0	0.40145
HAP3	0.04609
HAP4	0.04086
Sum	**0.48840**

Haplotype	Frequency
HAP1	0.24842
HAP2	0.19669
HAP5	0.02739
HAP6	0.02462
Sum	**0.49712**

Figure 21. *LamB* gene evolves like a cloud of similar sequences. Colonies of a cultured clinical *K. pneumoniae* isolate were collected, combined and sequenced with Ion Proton Sequencer. Sequencing reads were reconstructed with QuasQ to derive the haplotypes within the *LamB* gene region. Based on the QuasQ output, the major allele frequencies along the complete gene sequence were plotted (A), proving the existence of polymorphic sites within the gene with differing minor allele frequencies. A minimum spanning tree (B) was built to illustrate that the *LamB* gene copies, as a pool, evolve like a sequence cloud. In B, larger dots are haplotypes with a frequency larger than 1%. Those haplotypes were extracted to build a Neighbor-Joining SNP tree (C). The tree splits into to two parts, both of which have their frequencies added up to around 50%. The two parts are supposed to be the two copies of the *LamB* gene, each with their neighbors similar yet with point mutations generated by microevolution.

Table 9. *K. pneumoniae* whole genome sequencing statistics and MLST.

Name	# Reads	# Bases	Estimated coverage*	MLST
iso_1	5,902,846	863,934,298	172.79	ST-231
iso_2	6,236,046	907,137,765	181.43	ST-231
iso_3	5,959,694	869,656,713	173.93	ST-231
iso_4	5,570,416	816,925,942	163.39	ST-231
iso_5	5,784,574	850,574,891	170.11	ST-231
iso_6	6,042,300	881,048,927	176.21	ST-231
iso_7	6,301,109	913,693,564	182.74	ST-231
iso_8	5,566,986	814,096,617	162.82	ST-231
iso_9	5,128,619	742,056,540	148.41	ST-231
iso_10	6,374,614	935,483,828	187.1	ST-231
iso_11	5,546,520	763,132,383	152.63	ST-231
iso_12	4,669,240	646,232,907	129.25	ST-231
iso_13	5,210,950	768,238,739	153.65	ST-11
iso_14	5,222,167	764,019,341	152.8	ST-273
iso_15	7,046,688	971,676,087	194.34	ST-14
iso_16	6,227,353	865,715,146	173.14	ST-16
iso_17	5,308,787	735,601,857	147.12	Unknown
illumina_1	1,313,970	394,191,000	78.84	ST-231
illumina_2	1,372,709	411,812,700	82.36	ST-231
illumina_3	1,240,225	372,067,500	74.41	ST-231
illumina_4	1,191,909	357,572,700	71.51	ST-231
illumina_5	1,502,732	450,819,600	90.16	ST-231

* The coverage is estimated by # bases/5,000,000

5.3.5 Divergence after gene duplication

For each chromosome, the number of amino acid changes from the historical sample, *K. pneumoniae* str. Kp52.145, was counted for each copy. When taking all chromosomes into consideration, the number of changes for the two copies regressed to the line $y = 0.9927x+2.3242$ with a R^2 of 0.9786 (Figure 22A). With the slope next to 1, this result suggests that when passing from species to species, the two copies evolve at a similar pace. When, however, looking into only the *K. pneumoniae* chromosomes, the two copies exhibit different patterns and are badly correlated (Figure 22B), denying the possibility that they are under the same selection pressure. An examination of the pair-wise amino acid difference within each cluster showed a significantly

(p-value = 1.215e-09) different mean values of difference for the two clusters, suggesting that Cluster1 copies, though forming a distinct cluster, have a larger variation that Cluster2 copies. This, again, suggests that the two copies are under different selection pressure in *K. pneumoniae* isolates. The same experiments were done with nucleotide differences and showed similar results, supporting the divergence driven by different selection pressures. This uneven evolution rate was also reported in rodent genes that there is an increased divergence in the novel daughter copies after duplication, which can be attributed to positive selection [146].

Amino acid sequences of all LamB gene copies in Figure 18 were aligned and compared with a number of amino acid differences observed. Positions with at least five sequences having differing residuals from the majority were plotted in Figure 23. While some of these differences feature a specific species, some residuals, like those of Positions 2, 3, 4, 10, 14, 17, 18, 20, and 21 at the N-terminus, are different in the two copies, which means that the two copies within the same chromosome differ at these residuals from one another. It is noteworthy that at position 21 of the aligned sequences, one copy has a deletion compared to the other which has a threonine.

Structures for the two LamB copies on *K. pneumoniae* 1084 were predicted with I-TASSER server. Predicted secondary structures both have 18 strands, which is true for *LamB* as a specific porin (Figure 24A). For the initial 60 amino acid residuals, however, there is a difference in the predicted helices. The predicted solvent accessibility (Figure 24B), at the same time, shows various regions of difference across the gene region.

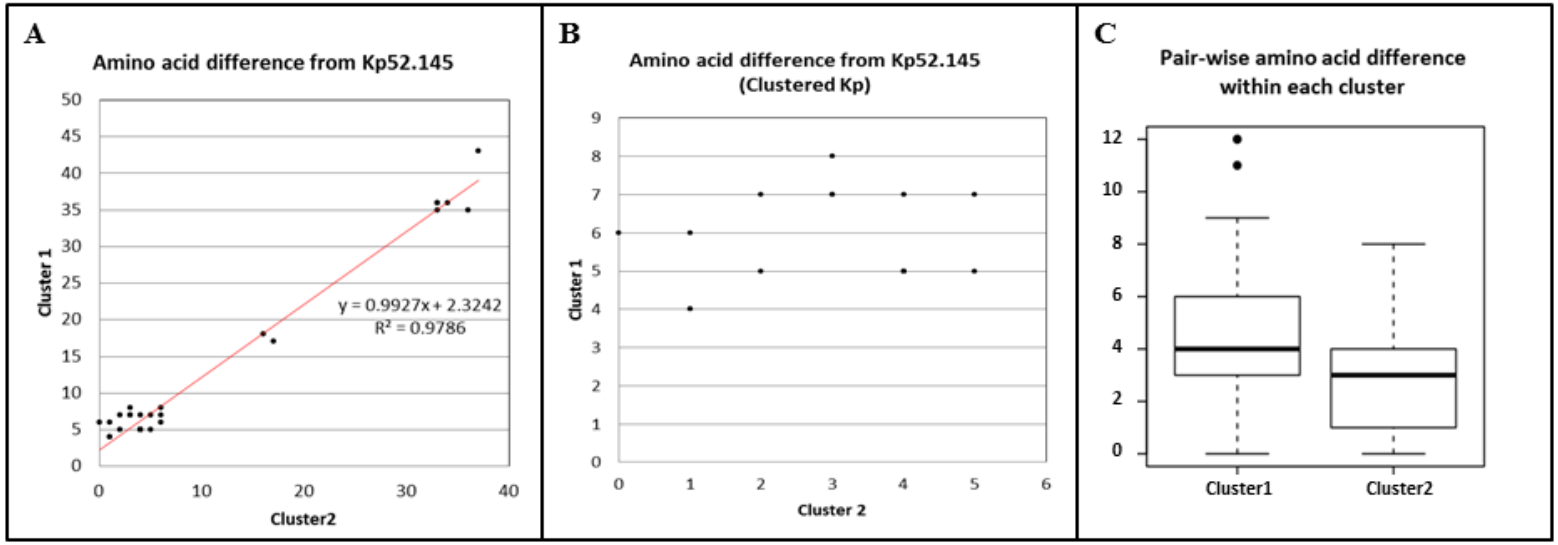


Figure 22. Amino acid changes of *LamB* gene sequences in each cluster. Using the *LamB* gene copies in the historical sample *K. pneumoniae* Kp52.145 as the references, the number of amino acid changes was calculated for each gene copy in each chromosome. In A and B, each dot represents one chromosome, with the x-axis value its difference from the Cluster2 *LamB* of the *K. pneumoniae* Kp52.145 and the y-axis value its difference from the Cluster1 copy. When taking all species into consideration (A), the dots fit well to the linear trend line with a slope of 0.9927, manifesting that the evolution rates are similar for the two gene copies while passing from species to species. When only the clustered *K. pneumoniae* chromosomes were considered (B), however, the two copies show different degrees of variation that the copies in Cluster2 varies in a range of 0-5 while the copies in Cluster1 varies by 4-8 amino acids. The intra-cluster pair-wise amino acid differences were also calculated for both clusters and were plotted in C. Cluster1, compared to Cluster2, has a higher pair-wise difference, which was tested significant with a t-test p-value of 1.215e-09.

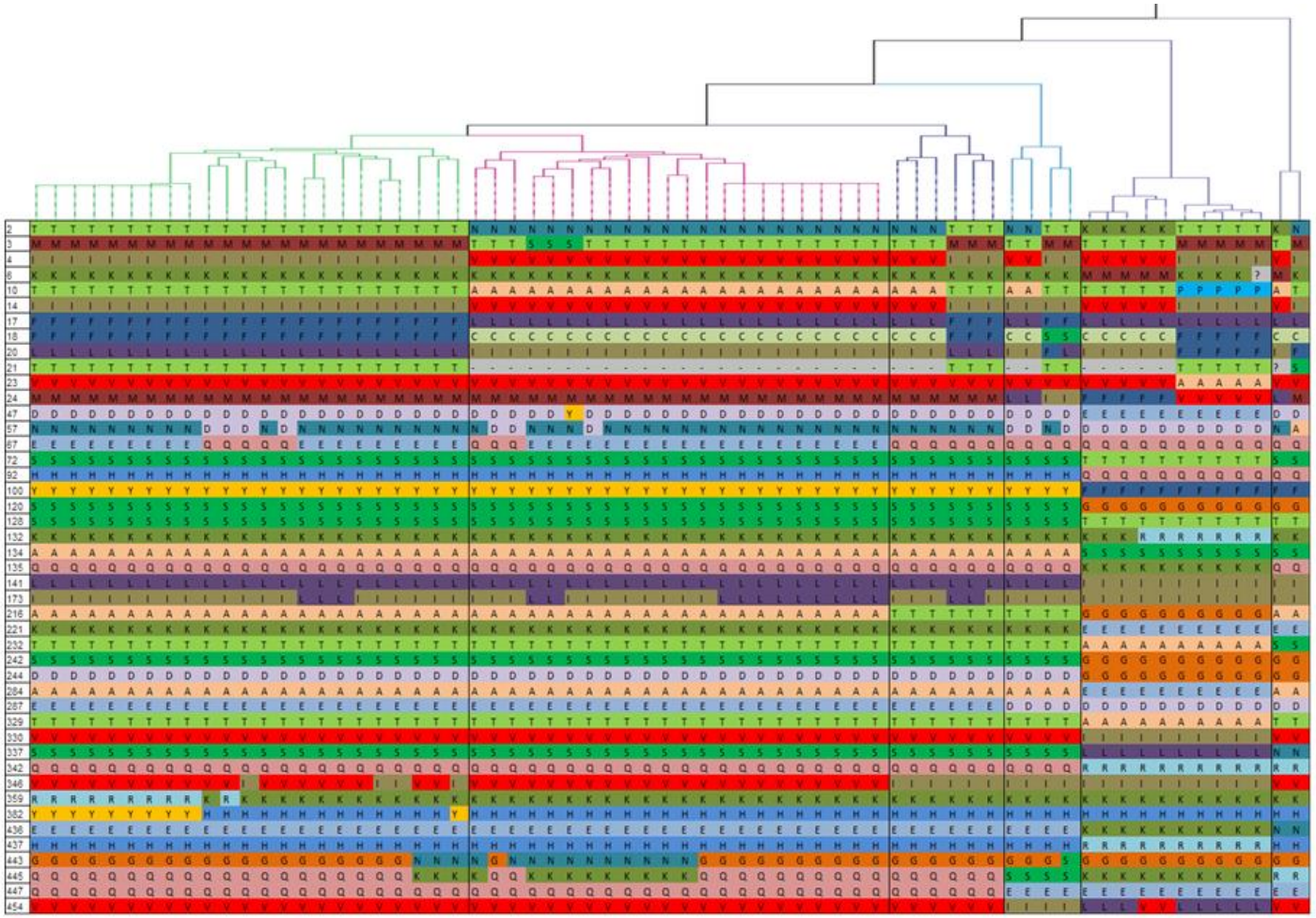


Figure 23. Positions with at least five sequences having different residuals from the major residual. A number of amino acid differences were observed for all *LamB* sequences summarized in Table 8. While some of these are species-specific, there are some residuals (Positions 2, 3, 4, 10, 14, 17, 18, 20, and 21, for example) at the N-terminus that are copy-specific, which suggests the two copies within the same chromosome differ at these residuals.

