Use of Next Generation Sequencing to Investigate the Genomics of Bacterial Pathogens.

Submitted by James William Harrison to the University of Exeter

as a thesis for the degree of Doctor of Philosophy in Biological Sciences

July 2019

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

....

James William Harrison

Abstract

This thesis explores the use of next generation sequencing as a tool to investigate the genomics of bacterial pathogens of plants and humans.

Firstly, second-generation sequencing was applied to the evolution of distantly related bacterial species that have converged on common host plants (*Xanthomonas* bacteria on sugarcane and common-bean plants). This revealed evidence of recent horizontal gene transfer between *X. phaseoli* pv. *phaseoli* and *X. citri* pv. *fuscans* and between *X. axonopodis* pv. *vasculorum* and *X. vasicola*. distantly related sugarcane pathogens. Furthermore, we discovered that strains isolated from lablab bean (a close relative of common bean) form a previously unknown third distinct clade (and perhaps pathovar) and whole-genome comparisons suggested horizontal gene transfer played an important role in the evolution of host specificity in xanthomonad pathogens.

Next, second-generation sequencing was used to rapidly gain insight into novel emerging bacterial pathogens, namely unusually virulent Asian strains of the human pathogen *Campylobacter jejuni* and a xanthomonad causing unusual symptoms on common bean in African country of Rwanda. A type six secretion system was shown to be associated with a more serious form of campylobacteriosis and a molecular marker for an intact type six secretion system was identified. This was shown to be more prevalent in strains isolated from Asia than strains isolated in the UK, a finding which has serious implications for chicken import. Further to this the genome sequence of a newly emerging *Xanthomonas* bean pathogen isolated from a recent outbreak in Rwanda is presented. Analysis of the Rwandan *Xanthomonas* genome shows it represents the first sequenced isolate in a novel species level clade, which was subsequently named as *Xanthomonas cannabis* and is genetically distinct from previously known bean pathogens.

Lastly, the performance of the third-generation sequencing platform Oxford Nanopore MinION was assessed which will prove to be an exciting resource to perform bacterial genomic studies in the future. In summary, this work exemplifies the value of sequencing-based approaches for rapidly and cheaply gaining insights into evolution of bacterial pathogens.

Acknowledgements

I would like to acknowledge my supervisor Dr. David Studholme for giving me the opportunity to work on a truly fantastic project. His inspiration and unwavering support helped make the process an absolute joy.

I would also like to thank Dr. Thomas Laver for his support throughout and his keen editorial skill.

Also of course my children, for providing the inspiration for beginning the project and for keeping going when times were difficult.

Contents

List of figures and tables

Authors declaration

This thesis is concerned with the bioinformatics analysis of data largely generated by collaborators on the various project discussed here in. The author planned and carried out the bioinformatics analysis detailed here and unless otherwise stated other results mentioned were generated by collaborators and are mentioned in this document in order to support the conclusions of the project.

The work undertaken has been published in papers which are included where relevant in the manuscript.

Chapter 1:

Introduction

Introduction

The scientific field of molecular biology was revolutionised in the latter part of the last century with the advent of DNA sequencing. This revolutionary technology has brought about a rapid and still accelerating rate of discovery that shows no signs of slowing. In fact, in the field of modern molecular biology and ‘omics’ the limiting factor is no longer the generation of sequencing data or the constraints of wet bench laboratory work, but in data analysis and computational power. These technological advances have allowed the bloom of genomics research over the last few decades.

Sequencing

The development and refinement of DNA sequencing has opened up a wealth of possibilities. Researchers are now able to answer questions that were, until the advent of low cost next generation sequencing (NGS) technologies, at best difficult and more likely impossible. Recent studies have included data from thousands of bacterial genomes (e.g. [1,2]) and the publishing of full genome sequences happens on a regular basis [3–6]. In order to store and disseminate the huge amounts of data generated and published by the scientific community worldwide, massive online repositories have developed. These include the GenBank database hosted by the National Centre for Biotechnology Information (NCBI) [7], the European Nucleotide Archive hosted at the European Bioinformatics Institute (EBI) and many others. The development of specialist repositories for organism- or subject-specific genomic information have made available a huge amount of data which is at the fingertips (quite literally) of any researcher with internet access.

The first DNA sequence was announced by Frederick Sanger in 1975. Sanger, who died in 2013, later received two Nobel prizes for his work on DNA and protein sequencing which had changed the face of bioscience and biomedical research forever. Sanger continued to contribute to the field publishing his final research paper in 1982 containing the first large genome sequence at 48,502 b.p. of genomic DNA: that of the now famous bacteriophage lambda [8]. In 1986, the first automated DNA sequencing technology was pioneered by researchers at Caltech [9] and released commercially by a group comprising of several factions including Applied Biosciences (ABI) and The European Molecular Biology Laboratory (EMBL).