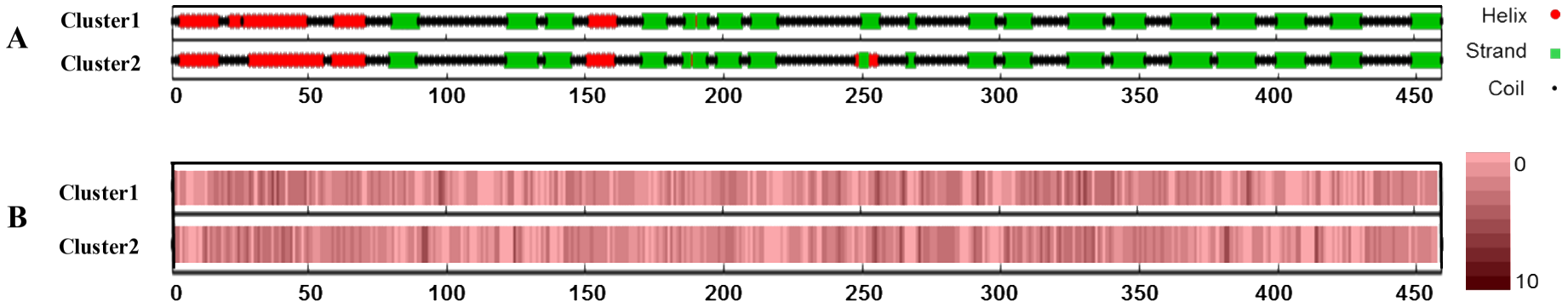


Figure 24. Predicted secondary structure (A) and solvent accessibility (B) for the two *LamB* copies in *K. pneumoniae* 1084 genome. I-TASSER server, a platform for protein structure and function predictions was used. The predicted secondary structures for the Cluster1 copy and the Cluster2 copy have similar structures in that they both show the typical 18 strands of maltoporin, but have different helix compositions for the first 60 residuals. Prediction of solvent accessibility gave values ranging from 0 (buried residue) to 10 (highly exposed residue). The prediction results for the two copies were plotted in B.

Different selection pressures as the two gene copies are potentially under, they are not diverging unboundedly from each other according to the samples available (Figure 25). There is a range of around 19-51 nucleotide differences between the gene pairs resulting in only 8-15 amino acid changes, especially for *K. pneumoniae,* which has only 8-12 amino acid differences within pairs. This, in its implications, states the potentially overlapping functions in certain regions of the gene given the difference in selection pressure the pairs are under.

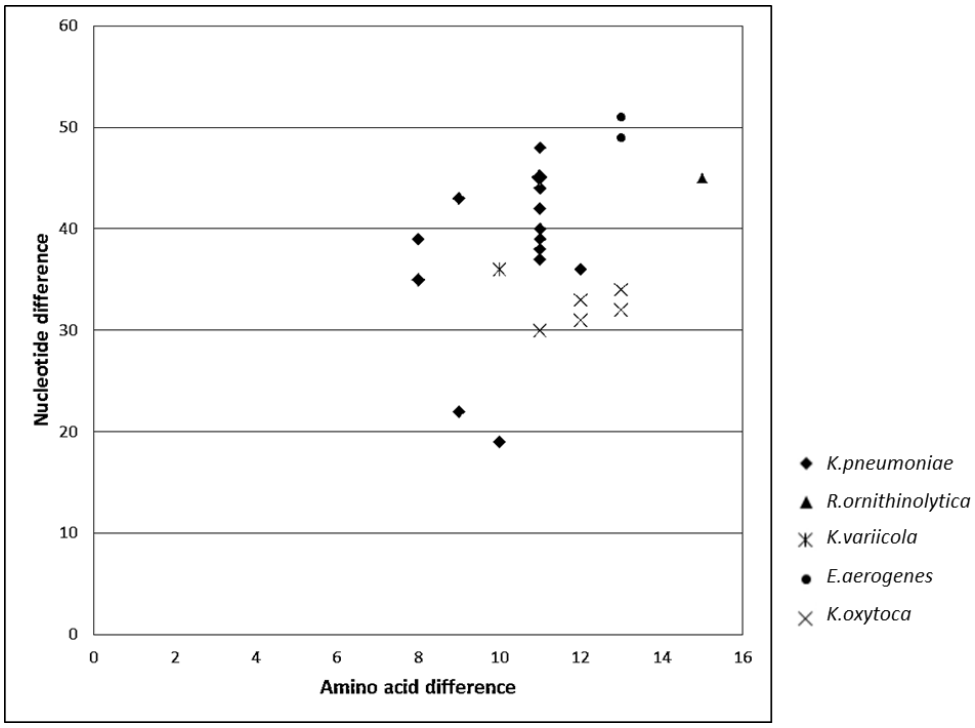


Figure 25. Difference between gene pairs within the same chromosome. Each chromosome is represented by a dot with the amino acid difference between the two *LamB* copies shown in the x-axis, the nucleotide difference in the y-axis, and the species denoted by the shape. Regardless of the number of nucleotide differences, the number of amino acid change is bounded, especially for *K. pneumoniae* samples, with a range of 7 to 12. This reflects that while the copies are evolving with their own selection pressures, the pressure may not be independent and it maintains a bounded level of amino acid difference. Another explanation may be that part of the proteins serves the same functionality that is too conservative to allow for amino acid changes.

5.4 Discussion

In summary, *LamB* gene duplication in *K. pneumoniae* and other related species was investigated using 34 complete genomes available in NCBI, together with whole genome sequencing data of 22 cultured clinical isolates. *LamB* gene duplication is found in *K. pneumoniae, K. oxytoca, K. variicola, E. aerogenes,* and *R. ornithinolytica* and is maintained in *K. pneumoniae* as two distinct copies lying from each other at a narrow range of genetic distances. During bacteria growth as a population, *LamB* gene copies are stably polymorphic for single-nucleotide variations, evolving like a cloud of similar sequences, providing the gene pool with more mutations for emergence of new functions. Under selection pressure, genes with survival advantages are preserved. When selection pressures are different for the two copies, they evolve at different rates. In this case, the two copies are evolving at different rates, while the potential overlap in functions limits their unbounded divergence from one another.

Based on this example, the IAID (Innovation-Amplification-Innovation-Divergence) model for genome evolution via gene duplication is proposed as comprised of the following four steps: (1) the gene in the population is undergoing constant microevolution to introduce mutations for innovation; (2) the gene is amplified; (3) innovation continues to take place after duplication; and (4) selection pressure drives the divergence of the gene copies.

While the fate of the majority of the duplicated genes is to be removed due to fitness cost, some are preserved in the genome, stably or temporarily. Various models have been proposed to explain the maintenance of duplicated genes in the genome. In the increased-dosage advantage model, the

duplication itself is an advantage. This model also features the instability of the duplication since once the selection pressure is removed so that the increased dosage is no longer an advantage, the duplication would be removed as well. In the neofunctionalization or subfunctionalization models, the functional divergence emerges after the duplication, which contradicts the Ohno's dilemma. The IAD model, however, proposes new side functions preceding the gene duplication. This is especially probable for bacteria, which live together in large amounts as colonies and are under constant microevolution. The IAID model is different from the IAD model in that microevolution is raised as a major resource for innovation, which is illustrated with whole genome sequencing data of *K. pneumoniae*. This microevolution happens before and after the duplication, providing source for divergence. Also, HGT is not taken as another different way of getting new genes but as a means of the amplification step. Although no evidence of HGT was discovered in this study, one plasmid harbors this *LamB* gene, making it potentially possible to be passed to other genomes.

Maltoporin, coded by the *LamB* gene, was first identified as a lambda phage receptor and later proved to be a channel for sugar transportation. Various hypotheses can be made to explain the duplication and microevolution of the *LamB* gene. It is true that porins, as channels for molecules to diffuse, abound in the cell surface. As a result, it is likely that the increased dosage may be an advantage for survival. In the case of maltoporin, the duplicated copy may be needed for elevated expression of the maltose system during glucose starvation. At the same time, cell surface proteins are subject to strong selection due to immune pressure from the host [147], thus making fast

mutation and evolution necessary. This may also be true for maltoporin, which, on the one hand, functions as a transporter, and on the other hand, has to escape the immune system. Given the obvious difference in *K. pneumoniae LamB* gene copies from human host and those isolated from plants, this may be a proper explanation. Some studies correlate maltoporins with antibiotic resistance. Maltoporin is reported to be a negative regulator for antibiotic resistance in *E. coli*, which functions to influx CTC (an antibiotic) in complex with Odp1 [148]. Another study showed that in two clinical multidrug-resistant *E. aerogenes* strains, the expression of major porins is reduced while *LamB* is overexpressed [149]. It is true that the hypotheses need further experiments to validate.

Microevolution is used to refer to the accumulation of genetic changes in a few loci. Based on the different ratios of recombination to spread, populations undergoing microevolution can be divided to three structures [129]: (1) clonal structure; (2) panmictic structure, in which genetic recombination causes random association of loci; and (3) appear to be clonal because of the rapid epidemic spread of panmictic bacteria. Though there are differences between the three structures and the two sources of genetic mutation (point mutation and recombination), no attempts have been made to make a distinction whether the cloud-like population of the *LamB* gene is shaped primarily by point mutation or recombination, or which structure it really takes. Microevolution, as a force creating genetic changes, is the source of innovation for new gene functions to evolve.

The IAID model proposed, although illustrated only with an example for bacteria, can also be extended to other organisms. Even though the mutation

rate and population size may vary from case to case, the accumulation of mutations for innovation works for other organisms. The amplification step may vary in its mechanisms, but produces the same result of gene duplication. With the constant accumulation of mutations after gene duplication, divergence can be driven by selection pressures to produce new genes or novel functions of the gene.

While *LamB* gene serves as a good example illustrating the IAID model, there is no denying the possibility that other genes cannot be explained by this model or that other models can account for gene evolution via duplication. Since different genes differ in their functions, mutation rates, and the selection pressures they are under, they may have different mechanism to generate variations and may be driven by different forces to diverge once duplicated, thus allowing for the existence of different models addressing the same phenomenon.

Haplotype reconstruction methods are designed for reconstructing highly similar sequences in a single sequencing experiment and estimating the relative frequency. This is most often used for inferring intra-host genetic variation when multiple genomes are sequenced together in a single sequencing experiment. Haplotype reconstruction methods are most widely used for sequencing experiments of RNA virus due to the error-prone nature of RNA viruses and thus the high intra-host diversity. Such methods include ShoRAH [136], ViSpA [150], QuRe [151], all of which implements a pre-processing step designed for quality filtering and sequencing error correction, a reconstruction step making use of overlap graph, and a frequency estimation step after inferring the sequences. QuasQ differs from existing software by

putting more efforts on the error correction step, which thus reduces sequencing errors and solves in part the inflation of population size in haplotype inference.

5.5 Conclusion

The IAID (Innovation-Amplification-Innovation-Divergence) model was described to explain the generation of new genes by duplication, especially in bacteria. In this model, a gene with side functions generated by microevolution is amplified, after which microevolution still brings about innovations for each copy as they diverge from each other under selection pressure. One example is *LamB* gene that is duplicated in *K. pneumoniae* and other related species. With 34 complete genome sequences from NCBI, I showed that the duplication arising by tandem duplication and passing on to different genomes is stably maintained and the copies are driven to diverge from each other by different selection pressures. Haplotype reconstruction of whole genome sequences from 22 clinical isolates pictured the gene in each isolate as a population of similar sequences. These results suggest the efficacy of the IAID model in explaining the gene evolution by duplication in bacteria.

Chapter 6

SpoTyping: fast and accurate *in silico Mycobacterium* spoligotyping from sequencing reads

The content of this chapter has been published as [152]. Reproduction of figures and tables is permitted by the publisher.

6.1 Background

TB is an infectious disease caused mainly by *Mtb*. It is a top infectious disease killer around the world and remains an acute international health problem, resulting in an estimated 9.6 million new cases and 1.5 million deaths globally in 2014 [153]. Though TB burdens have decreased by nearly a half in the past 20 years, the global emergence and spread of drug-resistant TB have compounded the difficulty of treating and eradicating this disease.

Spoligotyping (spacer oligonucleotide typing) is a widely used genotyping method for *Mtb*, which exploits the genetic diversity in the clustered regularly interspersed short palindromic repeats (CRISPR) locus, which is also known as the direct repeat (DR) locus in *Mtb* genome [154]. Each DR region consists of several copies of the 36 bp DR sequence, which are interspersed with 34 bp to 41 bp non-repetitive spacers [155]. A set of 43 unique spacer sequences is used to classify *Mtb* strains based on their presence or absence. The patterns of presence and absence in each of the 43 spacer sequences can be summarized with a 43-digit binary code with '1' denoting the presence and '0' denoting the absence for each spacer, which can also be translated into a 15-digit octal code [156] termed as the spoligotype. Spoligotypes can be used to compare *Mtb* isolates collected between different laboratories. Traditionally, spoligotyping is conducted using PCR-based reverse line hybridization blotting technique [154]. Various new spoligotyping methods have been proposed recently, the most of which are microarrays, such as the PixSysn QUAD 4500 Microarrayer [157], DNA microarray [158], hydrogel microarray (biochip) [159], *Spoligorifytyping*

[160] and its follow up TB-SPRINT [161]. Other spoligotyping methods include those based on a matrix-assisted laser desorption/ionization time-of-flight mass-spectrometry (MALDI-ToF MS) platform [162, 163]. Spoligotyping has also been applied to strain typing in other bacterial species such as *Campylobacter jejuni* [164, 165], *Legionella pneumophila* [166], and *Salmonella* [167].

Though technological advancements in next-generation sequencing have enabled single-nucleotide resolution for *Mtb* phylogenetic studies by allowing the construction of a SNP-based phylogenetic tree, genotyping of bacteria is still needed for fast strain identification and correlation with previous isolates. For previous isolates, particularly the historical ones, genotypes including the spoligotype may have been determined as a routine, but whole genome sequencing data is not available and some isolates are not able to be sequenced. Under such circumstances, *in silico* genotyping from the whole genome sequences is necessary for correlating current isolates with previously genotyped ones. There are several molecular genotyping techniques for *Mtb*, of which the most widely used are: (1) spoligotyping; (2) Mycobacterial Interspersed Repetitive Units - Variable Numbers of Tandem Repeat (MIRU-VNTR) and (3) IS6110-based restriction fragment length polymorphism (IS6110-RFLP) [168]. Since the determination of MIRU-VNTR depends on determining the repeat number of tandem repeats, inferring MIRU-VNTR from next-generation sequencing reads involves resolving tandem repeats, which is extremely challenging for the current sequencing reads generated by the most widely used sequencing platforms due to their short lengths. IS6110-RFLP commonly has its result based on DNA fragment

blots on electrophoresis gel image and focuses on the determination of fragment lengths, which is also extremely challenging to infer since short-read sequencing cannot be used alone to construct finished genomes. Spoligotyping, therefore, provides a unique chance to obtain the same result from whole genome sequences as the molecular genotyping result achieved in laboratories, which can correlate isolates investigated using different approaches. *In silico* spoligotyping is also important in investigations using public data, where sequencing reads or complete genomic sequences are available but the spoligotypes of the isolates are not reported.

SpolPred [169] is a tool capable of accurately predicting the spoligotype of *Mtb* isolates using sequencing reads of uniform length obtained from platforms such as Illumina GAII and HiSeq. However, for sequencing reads generated by platforms marketed for clinical diagnostics such as Illumina MiSeq and Ion sequencers, where throughput is moderate and read lengths are non-uniform, the accuracy of SpolPred is significantly reduced. SpoTyping improves the performance of SpolPred in three ways: (1) SpolPred reads in a fixed number of bases from each sequencing read as specified by the user. As a result, for sequencing experiments with non-uniform read length, prediction accuracy is highly dependent on the choice of the read length by the user, which is hard to determine. SpoTyping, by reading in the full length of every read, makes use of all the available sequencing data. (2) SpolPred requires the user to specify a direction for the reads, which can be either direct or reverse. However, since each FASTQ file consists of both direct and reverse reads, SpolPred only utilizes a fraction of the input sequencing reads which can lead to incorrect predictions for

sequencing experiments with low throughput. SpoTyping explicitly considers the reads in both directions, thereby using all the information presented in the sequencing reads. (3) SpolPred relies on an inefficient sequence search algorithm; whereas SpoTyping integrates the BLAST algorithm in the search which can considerably reduce the time of the search. In addition to the improvements listed above, SpoTyping also comes with novel functions not found in SpolPred or other software previously: (1) For TB disease outbreak investigation, it is necessary to quickly identify isolates with matching spoligotypes. SpoTyping thus automatically queries SITVIT [170], a global *Mtb* molecular markers database to download associated epidemiological data for isolates with matched spoligotypes in an Excel spreadsheet, which can be presented as a graphical report showing the distribution summaries of the meta-data corresponding to the clades, years and countries of isolation for these isolates. (2) SpoTyping works on different input files such as next-generations sequencing reads in FASTQ format, and complete genomic sequences or assembled contigs in FASTA format. (3) SpoTyping can work on most operating systems such as Windows, Linux and Mac OS, either as a non-interactive script which can be integrated into individual analysis pipelines or as an interactive application with a graphical user interface. Thus, we believe SpoTyping would be a useful tool for public health surveillance and genotyping from next-generation sequencing data in clinical diagnostic of *Mtb* strains.

SpoTyping is written in Python, and is freely available at: https://github.com/xiaeryu/SpoTyping-v2.0.

6.2 Methods

6.2.1 Implementation

SpoTyping is implemented with Python and accepts two kinds of input files: single-end or pair-end sequencing reads in FASTQ format, and complete genomic sequences or assembled contigs in FASTA format. A schematic representation of the SpoTyping workflow is shown in Figure 26. When the input files are sequencing reads, SpoTyping first concatenates all sequencing reads in the input FASTQ file(s) into a single contiguous sequence in FASTA format, which would then be constructed into a BLAST [70] nucleotide database. The current program uses the swift mode by default, which, instead of processing all sequencing reads, reads in no more than 250 Mbp of the sequencing reads, which corresponds to a read depth of ~55X of the *Mtb* genome and would be sufficient in most situations. Disabling the swift mode would require SpoTyping to utilize all sequencing reads with increased execution time. The set of 43 spacer sequences, each of 25 bp in length, would be queried against the constructed database using nucleotide BLAST. The BLAST output is then parsed to determine the number of hits for each spacer sequence in the input file(s). At most one mismatch out of 25 bp of the spacer sequence is allowed for a BLAST match to be considered as a hit. For sequencing reads, if a spacer sequence is absent in the *Mtb* isolate, then no or very few hits would be identified, while if the number of hits exceeds a threshold (hit threshold, with a default of 5 error-free hits and 6 1-error-tolerant hits), it indicates the presence of the spacer sequence where the number of hits correlates with the

sequencing read depth of the locus. For genomic sequences or assembled contigs, the presence of one hit for a spacer sequence indicates the presence of the spacer. The 43-digit binary string, each digit representing one of the 43 spacer sequences with '0' indicating absence and '1' indicating presence, can therefore be written into an octal code that defines the spoligotype of the *Mtb* isolate. The predicted spoligotype is then automatically queried in the SITVIT database to retrieve all reported isolates having identical spoligotypes, where associated data corresponding to the MIRU12, VNTR, SIT, MIT, VIT, clade, country of origin, country of isolation, and year of report for these isolates would be downloaded in an Excel spreadsheet. SpoTyping also includes an R script that can present summary statistics of the associated meta-data as a pdf report.

6.2.2 Performance assessment: accuracy

The accuracy of SpoTyping was assessed in comparison with SpolPred on 3 datasets: (1) 161 isolates sequenced on Illumina HiSeq [SRA: SRA065095]; (2) 30 isolates sequenced on Illumina MiSeq [ENA: PRJNA218508]; and (3) 16 isolates sequenced on Ion Torrent [ENA: PRJEB6576].

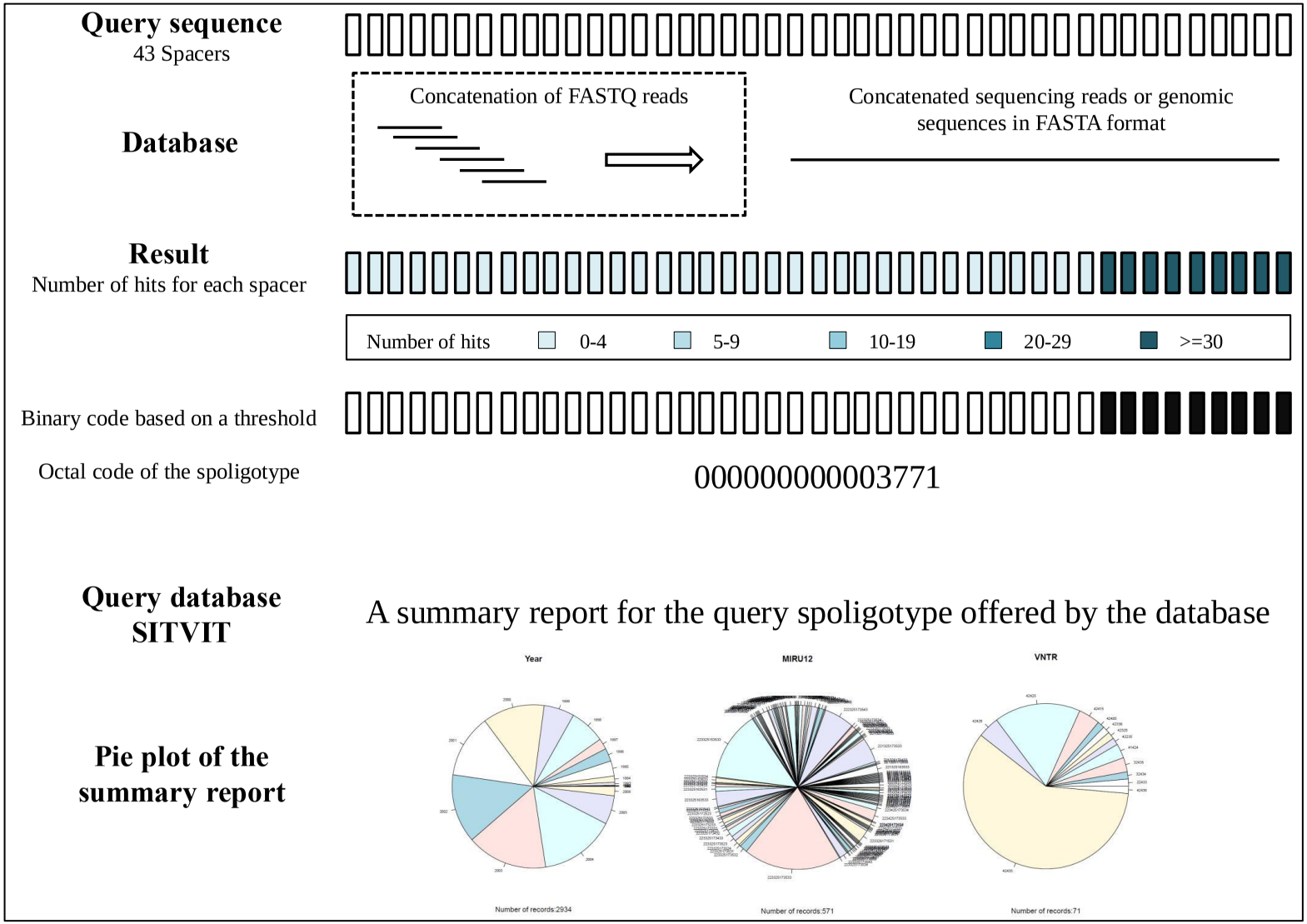

Query sequence
43 Spacers
Database
Concatenation of FASTQ reads
Concatenated sequencing reads or genomic sequences in FASTA format
Result
Number of hits for each spacer
Number of hits
0-4
5-9
10-19
20-29
>=30
Binary code based on a threshold
Octal code of the spoligotype
000000000003771
Query database SITVIT
A summary report for the query spoligotype offered by the database
Pie plot of the summary report
Year
MIRU12
VNTR
Number of records:2934
Number of records:571
Number of records:71

Figure 26. A schematic representation of the SpoTyping workflow. If the specified input contains sequencing reads, SpoTyping first concatenates the sequencing reads to form an artificial sequence. The artificial sequence, or genetic sequences when the input contains a complete genomic sequence or assembled contigs, would be built into a BLAST database. After querying the 43 spacer sequences in the database, the results are parsed to count the number of hits for each spacer sequence. A hit threshold is set to define a spacer as 'present' in the genome, resulting in a 43-digit binary code with '1' as present and '0' as absent, which is further translated into an octal code of the spoligotype. SITVIT database is then queried to identify matching isolates having the same spoligotype, where the associated data of the matched isolates are downloaded and summarized with pie charts.

The first assessment was conducted on a dataset of 161 *Mtb* isolates sequenced on Illumina HiSeq whose spoligotypes have been experimentally determined and reported [171]. Both SpoTyping and SpolPred were run with default parameters. The predicted octal codes were each queried in the SITVIT database to identify the matching spoligotype to compare with the reported spoligotype. Isolates with discordant results were examined by searching the spacer sequences on the contigs assembled using the *de novo* assembly software Velvet [172].

The next assessment was conducted on a dataset of 30 *Mtb* isolates sequenced on Illumina MiSeq without reported spoligotypes. The reference spoligotype for each isolate was determined by manual inspection of the BLAST output file to determine the number of hits for each spacer sequence in the sequencing reads. Given that the sequencing read depths are above 20X for all isolates, no hit for a spacer sequence is a strong indication of its absence while more than 5 hits is a strong indication of its presence. While a judgement cannot be safely made based on a hit number of 1-5, isolates with at least one such case were removed from the assessment, leaving only isolates with confident reference spoligotypes. SpoTyping was run with default

parameters while SpolPred calls for a specified read length, where a range of read lengths were used based on the read length percentiles from 0.04 to 1 at a step of 0.04, resulting in a total of 25 predictions for each isolate.

The accuracy of SpoTyping was also assessed in comparison with SpolPred on a dataset of 16 *Mtb* isolates sequenced on Ion Torrent. The reference spoligotypes were determined the same as those for Illumina MiSeq data. The running parameters were also similar as those for Illumina MiSeq data.

6.2.3 Performance assessment: execution time

The time performance of SpoTyping was compared with SpolPred based on the first dataset described above. The programs were run on a 64-bit Fedora Linux server workstation having a 2.0GHz quad processor and 32GB RAM. Both SpoTyping and SpolPred were run twice for each isolate with the swift mode either on or off. Default parameters were used for SpoTyping swift mode, while for non-swift mode, 10 error-free hits or 12 1-error-tolerant hits (options of -m 10 -r 12) was taken as the hit threshold due to the high sequencing read depth to eliminate false positives. For SpolPred, the pair-end sequencing reads were first concatenated (concatenation time was not counted toward the execution time). The read lengths were set to be the actual read lengths. The hit threshold was similarly set to be 10 (option of -m 10) in the non-swift mode.