The history of Sanger sequencing is littered with seminal discoveries. Beginning with the first phage sequence [8] there have since been many exciting discoveries such as the sequence of the 320000 bp chromosome III of *Saccharomyces cerevisiae* in 1992 [10]. The first completed bacterial genome sequence was that of *Haemophilus influenzae,* published in 1995 [11] beginning a trend which is still gaining momentum. Soon completed genomes began to be published regularly; notable examples include the sequences for model organisms such as the bacterium *Escherichia coli* in 1997 [12] and the model multi-cellular eukaryote *Caenorhabditis elegans* in 1998 [13]. This explosion in the volume of sequence data generated necessitated the creation of the modern fields of bioinformatics and computational biology. In the early stages of DNA sequencing sequences deposited in repositories such as NCBI were largely short partial gene sequences. This was until the early 1990s when whole genome sequence data began to be generated. The amount of genome sequence data held in public data bases has grown at an astonishing rate, with (as of June 2016) 71296 prokaryotic genomes, 3275 eukaryotic genomes and 5576 viral genomes available at various stages of completion. As time and technology progresses the volume of sequence data increases (see Figure 1).

Not only did the advent of DNA sequencing facilitate the sequencing of genomes of single organisms, but many novel techniques were made possible. The development of molecular microbial profiling made a significant, far reaching and long-term contribution to fields such as environmental ecology and clinical biology. This method was applied by Giovannoni *et al.* in their 1990 study using primers

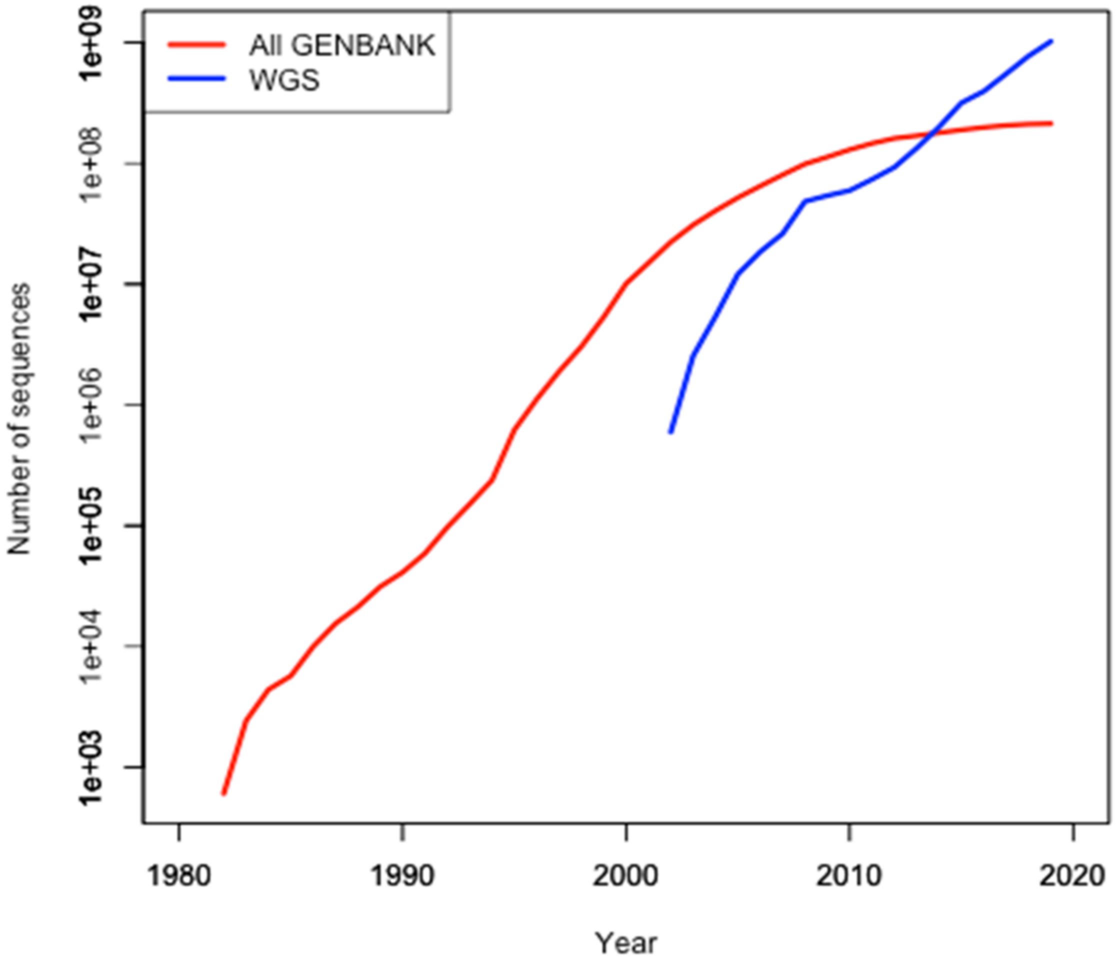


Figure 1: Graph of number of sequences held in the NCBI genbank repository. The lines represent sequences in GenBank (red) and whole genome sequences (blue). The y axis is a log scale. Data obtained from [124].

targeting the 16S rDNA gene, ubiquitously conserved among prokaryotes, to interrogate marine samples from the Sargasso Sea [14]. They discovered the SAR11 clade of bacteria (now known as *Pelagibacterales*) which have since been suggested to be the most numerous bacteria in nature [15].