6.2.4 Performance assessment: downsampling experiment

The performance of SpoTyping was next assessed at various sequencing read depths to determine its applicable range, where SpoTyping prediction accuracy was determined for: (1) an H37Ra *Mtb* isolate that was sequenced at a sequencing throughput of 3,000 Mbp (~670X); and (2) a Beijing-genotype *Mtb* isolate with a sequencing throughput of 2,700 Mbp (~600X) by performing 50 iterations each for six downsampling ratios of 50%, 20%, 10%, 5%, 2% and 1% of the initial number of reads for each isolate. In each downsampling experiment, a certain percent of the sequencing reads were randomly selected from the original FASTQ file to form a new file with a lower read depth, where the percentage is called the downsampling ratio. For all downsampling experiments, default settings were used except for the categories of 2% and 1% where the hit threshold was set to 2 error-free hits and 3 1-error-tolerant hits (options of –m 2 –r 3) due to the low read depths. The false positives caused by the concatenation of sequencing reads were also assessed in the downsampling experiment.

Sequencing reads of the Beijing-genotype isolate are deposited in European Nucleotide Archive under the code of ERP006354. The H37Ra isolate is a laboratory strain and was sequenced as part of a validation sequencing run.

6.2.5 Hit threshold selection

The selection of the hit thresholds was also based on the downsampling experiments. In each downsampling experiment, the number of both error-free hits and 1-error-

tolerant hits for each spacer identified by SpoTyping were divided by the estimated read depth (number of sequence bases/ 4,500,000) of the experiment, representing the number of hits as a percentage of the estimated read depth. For each spacer sequence in each experiment, the percentage is used as the feature to classify a spacer as present or absent, while the spacer's actual class of presence or absence is used to assess whether the classification is correct. A set of percentages was used as the thresholds to calculate the respective true positive rates and false positive rates, which were plotted as a receiver operating characteristic curve (ROC curve). The thresholds were selected to maximize the true positive rate while minimizing the false positive rate.

6.3 Results

6.3.1 *In silico* spoligotyping of 161 *Mtb* isolates sequenced on Illumina HiSeq

For all the 161 *Mtb* isolates, SpoTyping and SpolPred predicted the same spoligotypes, of which 20 isolates either without a match in the SITVIT database or reported as "New" were excluded from subsequent comparisons. Of the remaining 141 isolates, predictions of SpoTyping and spoligotypes determined in laboratory for 127 isolates (90.07%) were identical. For the 14 discordant isolates, the spacer sequences were searched in the assembled contigs to determine the spoligotypes, which are all concordant with the predictions from SpoTyping.

6.3.2 *In silico* spoligotyping of 30 *Mtb* isolates sequenced on Illumina MiSeq

The accuracy of SpoTyping was then assessed in comparison with SpolPred on 30 *Mtb* isolates sequenced on Illumina MiSeq, among which 21 passed filtering for having reference spoligotypes confidently determined. SpoTyping correctly inferred the spoligotypes for all 21 isolates. Since SpolPred requires a read length to be specified, a range of read lengths were assessed based on the percentiles from 0.04 to 1 at a step of 0.04, resulting in a total of 25 predictions for each isolate. At each percentile, the predictions for the 21 isolates were analyzed to calculate the prediction accuracy, which is summarized in Figure 27. SpolPred performs the best using the read lengths at the 0.36, 0.40 or 0.44 percentiles, with accuracies around 50%. The prediction accuracy of SpolPred is significantly lower than that obtained by SpoTyping and is also highly dependent on the choice of read length used as input, which, in itself, is difficult to determine.

6.3.3 *In silico* spoligotyping of 16 *Mtb* isolates sequenced on Ion Torrent

The accuracy for spoligotype inference was also determined on 16 *Mtb* isolates sequenced on Ion Torrent with spoligotypes reported to be all Beijing genotype [173]. Of the 16 isolates, 11 have confidently determined spoligotypes, which are all of the spoligotype ‘000000000003771’ as are consistent with the reported Beijing genotype. SpoTyping makes correct prediction for all the 11 isolates. The performance of SpolPred is summarized in Figure 27. SpolPred performs best using the read length at the 0.08 and 0.12 percentile, with accuracies of only around 10%.

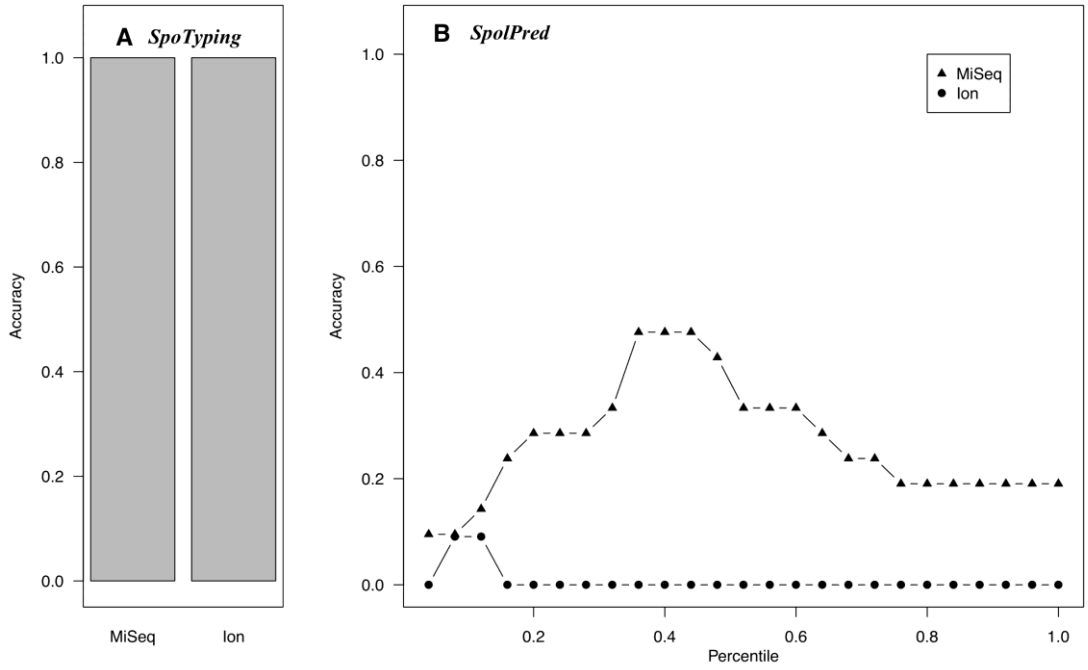


Figure 27. Prediction accuracy of *Mtb* isolates sequenced on Illumina MiSeq and Ion Torrent. SpolPred requires a read length to be specified, which results in inconsistent predictions when using different specifications. The accuracy assessment was conducted between SpoTyping (A) and SpolPred (B) on 21 MiSeq-sequenced isolates and 11 Ion-sequenced isolates, with SpoTyping predictions using default parameters and SpolPred predictions using different read length percentiles as the input read lengths. While SpoTyping have perfect accuracies for both datasets, SpolPred gives varying accuracies depending on the read length, but are always lower than 50%.

6.3.4 Comparison of time performance for SpoTyping and SpolPred on 161 *Mtb* isolates

For the 161 *Mtb* isolates assessed, SpoTyping is about 20-40 times faster than SpolPred, with SpoTyping taking an average of 28.8 sec (standard deviation is 5.3 sec) in its swift mode, and an average of 56.4 sec (standard deviation is 8.0 sec) to process all reads, while SpolPred took an average of 17 min 19.3 sec (standard deviation is 1 min 35.3 sec) by using the –s option, or an average of 18 min 20.0s (standard deviation is 50.2 sec) to process all reads.

6.3.5 Downsampling experiments

Based on the downsampling experiments which first explore the applicable throughput for accurate spoligotype inference, SpoTyping is able to efficiently and accurately predict the spoligotype for isolates having sequencing throughput over 54 Mbp (read depth of ~12X) with accuracies above 98% (Figure 28, Table 10 for H37Ra, and Table 11 for Beijing). However, in experiments with very low throughput (read depth below 10X), lowering hit thresholds is still not sufficient to make accurate predictions as some of the spacer sequences would not be adequately sequenced and represented in the input FASTQ file(s).

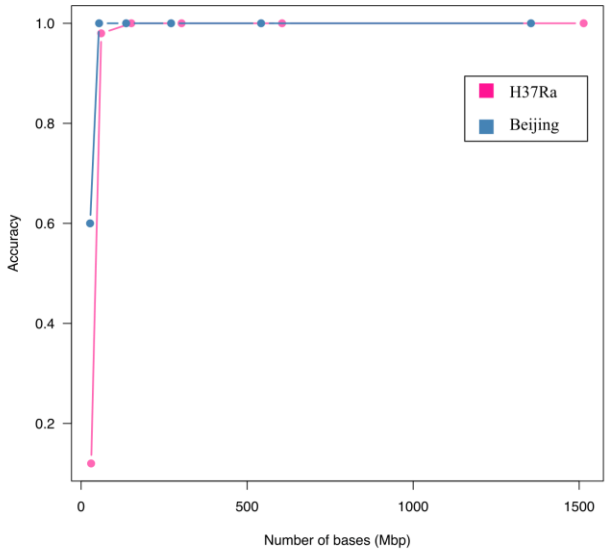


Figure 28. Assessing the accuracy of SpoTyping across various sequencing read depths for H37Ra and Beijing-genotype isolates. With blue points denoting the Beijing genotype, pink points denoting H37Ra, the prediction accuracies were assessed with the sequencing throughput measured by the number of bases for all the downsampling experiments. SpoTyping is suitable for sequencing runs whose throughput are over 54 Mbp (read depth of ~12X), where the accuracy is almost 100%.

Table 10. Statistics of time and accuracy of running SpoTyping on 50 iterations each for various downsampling ratios of an H37Ra *Mtb* isolate.

Downsampling ratio		1+	0.5	0.2	0.1	0.05	0.02	0.01
# Read pairs	Mean (M^)	199.1	99.58	39.83	19.91	9.96	3.98	1.99
	SD	NA	2,901	1,765	1,372	1,173	626	356
# Bases	Mean (M^)	3,027	1,513	605	302	151	60	30
	SD	NA	440,98	268,22	208,49	178,22	95,223	54,112
Estimated coverage*	Mean	672.7	336.35	134.53	67.26	33.63	13.46	6.73
	SD	NA	0.1	0.06	0.05	0.04	0.02	0.01
Time elapsed (s)	Mean	25.936	40.476	40.068	50.663	24.351	7.705	4.698
	SD	NA	1.534	1.257	2.148	2.169	0.834	0.639
Accuracy		1	100%	100%	100%	100%	98%	12%

+ No downsampling was performed
* The coverage is estimated by (#bases/4,500,000)
^ In the unit of a factor of one million

Table 11. Statistics of time and accuracy of running SpoTyping on 50 iterations each for various downsampling ratios of a Beijing-genotype *Mtb* isolate.

Downsampling ratio		1+	0.5	0.2	0.1	0.05	0.02	0.01
# Read pairs	Mean (M^)	17.83	8.91	3.57	1.78	0.89	0.36	0.18
	SD	NA	1,921	1,683	1,265	845	528	410
# Bases	Mean (M^)	2,710	1,355	542	271	136	54	27
	SD	NA	292,03	255,76	192,30	128,51	80,184	62,321
Estimated coverage*	Mean	602.24	301.12	120.45	60.22	30.11	12.05	6.02
	SD	NA	0.06	0.06	0.04	0.03	0.02	0.01
Time elapsed (s)	Mean	25.301	38.778	38.506	42.15	20.276	6.427	3.977
	SD	NA	1.732	1.945	2.098	0.807	0.296	0.535
Accuracy		1	100%	100%	100%	100%	100%	60%

+ No downsampling was performed
* The coverage is estimated by (#bases/4,500,000)
^ In the unit of a factor of one million

Since SpoTyping concatenates sequencing reads into an artificial sequence to create the BLAST database, an immediate concern is the false positives created due to chimera sequences. In all of 600 downsampling experiments performed for both H37Ra and Beijing genotype *Mtb* isolates, the maximum number of false positive hit is 1 for both error-free hits and 1-error-tolerant hits. Of the experiments, 98.3%

(590/600) show no false-positive error-free hits while 95.7% (574/600) show no false-positive 1-error-tolerant hits. The likelihood of false positives created due to chimera sequences is thus low, which can be further reduced by setting more stringent hit thresholds.

6.3.6 Hit threshold selection

The choice of hit thresholds to determine the presence or absence of a spacer sequence used in SpoTyping was evaluated. The evaluation was conducted in the downsampling experiments, based on the groups with downsampling ratios from 2% to 50% (read depths between ~12X and ~300X) where accurate inferences for the spacer sequences are possible to be made. A total of 21,586 spacer sequence instances ((5 downsampling ratios * 50 rounds for each downsampling ratio * 43 spacer for each round + 43 spacers without downsampling) = 10,793 spacers for each of the two strains) with their respective number of hits identified by SpoTyping were included in the analysis, of which 10,040 are absent cases and 11,546 are present cases. The number of hits was divided by the estimated read depth to represent the number of hits as a percentage of the read depth in order to adjust for the difference in sequencing throughput. A set of percentages was used as the thresholds to calculate the respective true positive rates and false positive rates, which were plotted as an ROC curve (Figure 29). The ROC curves for both the error-free hits (Figure 29A) and 1-error-tolerant hits (Figure 29B) show very high true positive rates and very low false positive rates, with the areas under the ROC being 0.9999997 and 0.9999998,

respectively. False positive rates are always nearly 0, while the true positive rates are above 99% by setting the thresholds to be 1.80% to 14.86% of the read depth for error-free hits and 1.80% to 14.88% of the read depth for 1-error-tolerant hits. Thus the default thresholds of 5 error-free hits and 6 1-error-tolerant hits are applicable to sequencing experiments with estimated read depths between ~30X and ~280X. The thresholds can be adjusted accordingly given sequencing throughputs beyond this range.

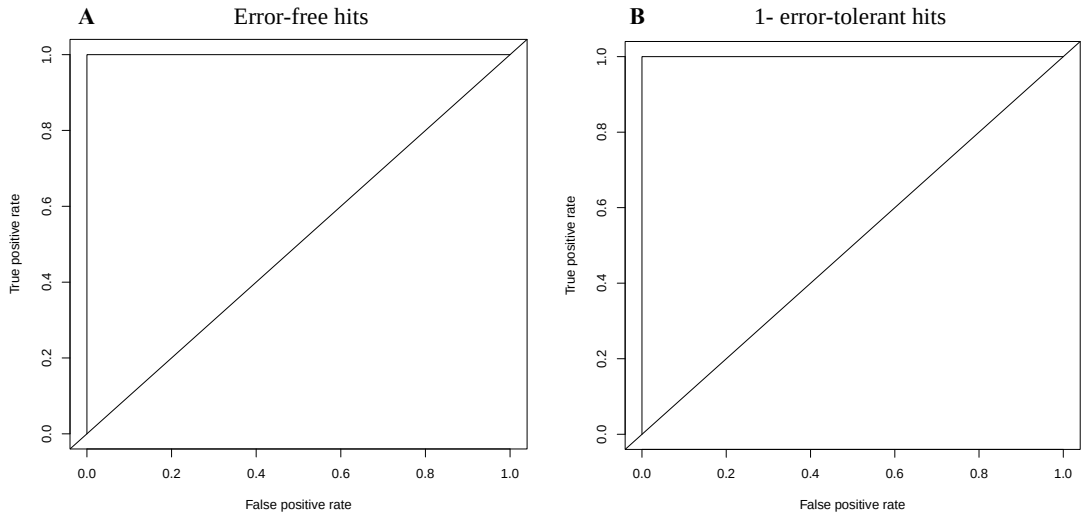


Figure 29. ROC curves for the selection of hit thresholds. The ROC curves were plotted for both error-free hits (A) and 1-error-tolerant hits (B) to select the hit thresholds. Diagonal lines, also known as lines of no discrimination, were plotted as references of random guess. The threshold evaluation was based on a percentage calculated as the number of hits divided by the estimated read depth. A set of percentages was used as the thresholds to calculate the respective true positive rates and false positive rates, which were plotted as the ROC curves. Both ROC curves show constantly high true positive rates and low false positive rates, with the areas under the ROC curve being 0.9999997 and 0.9999998, respectively.

6.4 Discussion

The global burden of TB, especially drug-resistant strains, has put a significant spotlight on pathogen whole genome sequencing as a rapid diagnostic tool, which is of great relevance to both public health surveillance and clinical treatment. The application of next-generation sequencing in clinical microbiology requires fast and easy-to-use software that is able to accurately produce easily comprehensible results. As shown, SpoTyping is able to accurately determine the spoligotype of *Mtb* isolates rapidly. Contrary to SpolPred which is sensitive to the user-specified read length and gives inconsistent predictions at different read lengths, SpoTyping gives accurate predictions based on sequencing reads produced from different sequencing platforms regardless of the length uniformity of the sequencing reads and is 20 to 40 times faster than SpolPred. Additional functions of SpoTyping include: (1) database query, where the predicted spoligotype is automatically queried in the SITVIT database to retrieve all associated epidemiological data corresponding to the MIRU12, VNTR, SIT, MIT, VIT, clade, country of origin, country of isolation, and year of report; and (2) information visualization, where the retrieved information would be summarized, visualized, and presented as a report. These additional functions would be useful for public health surveillance of *Mtb* strains causing TB.

While there are several molecular typing techniques for *Mtb*, the most widely used are spoligotyping, MIRU-VNTR and IS6110-RFLP. Spoligotyping, though being a relatively simple, cost-effective, and high-throughput method, suffers from the limitations of: (1) having relatively low discriminatory power [174] due to its use of

only a single genetic locus for genotyping; and (2) having limited use in phylogenetic study. Among the genotyping methods for *Mtb*, a combination of spoligotyping and MIRU-VNTR was reported to be the best strategy [175, 176]. However, significant technical challenges currently exist for accurate *in silico* typing from next-generation sequencing reads of MIRU-VNTR which involves resolving tandem repeats and IS6110-RFLP whose result is based on DNA fragment blots on electrophoresis gel image and thus involves the determination of DNA fragment lengths. Spoligotyping, as a result, provides a unique chance to obtain the same result from whole genome sequences as the molecular typing result achieved in laboratories, which can correlate the isolates investigated with different approaches. Though spoligotyping has less discrimination power than SNP phylogeny inferred from whole genome sequences, it is unique in correlating the genomic data produced in research laboratories and the molecular typing data from clinical laboratories. Thus *in silico* spoligotyping is not only a genotyping method for *Mtb* isolate differentiation, but also a bridge between isolates investigated with whole genome sequencing and isolates investigated with traditional laboratory protocols, especially those historical isolates that are not sequenced. Inexorably, clinical surveillance and management of TB, particularly for disease diagnosis and treatment, will progress towards the use of direct *Mtb* sequencing. Thus the ease of use and interpretability of the results will be of considerable importance to users within a public health setting, which is well achieved with SpoTyping.

A recently published letter reported CASTB, an analysis server for the *Mycobacterium tuberculosis* complex, which provides next-generation sequencing data analysis tools for virtual typing (spoligotyping included), virtual drug resistance analysis, and phylogenetic analysis [177]. While the webserver provides a comprehensive overview of the sequencing data, the performance of each tool is not well evaluated in the publication. More accurate and well assessed tools are thus needed for further analysis. SpoTyping is here assessed to provide high accuracy for *in silico* spoligotyping and thus demonstrates the reliability of the results. SpoTyping also benefits from its open source nature that it can be easily integrated into in house analysis pipelines for in-depth analysis of the sequencing data. When talking about execution time, services provided by webservers may be very slow due to the inherent issues such as the process of data uploading and the availability of the computational resources. SpoTyping, on the other hand, can be setup locally and provides the spoligotyping result within a minute.

For the 14 discordant spoligotypes between the laboratory tests and the *in silico* predictions made by SpoTyping in the 161 *Mtb* isolates sequenced on Illumina HiSeq, the SNP-based phylogenetic tree of these 161 Mtb isolates in the original article [171] was examined to compare the lineage with the spoligotyping results. Out of the 14 discordant results, 3 showed better concordance of the *in silico* prediction with the lineage shown on the tree. As an example, an isolate (Accession: SRR671868, Strain: 143) located at Lineage 4.2 on the SNP-based phylogenetic tree is reported to be Beijing genotype based on the laboratory test in the publication, while predicted to be

T2 genotype by SpoTyping. However, Beijing genotype is usually found at East Asia Lineage 2, while Lineage 4 usually harbors the Euro-American genotypes. One of the discrepancies may be caused by the different naming of spoligotypes in different databases (Beijing and Beijing-like). Definite conclusion cannot be made for the remaining 10 isolates for which the reported spoligotype and *in silico* predicted spoligotype are different while the lineages for both spoligotypes are similar (T2 and H3, for example). For such isolates, the difference could be due to the discrepancy between laboratory tests and the genomic features.

SpoTyping would not be able to differentiate between mixed infections as spacers deleted in one strain may be compensated by reads from another strain, thus making an incorrect inference of presence of the spacer sequence.

6.5 Conclusion

SpoTyping is an accurate, fast and easy-to-use program for *in silico* spoligotyping of *Mtb* isolates from next-generation sequencing reads, complete genomic sequences, and assembled contigs. In addition, SpoTyping automatically queries the global *Mtb* molecular markers database SITVIT to retrieve associated data for matching isolates with the inferred spoligotypes, which can be summarized graphically to generate a report. SpoTyping would be a useful tool for public health surveillance and genotyping of *Mtb* strains.

Chapter 7

Discussion

7.1 Longer reads can do more

Illumina sequencing has been the most widely used sequencing technique in bacteria genomics. While bearing the merit of high accuracy, reads generated by Illumina sequencing has relatively short length (pair-end reads of up to 150 bp in HiSeq, and 250 bp in MiSeq). The short read lengths may not cause problems for reference-based reads mapping and variant calling, but may be a limitation in bacteria genomics, where *de novo* assembly is widely used. Repeats are notoriously hard to resolve when their lengths are longer than the sequencing read lengths. Tandem repeats are repeats where repetitions are directly adjacent to each other, and may describe patterns that help to determine an individual's traits. MIRU-VNTR, a genotyping method for *Mtb*, for example, involves the determination of repetition numbers in tandem repeats, and is not feasible with short sequencing reads. There are also repeat sequences like insertion sequences, transposable elements, and duplicated genes that cannot be adequately resolved by short sequencing reads, thus confounding *de novo* assembly, and making it extremely difficult to construct complete genomes with only these reads. Accuracy of haplotype reconstruction described in Chapter 5 is also limited by read length. Thus longer sequencing reads can achieve more in bacteria genomics if sequencing quality is not undermined.

Efforts have been made to increase sequencing read length. The company Pacific Biosciences has achieved the success by using the SMRT technology for sequencing, which was reported to have a throughput of 500Mbp to 1Gbp per cell with half of the reads longer than 14Kbp, 5% of the reads longer than 24Kbp and a maximum read length of longer than 40Kbp. Other attempts like

the Oxford Nanopore sequencing also provide increased read length. Though generating long reads in several kilo bases, single-molecule sequencing approaches have quite high error rates (15.4% to 17.9% [178]). As a result, methods have been proposed [178, 179] to finish bacterial genomes using a combination of high quality short reads from next-generation sequencing and less accurate long sequencing reads, exploiting both merits of higher accuracy and longer read lengths, respectively.

7.2 Experience with different sequencing platforms

During my PhD training, I have encountered sequencing reads from multiple platforms: Illumina MiSeq sequencing, Illumina HiSeq sequencing, 454 sequencing, Ion Proton sequencing, and PacBio SMRT sequencing.

Illumina MiSeq sequencing is most widely used in sequencing bacterial genomes, which provides highly accurate pair-end reads of up to 250 bp in length. The accuracy and the relatively long read length make MiSeq optimum among the platforms for *de novo* assembly when used alone, though longer reads will still help to improve assembly quality. Compared to MiSeq, HiSeq have higher throughput but shorter read length. Given the importance of read length in *de novo* assembly, HiSeq is more often used in sequencing of chromosomes of clonal bacterial like *Mtb,* where reference-based reads mapping would be used. The major error type for Illumina sequencing is substitutions, which does not call for special pre-processing given sufficient read depth (~50X).

Roche's 454 sequencing was used once for sequencing dengue virus in order to do haplotype reconstruction. Back in 2012, 454 and Illumina were the

most used sequencing platforms. Reads generated by 454 sequencers have the advantage of being longer (~700 bp), but also the disadvantages of having: (1) higher error rates; and (2) much higher cost. The major error types of 454 are insertions and deletions, which needs to be considered in the pre-processing step. Reads of extreme lengths, which are correlated with low sequencing quality [134], may also need to be removed at the pre-processing step. As Roche announced the plan to shut down the 454 sequencing business, people tend to use it less and less.

Ion Torrent sequencing was used once as a trial run, where the performance appeared similar to that of Illumina in terms of the relatively short read length, and similar to that of 454 in terms of the higher error rates (insertions and deletions, primarily), thus not optimum for our research purposes. However, Ion Torrent sequencing has the advantage of having very fast speed and relatively low throughput per run, making it ideal for clinical diagnostics laboratories where rapid sequencing of a small number of isolates is required.

PacBio SMRT sequencing was also used as a trial run, which managed to complete 4 out of 5 pieces of DNA in a *K. pneumoniae* isolate though raw sequencing reads and sequencing design like how many cells were actually used were not provided. It would be very useful in bacteria genomics studies where genomes are plastic and complete genomes are needed for better characterization of the isolates if the price is not that high.

Bibliography

1. Sanger F, Coulson AR: **A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.** *J Mol Biol* 1975, **94**:441–448.

2. Sanger F, Nicklen S: **DNA sequencing with chain-terminating**. 1977, **74**:5463–5467.

3. Maxam AM, Gilbert W: **A new method for sequencing DNA.** *Proc Natl Acad Sci U S A* 1977, **74**:560–564.

4. Elaine R. Mardis: **Next generation sequencing methods.pdf**. 2008:387–402.

5. Morey M, Fernández-Marmiesse A, Castiñeiras D, Fraga JM, Couce ML, Cocho J a: **A glimpse into past, present, and future DNA sequencing.** *Molecular genetics and metabolism* 2013:3–24.

6. Schadt EE, Turner S, Kasarskis A: **A window into third-generation sequencing**. *Hum Mol Genet* 2010, **19**.

7. Kircher M, Kelso J: **High-throughput DNA sequencing--concepts and limitations.** *Bioessays* 2010, **32**:524–536.

8. McCutcheon JP, Von Dohlen CD: **An interdependent metabolic patchwork in the nested symbiosis of mealybugs**. *Curr Biol* 2011, **21**:1366–1372.

9. Han K, Li Z, Peng R, Zhu L, Zhou T, Wang L, Li S, Zhang X, Hu W, Wu Z, Qin N, Li Y: **Extraordinary expansion of a Sorangium cellulosum genome from an alkaline milieu.** *Sci Rep* 2013, **3**:2101.

10. Brilli M, Mengoni A, Fondi M, Bazzicalupo M, Liò P, Fani R: **Analysis of plasmid genes by phylogenetic profiling and visualization of homology relationships using Blast2Network.** *BMC Bioinformatics* 2008, **9**:551.

11. Pallen MJ, Wren BW: **Bacterial pathogenomics.** *Nature* 2007, **449**:835–842.

12. Dobrindt U, Hochhut B, Hentschel U, Hacker J: **Genomic islands in pathogenic and environmental microorganisms.** *Nat Rev Microbiol* 2004, **2**:414–424.

13. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW: **Transforming clinical microbiology with bacterial genome sequencing**. *Nature Reviews Genetics* 2012:601–612.

14. He M, Miyajima F, Roberts P, Ellison L, Pickard DJ, Martin MJ, Connor TR, Harris SR, Fairley D, Bamford KB, D'Arc S, Brazier J, Brown D, Coia JE, Douce G, Gerding D, Kim HJ, Koh TH, Kato H, Senoh M, Louie T, Michell S, Butt E, Peacock SJ, Brown NM, Riley T, Songer G, Wilcox M, Pirmohamed M, Kuijper E, et al.: **Emergence and global spread of epidemic healthcare-associated Clostridium difficile.** *Nat Genet* 2013, **45**:109–13.

15. Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, Paxinos EE, Sebra R, Chin C-S, Iliopoulos D, Klammer A, Peluso P, Lee L, Kislyuk AO, Bullard J, Kasarskis A, Wang S, Eid J, Rank D, Redman JC, Steyert SR, Frimodt-Møller J, Struve C, Petersen AM, Krogfelt KA, Nataro JP, Schadt EE, Waldor MK: **Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany.** *N Engl J Med* 2011, **365**:709–717.

16. von Mentzer A, Connor TR, Wieler LH, Semmler T, Iguchi A, Thomson NR, Rasko D a, Joffre E, Corander J, Pickard D, Wiklund G, Svennerholm A-M, Sjöling Å, Dougan G: **Identification of enterotoxigenic Escherichia coli (ETEC) clades with long-term global distribution**. *Nat Genet* 2014, **46**:1321–1326.

17. Salipante SJ, Roach DJ, Kitzman JO, Snyder MW, Stackhouse B, Butler-wu SM, Lee C, Cookson BT, Shendure J: **Large-scale genomic sequencing of extraintestinal pathogenic Escherichia coli strains**. *Genome Res* 2015:119–128.

18. Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Henderson DK, Palmore TN, Segre JA: **Tracking a Hospital Outbreak of Carbapenem-Resistant Klebsiella pneumoniae with Whole-Genome Sequencing**. *Science Translational Medicine* 2012:148ra116–148ra116.

19. Harris S, Feil E, Holden M: **Evolution of MRSA during hospital transmission and intercontinental spread**. *Science (80-)* 2010.

20. Köser CU, Holden MTG, Ellington MJ, Cartwright EJP, Brown NM, Ogilvy-Stuart AL, Hsu LY, Chewapreecha C, Croucher NJ, Harris SR, Sanders M, Enright MC, Dougan G, Bentley SD, Parkhill J, Fraser LJ, Betley JR, Schulz-Trieglaff OB, Smith GP, Peacock SJ: **Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak.** *The New England journal of medicine* 2012:2267–75.

21. Chan JZ-M, Sergeant MJ, Lee OY-C, Minnikin DE, Besra GS, Pap I, Spigelman M, Donoghue HD, Pallen MJ: **Metagenomic Analysis of Tuberculosis in a Mummy**. *NEJM* 2013, **369**:289–290.

22. Eyre DW, Cule ML, Griffiths D, Crook DW, Peto TEA, Walker AS, Wilson DJ: **Detection of Mixed Infection from Bacterial Whole Genome Sequence Data Allows Assessment of Its Role in Clostridium difficile Transmission**. *PLoS Comput Biol* 2013, **9**.

23. Lieberman TD, Flett KB, Yelin I, Martin TR, McAdam AJ, Priebe GP, Kishony R: **Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures.** *Nat Genet* 2014, **46**:82–7.

24. Loman NJ, Constantinidou C, Christner M, Rohde H, Chan JZ-M, Quick J, Weir JC, Quince C, Smith GP, Betley JR, Aepfelbacher M, Pallen MJ: **A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104:H4.** *JAMA* 2013, **309**:1502–10.

25. Andrews S: **FastQC: A quality control tool for high throughput sequence data**. *available from http//www.bioinformatics.babraham.ac.uk/projects/fastqc/* 2010:1.

26. Bolger AM, Lohse M, Usadel B: **Trimmomatic: A flexible trimmer for Illumina sequence data**. *Bioinformatics* 2014, **30**:2114–2120.

27. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18**:821–9.

28. Bankevich A, Nurk S, Antipov D, Gurevich A a., Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin A V., Sirotkin A V., Vyahhi N, Tesler G, Alekseyev M a., Pevzner P a.: **SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing**. *J Comput Biol* 2012, **19**:455–477.

29. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu YYYY, Tang J, Wu G, Zhang H, Shi Y, Liu YYYY, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu S-M, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang JJ, Lam T-W, Wang JJ: **SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler.** *Gigascience* 2012, **1**:18.

30. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9**:357–9.

31. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**:1754–60.

32. Ning Z, Cox AJ, Mullikin JC: **SSAHA: A fast search method for large DNA databases**. *Genome Res* 2001, **11**:1725–1729.

33. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup 1000 Genome Project Data Processing: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**:2078–2079.

34. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res* 2010, **20**:1297–303.

35. Huelsenbeck JP, Ronquist F: **MRBAYES: Bayesian inference of phylogeny**. *Bioinformatics* 2001, **17**:754–755.

36. Swofford DL: **Phylogenetic Analysis Using Parsimony**. *Options* 1996, **42**:294–307.

37. Stamatakis A: **RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies**. *Bioinformatics* 2014, **30**:1312–1313.

38. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O: **New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0**. *Syst Biol* 2010, **59**:307–321.

39. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods**. *Mol Biol Evol* 2011, **28**:2731–2739.

40. Felsenstein J: **Phylip: phylogeny inference package (version 3.2)**. *Cladistics* 1989, **5**:164–166.

41. Daubin V, Gouy M, Perrière G: **A phylogenomic approach to bacterial phylogeny: Evidence of a core of genes sharing a common history**. *Genome Res* 2002, **12**:1080–1090.

42. Jain R, Rivera MC, Lake JA: **Horizontal gene transfer among genomes: the complexity hypothesis.** *Proc Natl Acad Sci U S A* 1999, **96**:3801–3806.

43. Kaas RS, Friis C, Ussery DW, Aarestrup FM: **Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse Escherichia coli genomes**. *BMC Genomics* 2012:577.

44. Sanderson M, Doyle JJ: **Reconstruction of Organismal and Gene Phylogenies from Data on Multigene Families: Concerted Evolution, Homoplasy, and Confidence**. *Syst Biol* 1992, **41**:4–17.

45. Achtman M: **Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens.** *Annu Rev Microbiol* 2008, **62**:53–70.

46. Croucher NJ, Harris SR, Fraser C, Quail M a, Burton J, van der Linden M, McGee L, von Gottberg A, Song JH, Ko KS, Pichon B, Baker S, Parry CM, Lambertsen LM, Shahinas D, Pillai DR, Mitchell TJ, Dougan G, Tomasz A, Klugman KP, Parkhill J, Hanage WP, Bentley SD: **Rapid pneumococcal evolution in response to clinical interventions.** *Science (New York, N.Y.)* 2011:430–4.

47. Dykhuizen DE, Green L: **Recombination in Escherichia coli and the definition of biological species**. *J Bacteriol* 1991, **173**:7257–7268.

48. Ochman H, Lawrence JG, Groisman EA: **Lateral gene transfer and the nature of bacterial innovation.** *Nature* 2000, **405**:299–304.

49. Eisenstark a: **Genetic recombination in bacteria.** *Annu Rev Genet* 1977, **11**:369–396.

50. Llosa M, Gomis-Rüth FX, Coll M, De la Cruz F: **Bacterial conjugation: A two-step mechanism for DNA transport**. *Molecular Microbiology* 2002:1–8.

51. Chen I, Christie PJ, Dubnau D: **The ins and outs of DNA transfer in bacteria.** *Science* 2005, **310**:1456–1460.

52. Chen I, Dubnau D: **DNA uptake during bacterial transformation.** *Nat Rev Microbiol* 2004, **2**:241–249.

53. Ozeki H, Ikeda H: **Transduction Mechanisms**. *Annual Review of Genetics* 1968:245–278.

54. Croucher NJ, Page a. J, Connor TR, Delaney a. J, Keane J a., Bentley SD, Parkhill J, Harris SR: **Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins**. *Nucleic Acids Res* 2014, **43**:e15–e15.

55. Spratt BG, Hanage WP, Feil EJ: **The relative contributions of recombination and point mutation to the diversification of bacterial clones**. *Current Opinion in Microbiology* 2001:602–606.

56. Schierup MH, Hein J: **Consequences of recombination on traditional phylogenetic analysis**. *Genetics* 2000, **156**:879–891.

57. Schierup MH, Hein J: **Recombination and the molecular clock.** *Molecular biology and evolution* 2000:1578–1579.

58. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW: **GARD: A genetic algorithm for recombination detection**. *Bioinformatics* 2006, **22**:3096–3098.

59. Awadalla P: **The evolutionary genomics of pathogen recombination.** *Nat Rev Genet* 2003, **4**:50–60.

60. Sawyer S: **Statistical tests for detecting gene conversion.** *Mol Biol Evol* 1989, **6**:526–538.

61. Smith JM: **Analyzing the mosaic structure of genes.** *J Mol Evol* 1992, **34**:126–129.

62. Martin D, Rybicki E: **RDP: detection of recombination amongst aligned sequences.** *Bioinformatics* 2000, **16**:562–563.

63. Maynard Smith J, Smith NH: **Detecting recombination from gene trees.** *Mol Biol Evol* 1998, **15**:590–599.

64. Fang F, Ding J, Minin VN, Suchard MA, Dorman KS: **cBrother: Relaxing parental tree assumptions for Bayesian recombination detection**. *Bioinformatics* 2007, **23**:507–508.

65. Grassly NC, Holmes EC: **A likelihood method for the detection of selection and recombination using nucleotide sequences.** *Mol Biol Evol* 1997, **14**:239–247.

66. Stephens JC: **Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion.** *Mol Biol Evol* 1985, **2**:539–556.

67. Didelot X, Falush D: **Inference of bacterial microevolution using multilocus sequence data.** *Genetics* 2007, **175**:1251–66.

68. Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR, Bentley SD, Corander J: **Detection of recombination events in bacterial genomes from large population samples**. *Nucleic Acids Res* 2012, **40**.

69. Miller JH: **Mutational specificity in bacteria.** *Annu Rev Genet* 1983, **17**:215–238.

70. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic Local Alignment Search Tool**. *J Mol Biol* 1990:403–410.

71. Zhou Y, Call DR, Broschat SL: **Genetic relationships among 527 Gram-negative bacterial plasmids.** *Plasmid* 2012, **68**:133–41.

72. Call DR, Singer RS, Meng D, Broschat SL, Orfe LH, Anderson JM, Herndon DR, Kappmeyer LS, Daniels JB, Besser TE: **blaCMY-2-positive IncA/C plasmids from Escherichia coli and Salmonella enterica are a distinct component of a larger lineage of plasmids**. *Antimicrob Agents Chemother* 2010, **54**:590–596.

73. Grubbs FE: **Procedures for Detecting Outlying Observations in Samples**. *Technometrics* 1969, **11**:1–21.

74. Ramaswamy S, Rastogi R, Shim K: **Efficient algorithms for mining outliers from large data sets**. *ACM SIGMOD Rec* 2000, **29**:427–438.

75. Clermont O, Bonacorsi S, Bingen E: **Rapid and simple determination of the Escherichia coli phylogenetic group.** *Appl Environ Microbiol* 2000, **66**:4555–8.

76. Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MCJ, Ochman H, Achtman M: **Sex and virulence in Escherichia coli: an evolutionary perspective.** *Mol Microbiol* 2006, **60**:1136–51.

77. Holland BR, Huber KT, Moulton V, Lockhart PJ: **Using consensus networks to visualize contradictory evidence for species phylogeny.** *Mol Biol Evol* 2004, **21**:1459–61.

78. Huson DH, Bryant D: **Application of phylogenetic networks in evolutionary studies**. *Mol Biol Evol* 2006, **23**:254–267.

79. Treangen TJ, Ondov BD, Koren S, Phillippy AM: **The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes.** *Genome Biol* 2014, **15**:524.

80. Uchiyama I: **Multiple genome alignment for identifying the core structure among moderately related microbial genomes.** *BMC Genomics* 2008, **9**:515.

81. Zhang J: **Evolution by gene duplication: an update**. *Trends Ecol Evol* 2003, **18**:292–298.

82. Khong WX, Xia E, Marimuthu K, Xu W, Teo Y-Y, Tan EL, Neo S, Krishnan PU, Ang BSP, Lye DCB, Chow ALP, Ong RT-H, Ng OT, Perez F, Endimiani A, Ray A, Decker B, Wallace C, Hujer K, Ecker D, Adams M, Toltzis P, Dul M, Windau A, Bajaksouzian S, Jacobs M, Salata R, Bonomo R, Hsu L, TAN T, et al.: **Local transmission and global dissemination of New Delhi Metallo-Beta-Lactamase (NDM): a whole genome analysis**. *BMC Genomics* 2016, **17**:452.

83. Perez F, Endimiani A, Ray AJ, Decker BK, Wallace CJ, Hujer KM, Ecker DJ, Adams MD, Toltzis P, Dul MJ, Windau A, Bajaksouzian S, Jacobs MR, Salata RA, Bonomo RA: **Carbapenem-resistant Acinetobacter baumannii and Klebsiella pneumoniae across a hospital system: impact of post-acute care facilities on dissemination**. *J Antimicrob Chemother* 2010, **65**:1807–1818.

84. Hsu LY, Tan TY, Jureen R, Koh TH, Krishnan P, Lin RTP, Tee NWS, Tambyah PA: **Antimicrobial drug resistance in Singapore hospitals**. *Emerg Infect Dis* 2007, **13**:1944–1947.

85. Xiao Y-H, Giske CG, Wei Z-Q, Shen P, Heddini A, Li L-J: **Epidemiology and characteristics of antimicrobial resistance in China.** *Drug Resist Updat* 2011, **14**:236–50.

86. van Duijn PJ, Dautzenberg MJD, Oostdijk EAN: **Recent trends in antibiotic resistance in European ICUs.** *Curr Opin Crit Care* 2011, **17**:658–665.

87. Rhomberg PR, Jones RN: **Summary trends for the Meropenem Yearly Susceptibility Test Information Collection Program: a 10-year experience in the United States (1999-2008)**. *Diagn Microbiol Infect Dis* 2009, **65**:414–426.

88. Prabaker K, Weinstein RA: **Trends in antimicrobial resistance in intensive care units in the United States.** *Curr Opin Crit Care* 2011, **17**:472–9.

89. Schwaber MJ, Klarfeld-Lidji S, Navon-Venezia S, Schwartz D, Leavitt A, Carmeli Y: **Predictors of carbapenem-resistant Klebsiella pneumoniae acquisition anions hospitalized adults and effect of acquisition on mortality**. *Antimicrob Agents Chemother* 2008, **52**:1028–1033.

90. Bratu S, Landman D, Haag R, Recco R, Eramo A, Alam M, Quale J: **Rapid Spread of Carbapenem-Resistant Klebsiella pneumoniae in New York City**. *Arch Intern Med* 2005, **165**:1430.

91. Pitout JDD, Laupland KB: **Extended-spectrum beta-lactamase-producing Enterobacteriaceae: an emerging public-health concern.** *Lancet Infect Dis* 2008, **8**:159–66.

92. Yong D, Toleman MA, Giske CG, Cho HS, Sundman K, Lee K, Walsh TR: **Characterization of a new metallo-β-lactamase gene, *bla*(NDM-1), and a novel erythromycin esterase gene carried on a unique genetic structure in *Klebsiella pneumoniae* sequence type 14 from India.** *Antimicrob Agents Chemother* 2009, **53**:5046–5054.

93. Jones LS, Toleman M a, Weeks JL, Howe R a, Walsh TR, Kumarasamy KK: **Plasmid carriage of bla NDM-1 in clinical Acinetobacter baumannii isolates from India.** *Antimicrob Agents Chemother* 2014, **58**:4211–3.

94. Chen Y-T, Lin A-C, Siu LK, Koh TH: **Sequence of closely related plasmids encoding bla(NDM-1) in two unrelated Klebsiella pneumoniae isolates in Singapore.** *PLoS One* 2012, **7**:e48737.

95. Kumarasamy KK, Toleman M a., Walsh TR, Bagaria J, Butt F, Balakrishnan R, Chaudhary U, Doumith M, Giske CG, Irfan S, Krishnan P, Kumar A V., Maharjan S, Mushtaq S, Noorie T, Paterson DL, Pearson A, Perry C, Pike R, Rao B, Ray U, Sarma JB, Sharma M, Sheridan E, Thirunarayan M a., Turton J, Upadhyay S, Warner M, Welfare W, Livermore DM, et al.: **Emergence of a new antibiotic resistance mechanism in India, Pakistan, and the UK: A molecular, biological, and epidemiological study**. *Lancet Infect Dis* 2010, **10**:597–602.

96. Paterson DL, Wailan AM: **The spread and acquisition of NDM-1: a multifactorial problem.** *Expert Rev Anti Infect Ther* 2014, **12**:91–115.

97. Grad YH, Kirkcaldy RD, Trees D, Dordel J, Harris SR, Goldstein E, Weinstock H, Parkhill J, Hanage WP, Bentley S, Lipsitch M: **Genomic epidemiology of Neisseria gonorrhoeae with reduced susceptibility to cefixime in the USA: a retrospective observational study.** *Lancet Infect Dis* 2014, **14**:220–6.

98. Croucher NJ, Harris SR, Barquist L, Parkhill J, Bentley SD: **A high-resolution view of genome-wide pneumococcal transformation.** *PLoS Pathog* 2012, **8**:e1002745.

99. Harris SR, Clarke I, Seth-Smith HMB, Solomon AW, Cutcliffe LT, Marsh P, Skilton RJ, Holland MJ, Mabey D, Peeling RW, Lewis D a., Spratt BG, Unemo M, Persson K, Bjartling C, Brunham R, Vries D, Henry JC, Morré S a., Speksnijder A, Bébéar CM, Clerc M, de Barbeyrac B, Parkhill J, Thomson NR: **Whole-genome analysis of diverse Chlamydia trachomatis strains identifies phylogenetic relationships masked by current clinical typing**. *Nat Genet* 2012, **44**(February):413–419.

100. Harris SR, Feil EJ, Holden MTG, Quail M a, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay J a, Edgeworth JD, de Lencastre H, Parkhill J, Peacock SJ, Bentley SD: **Evolution of MRSA during hospital transmission and intercontinental spread.** *Science* 2010, **327**:469–474.

101. Poirel L, Dortet L, Bernabeu S, Nordmann P: **Genetic Features of blaNDM-1-Positive Enterobacteriaceae**. *Antimicrob Agents Chemother* 2011, **55**:5403–5407.

102. Carattoli A: **Resistance plasmid families in Enterobacteriaceae.** *Antimicrob Agents Chemother* 2009, **53**:2227–38.

103. Cambray G, Guerout A-M, Mazel D: **Integrons.** *Annu Rev Genet* 2010, **44**:141–166.

104. Schultsz C, Geerlings S: **Plasmid-mediated resistance in Enterobacteriaceae: changing landscape and implications for therapy.** *Drugs* 2012, **72**:1–16.

105. Xia E, Khong X, Marimuthu K, Xu W, Ong RT, Tan L, Krishnan U: **Draft Genome Sequence of a Multidrug-Resistant New Delhi Metallo- in Singapore**. *Genome Announc* 2013, **1**:1–2.

106. National Center for Emerging and Zoonotic Infectious Diseases: **CRE Toolkit - Guidance for Control of Carbapenem-resistant Enterobacteriaceae (CRE)**. 2012.

107. Larsen M V, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Pontén TS, Ussery DW, Aarestrup FM, Lund O: **Multilocus Sequence Typing of Total Genome Sequenced Bacteria.** *J Clin Microbiol* 2012, **50**:1355–61.

108. Johnson TJ, Nolan LK: **Plasmid replicon typing.** *Methods in molecular biology (Clifton, N.J.)* 2009:27–35.

109. Carver TJ, Rutherford KM, Berriman M, Rajandream M-A, Barrell BG, Parkhill J: **ACT: the Artemis Comparison Tool.** *Bioinformatics* 2005, **21**:3422–3.

110. Teo JWP, Tan P, La M-V, Krishnan P, Tee N, KOH TH, Deepak RN, TAN TY, Jureen R, Lin RTP: **Surveillance trends of carbapenem-resistant Enterobacteriaceae from Singapore, 2010–2013**. *J Glob Antimicrob Resist* 2014, **2**:99–102.

111. Bryant J, Chewapreecha C, Bentley SD: **Developing insights into the mechanisms of evolution of bacterial pathogens from whole-genome sequences**. *Future Microbiol* 2012, **7**:1283–1296.

112. Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, McGee L, von Gottberg A, Song JH, Ko KS, Pichon B, Baker S, Parry CM, Lambertsen LM, Shahinas D, Pillai DR, Mitchell TJ, Dougan G, Tomasz A, Klugman KP, Parkhill J, Hanage WP, Bentley SD: **Rapid pneumococcal evolution in response to clinical interventions.** *Science* 2011, **331**:430–4.

113. Stoesser N, Giess A, Batty EM, Sheppard AE, Walker AS, Wilson DJ, Didelot X, Bashir A, Sebra R, Kasarskis A, Sthapit B, Shakya M, Kelly D, Pollard AJ, Peto TEA, Crook DW, Donnelly P, Thorson S, Amatya P, Joshi S: **Genome sequencing of an extended series of NDM-producing Klebsiella pneumoniae isolates from neonatal infections in a Nepali hospital characterizes the extent of community- versus hospital-associated transmission in an endemic setting.** *Antimicrob Agents Chemother* 2014, **58**:7347–57.

114. Conlan S, Thomas PJ, Deming C, Park M, Lau AF, Dekker JP, Snitkin ES, Clark T a., Luong K, Song Y, Tsai Y-C, Boitano M, Dayal J, Brooks SY, Schmidt B, Young AC, Thomas JW, Bouffard GG, Blakesley RW, Mullikin JC, Korlach J, Henderson DK, Frank KM, Palmore TN, Segre J a.: **Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae.** *Sci Transl Med* 2014, **6**:254ra126.

115. Snitkin ES, Zelazny a. M, Thomas PJ, Stock F, Henderson DK, Palmore TN, Segre J a.: **Tracking a Hospital Outbreak of Carbapenem-Resistant Klebsiella pneumoniae with Whole-Genome Sequencing**. *Sci Transl Med* 2012, **4**:148ra116–148ra116.

116. Netikul T, Sidjabat HE, Paterson DL, Kamolvit W, Tantisiriwat W, Steen JA, Kiratisin P: **Characterization of an IncN2-type blaNDM-$_1$-carrying plasmid in Escherichia coli ST131 and Klebsiella pneumoniae ST11 and ST15 isolates in Thailand.** *J Antimicrob Chemother* 2014, **69**:3161–3.

117. Magoc T, Pabinger S, Canzar S, Liu X, Su Q, Puiu D, Tallon LJ, Salzberg SL: **GAGE-B: An evaluation of genome assemblers for bacterial organisms**. *Bioinformatics* 2013, **29**:1718–1725.

118. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes.** *Genome Biol* 2004, **5**:R12.

119. Ohno S: **Evolution by Gene Duplication**. *(1970)* 1970.

120. Bergthorsson U, Andersson DI, Roth JR: **Ohno's dilemma: evolution of new genes under continuous selection.** *Proc Natl Acad Sci U S A* 2007, **104**:17004–9.

121. Walsh JB: **How often do duplicated genes evolve new functions?** *Genetics* 1995, **139**:421–428.

122. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J: **Preservation of duplicate genes by complementary, degenerative mutations**. *Genetics* 1999, **151**:1531–1545.

123. Innan H, Kondrashov F: **The evolution of gene duplications: classifying and distinguishing between models.** *Nat Rev Genet* 2010, **11**:97–108.

124. Lynch M, Force A: **The probability of duplicate gene preservation by subfunctionalization**. *Genetics* 2000, **154**:459–473.

125. Serres MH, Kerr ARW, McCormack TJ, Riley M: **Evolution by leaps: gene duplication in bacteria.** *Biol Direct* 2009, **4**:46.

126. Sanchez-Perez G, Mira A, Nyiro G, Pašić L, Rodriguez-Valera F: **Adapting to environmental changes using specialized paralogs**. *Trends in Genetics* 2008:154–158.

127. Kondrashov FA, Rogozin IB, Wolf YI, Koonin E V: **Selection in the evolution of gene duplications.** *Genome Biol* 2002, **3**:RESEARCH0008.1–0008.9.

128. N äsvall J, Sun L, Roth JR, Andersson DI: **Real-time evolution of new genes by innovation, amplification, and divergence.** *Science* 2012, **338**:384–7.

129. Hobbs M, Seiler A: **Microevolution within a clonal population of pathogenic bacteria: recombination, gene duplication and horizontal genetic exchange in the opa gene family of Neisseria**. *Mol ...* 1994, **12**:171–180.

130. Falush D: **Toward the use of genomics to study microevolutionary change in bacteria.** *PLoS Genet* 2009, **5**:e1000627.

131. Reams AB, Kofoid E, Savageau M, Roth JR: **Duplication frequency in a population of Salmonella enterica rapidly approaches steady state with or without recombination.** *Genetics* 2010, **184**:1077–94.

132. Schirmer T, Keller T, Wang Y, Rosenbusch J: **Structural basis for sugar translocation through maltoporin channels at 3.1 A resolution**. *Science (80-)* 1995, **1**:1–3.

133. Poh W-T, Xia E, Chin-Inmanu K, Wong L-P, Cheng AY, Malasit P, Suriyaphol P, Teo Y-Y, Ong RT-H: **Viral quasispecies inference from 454 pyrosequencing.** *BMC*

Bioinformatics 2013, **14**:355.

134. Huse SM, Welch DBM: **Accuracy and Quality of Massively Parallel DNA Pyrosequencing**. In *Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches*; 2011:149–155.

135. Eriksson N, Pachter L, Mitsuya Y, Rhee SY, Wang C, Gharizadeh B, Ronaghi M, Shafer RW, Beerenwinkel N: **Viral population estimation using pyrosequencing**. *PLoS Comput Biol* 2008, **4**.

136. Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N: **ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data.** *BMC Bioinformatics* 2011, **12**:119.

137. Larkin MA, Blackshields G, Brown NP, Chenna R, Mcgettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG: **Clustal W and Clustal X version 2.0**. *Bioinformatics* 2007, **23**:2947–2948.

138. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S: **MEGA6: Molecular evolutionary genetics analysis version 6.0**. *Mol Biol Evol* 2013, **30**:2725–2729.

139. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4**:406–425.

140. DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ: **A framework for variation discovery and genotyping using next-generation DNA sequencing data.** *Nat Genet* 2011, **43**:491–498.

141. Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N: **LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets**. *Nucleic Acids Res* 2012, **40**:11189–11201.

142. Zhang Y: **I-TASSER server for protein 3D structure prediction.** *BMC Bioinformatics* 2008, **9**:40.

143. Roy A, Kucukural A, Zhang Y: **I-TASSER: a unified platform for automated protein structure and function prediction.** *Nat Protoc* 2010, **5**:725–738.

144. Roy A, Yang J, Zhang Y: **COFACTOR: An accurate comparative algorithm for structure-based protein function annotation**. *Nucleic Acids Res* 2012, **40**.

145. Riottot MM, Fournier JM, Jouin H: **Direct evidence for the involvement of capsular polysaccharide in the immunoprotective activity of Klebsiella pneumoniae ribosomal preparations**. *Infect Immun* 1981, **31**:71–77.

146. Pegueroles C, Laurie S, Alb àMM: **Accelerated evolution after gene duplication: a time-dependent process affecting just one copy.** *Mol Biol Evol* 2013, **30**:1830–42.

147. Achtman M: **MicroReview Clonal spread of serogroup A meningococci : a paradigm for the analysis of microevolution in bacteria**. 1994, **11**:15–22.

148. Lin X, Yang M, Li H, Wang C, Peng X-X: **Decreased expression of LamB and Odp1 complex is crucial for antibiotic resistance in Escherichia coli.** *J Proteomics* 2014, **98**:244–53.

149. Gayet S, Chollet R, Molle G: **Modification of outer membrane protein profile and evidence suggesting an active drug pump in Enterobacter aerogenes clinical strains**. *Antimicrob agents …* 2003, **47**:1555–1559.

150. Astrovskaya I, Tork B, Mangul S, Westbrooks K, Mandoiu I, Balfe P, Zelikovsky A: **Inferring viral quasispecies spectra from 454 pyrosequencing reads**. *BMCBioinformatics* 2011, **12 Suppl 6**(1471-2105 (Electronic)):S1.

151. Prosperi MCF, Salemi M: **QuRe: Software for viral quasispecies reconstruction from next-generation sequencing data**. *Bioinformatics* 2012, **28**:132–133.

152. Xia E, Teo Y-Y, Ong RT-H: **SpoTyping: fast and accurate in silico Mycobacterium**

spoligotyping from sequence reads. *Genome Med* 2016, **8**:19.

153. World Health Organization: **Tuberculosis Fact sheet N °104**. 2015:1–5.

154. van der Zanden AG, Hoentjen AH, Heilmann FG, Weltevreden EF, Schouls LM, van Embden JD: **Simultaneous detection and strain differentiation of Mycobacterium tuberculosis complex in paraffin wax embedded tissues and in stained microscopic preparations.** *Mol Pathol* 1998, **51**:209–214.

155. Hermans PWM, Van Soolingen D, Bik EM, De Haas PEW, Dale JW, Van Embden JDA: **Insertion element IS987 from Mycobacterium bovis BCG is located in a hot-spot integration region for insertion elements in Mycobacterium tuberculosis complex strains**. *Infect Immun* 1991, **59**:2695–2705.

156. Dale JW, Brittain D, Cataldi AA, Cousins D, Crawford JT, Driscoll J, Heersma H, Lillebaek T, Quitugua T, Rastogi N, Skuce RA, Sola C, Van Soolingen D, Vincent V: **Spacer oligonucleotide typing of bacteria of the Mycobacterium tuberculosis complex: Recommendations for standardised nomenclature**. *International Journal of Tuberculosis and Lung Disease* 2001:216–219.

157. Song EJ, Jeong HJ, Lee SM, Kim CM, Song ES, Park YK, Bai G-H, Lee EY, Chang CL: **A DNA chip-based spoligotyping method for the strain identification of Mycobacterium tuberculosis isolates.** *J Microbiol Methods* 2007, **68**:430–3.

158. Ruettger A, Nieter J, Skrypnyk A, Engelmann I, Ziegler A, Moser I, Monecke S, Ehricht R, Sachse K: **Rapid spoligotyping of Mycobacterium tuberculosis complex bacteria by use of a microarray system with automatic data processing and assignment.** *J Clin Microbiol* 2012, **50**:2492–5.

159. Bespyatykh JA, Zimenkov D V, Shitikov EA, Kulagina E V, Lapa SA, Gryadunov DA, Ilina EN, Govorun VM: **Spoligotyping of Mycobacterium tuberculosis complex isolates using hydrogel oligonucleotide microarrays.** *Infect Genet Evol* 2014, **26**:41–6.

160. Gomgnimbou MK, Abadia E, Zhang J, Refrégier G, Panaiotov S, Bachiyska E, Sola C: **"Spoligoriftyping," a dual-priming-oligonucleotide-based direct-hybridization assay for tuberculosis control with a multianalyte microbead-based hybridization system.** *J Clin Microbiol* 2012, **50**:3172–9.

161. Gomgnimbou MK, Hernández-Neuta I, Panaiotov S, Bachiyska E, Palomino JC, Martin A, del Portillo P, Refregier G, Sola C: **Tuberculosis-spoligo-rifampin-isoniazid typing: an all-in-one assay technique for surveillance and control of multidrug-resistant tuberculosis on Luminex devices.** *J Clin Microbiol* 2013, **51**:3527–34.

162. Honisch C, Mosko M, Arnold C, Gharbia SE, Diel R, Niemann S: **Replacing reverse line blot hybridization spoligotyping of the Mycobacterium tuberculosis complex.** *J Clin Microbiol* 2010, **48**:1520–6.

163. Shitikov E, Ilina E, Chernousova L, Borovskaya A, Rukin I, Afanas'ev M, Smirnova T, Vorobyeva A, Larionova E, Andreevskaya S, Kostrzewa M, Govorun V: **Mass spectrometry based methods for the discrimination and typing of mycobacteria.** *Infect Genet Evol* 2012, **12**:838–45.

164. Price EP, Smith H, Huygens F, Giffard PM: **High-resolution DNA melt curve analysis of the clustered, regularly interspaced short-palindromic-repeat locus of Campylobacter jejuni.** *Appl Environ Microbiol* 2007, **73**:3431–6.

165. Schouls LM, Reulen S, Duim B, Wagenaar JA, Willems RJL, Dingle KE, Colles FM, Van Embden JDA: **Comparative Genotyping of Campylobacter jejuni by Amplified Fragment Length Polymorphism, Multilocus Sequence Typing, and Short Repeat Sequencing: Strain Diversity, Host Range, and Recombination**. *J Clin Microbiol* 2003, **41**:15–26.

166. Ginevra C, Jacotin N, Diancourt L, Guigon G, Arquilliere R, Meugnier H, Descours G, Vandenesch F, Etienne J, Lina G, Caro V, Jarraud S: **Legionella pneumophila sequence type 1/Paris pulsotype subtyping by spoligotyping.** *J Clin Microbiol* 2012, **50**:696–701.

167. Fabre L, Zhang J, Guigon G, Le Hello S, Guibert V, Accou-Demartin M, de Romans S,

Lim C, Roux C, Passet V, Diancourt L, Guibourdenche M, Issenhuth-Jeanjean S, Achtman M, Brisse S, Sola C, Weill F-X: **CRISPR typing and subtyping for improved laboratory surveillance of Salmonella infections.** *PLoS One* 2012, **7**:e36995.

168. Barnes PF, Cave MD: **Molecular epidemiology of tuberculosis.** *N Engl J Med* 2003, **349**:1149–56.

169. Coll F, Mallard K, Preston MD, Bentley S, Parkhill J, McNerney R, Martin N, Clark TG: **SpolPred: rapid and accurate prediction of Mycobacterium tuberculosis spoligotypes from short genomic sequences.** *Bioinformatics (Oxford, England)* 2012:2991–3.

170. Demay C, Liens B, Burguière T, Hill V, Couvin D, Millet J, Mokrousov I, Sola C, Zozio T, Rastogi N: **SITVITWEB - A publicly available international multimarker database for studying Mycobacterium tuberculosis genetic diversity and molecular epidemiology**. *Infect Genet Evol* 2012, **12**:755–766.

171. Zhang H, Li D, Zhao L, Fleming J, Lin N, Wang T, Liu Z, Li C, Galwey N, Deng J, Zhou Y, Zhu Y, Gao Y, Wang T, Wang S, Huang Y, Wang M, Zhong Q, Zhou L, Chen T, Zhou J, Yang R, Zhu G, Hang H, Zhang J, Li F, Wan K, Wang J, Zhang X-E, Bi L: **Genome sequencing of 161 Mycobacterium tuberculosis isolates from China identifies genes and intergenic regions associated with drug resistance.** *Nat Genet* 2013, **45**:1255–60.

172. Zerbino DR, Birney E: **Velvet: Algorithms for de novo short read assembly using de Bruijn graphs**. *Genome Res* 2008, **18**:821–829.

173. Witney A a., Gould K a., Arnold A, Coleman D, Delgado R, Dhillon J, Pond M, Pope CF, Planche TD, Stoker NG, Cosgrove C a., Butcher PD, Harrison TS, Hinds J: **Clinical application of whole genome sequencing to inform treatment for multi-drug resistant tuberculosis cases**. *J Clin Microbiol* 2015(February):JCM.02993–14.

174. Roetzer A, Schuback S, Diel R, Gasau F, Ubben T, Di Nauta A, Richter E, Rüsch-Gerdes S, Niemann S: **Evaluation of Mycobacterium tuberculosis typing methods in a 4-year study in Schleswig-Holstein, Northern Germany**. *J Clin Microbiol* 2011, **49**:4173–4178.

175. Oelemann MC, Diel R, Vatin V, Haas W, Rüsch-Gerdes S, Locht C, Niemann S, Supply P: **Assessment of an optimized mycobacterial interspersed repetitive-unit- variable-number tandem-repeat typing system combined with spoligotyping for population-based molecular epidemiology studies of tuberculosis**. *J Clin Microbiol* 2007, **45**:691–697.

176. Allix-Béguec C, Fauville-Dufaux M, Supply P: **Three-year population-based evaluation of standardized mycobacterial interspersed repetitive-unit-variable-number tandem-repeat typing of Mycobacterium tuberculosis.** *J Clin Microbiol* 2008, **46**:1398–406.

177. Iwai H, Kato-miyazawa M, Kirikae T, Miyoshi-akiyama T: **CASTB (the comprehensive analysis server for the Mycobacterium tuberculosis complex): A publicly accessible web server for epidemiological analyses , drug-resistance prediction and phylogenetic comparison of clinical isolates**. *Tuberculosis* 2015:9–10.

178. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, Phillippy AM: **Hybrid error correction and de novo assembly of single-molecule sequencing reads**. *Nature Biotechnology* 2012:693–700.

179. Chin C-S, Alexander DH, Marks P, Klammer A a, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J: **Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data.** *Nat Methods* 2013, **10**:563–9.