The Sanger method was refined and added-to for the following decades and this technology is still used today as a companion to more modern approaches. It offers a cost effective method of investigating specific questions targeted at particular regions of sequence. Often interesting or potentially controversial findings identified using less accurate NGS methods are confirmed using Sanger sequencing e.g.[16].

In 2001 possibly the most revolutionary advance in the field of molecular biology was announced, the first two sequences of the human genome [17,18]. Actually, completed by two competing teams of researchers one led by Craig Ventner and the other by Eric Lander, this momentous achievement heralded a new age of genomics.

DNA sequencing was again to take a significant leap forward in 2006 with the release of the GS20 by 454 life sciences [19]. This was the beginning of a new wave of sequencing technologies known as NGS. 454 life sciences pioneered pyrosequencing a technology still used today (e.g. [20]) although it has been largely superseded by more modern methods. Pyrosequencing facilitated the generation of previously undreamt of amounts of sequence data and opened the way for other NGS technologies. The characteristic benefits of this new methodology, particularly the high throughput massively parallel nature, democratized sequencing and allowed individual research groups access to resources previously only open to a few large sequencing centres worldwide [21]. 454 sequenced large amounts of short (100 to 700 base pairs [22]) lengths of DNA which could be processed bioinformatically in order to answer diverse biological questions.

Since Roche's ground-breaking method was released there have been several other competing technologies which have progressed the field of sequencing further. Arguably, the most successful NGS technology to date [23] is Illumina/Solexa – also released in 2006. The Illumina technology is again based around sequencing short lengths of DNA. Template DNA is prepared by fragmenting the input material using various methods such as restriction enzymes or sonication. These fragments are then size-selected to enrich for a specific size (typically around 600 – 800 bp). Adapters are then ligated to the fragmented strands of DNA. The sequencing actually takes place on a flowcell: the Illumina flowcell has millions of primers attached to its surface, which bind the corresponding adaptor on the input material. This allows the accurate positioning of each strand of DNA on the flowcell. *In situ* PCR is then carried out creating clusters of copies of each template strand. The sequencing is actually carried out by washing the flowcell with its clusters of template DNA with many rounds of fluorescently tagged nucleotides, taking a high-resolution image and using proprietary image analysis software to interpret the results at each nucleotide, generating a usable sequence. To generate paired end reads the molecules are inverted on the flowcell and the process is repeated, generating the pairs of reads with a 'known' gap [24] . The benefit Illumina displayed over its competitors was the volume of data generated and the cost per base pair of this sequence data [25]. At its inception, the Illumina GAII was able to produce 1GB of sequence data with reads of 32 bp in length, already a massive leap from previous instruments. However, soon the length of the reads generated and the amount of sequence data generated from a single run grew and by 2014 the HI-SEQ 2000 was routinely able to produce 5 billion reads of 150 pairs from a single flowcell (see table 1 for detail of Illumina technologies). A further benefit afforded by Illumina sequencing was the generation of different types of reads. Paired end and mate pair sequencing involved the generation of sequence reads of normal Illumina length from two ends of the same molecular fragment of DNA with a known distance in between [26]. Typically, for paired-end sequencing the fragment size is around 600 base pairs and for mate pair reads the fragment size could be much larger. These more complex read types allowed much more information to be gained from the sequence data. The known fragment size and therefore the gap between the reads generated facilitated the placing of reads in a genomic context. This improved the accuracy and effectiveness of assembling sequence reads into longer contiguous sequences better representing the template from which they were generated [26]. These improvements, along with advances in library preparation protocols and further improvements to the parallelisation and throughput of the system have made multiplexing many samples on one flowcell easy and cost effective, bringing the cost and availability of huge and varied sequencing projects into the realms of most research groups [26].

	MiniSeq System	MiSeq Series	NextSeq Series	HiSeq Series	HiSeq X Series*
Key Methods	Amplicon, targeted RNA, small RNA, and targeted gene panel sequencing.	Small genome, amplicon, and targeted gene panel sequencing.	Everyday exome, transcriptome, and targeted resequencing.	Production-scale genome, exome, transcriptome sequencing, and more.	Population- and production-scale whole-genome sequencing.
Maximum Output	7.5 Gb	15 Gb	120 Gb	1500 Gb	1800 Gb
Maximum Reads per Run	25 million	25 million†	400 million	5 billion	6 billion
Maximum Read Length	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp
Run Time	4–24 hours	4–55 hours	12–30 hours	<1–3.5 days (HiSeq 3000/HiSeq 4000) 7 hours–6 days (HiSeq 2500)	<3 days
Benchtop Sequencer	Yes	Yes	Yes	No	No
System Versions	MiniSeq System for low-throughput targeted DNA and RNA sequencing	• MiSeq System for targeted and small genome sequencing • MiSeq FGx System for forensic genomics • MiSeqDx Systemfor molecular diagnostics	• NextSeq 500 System for everyday genomics • NextSeq 550 System for both sequencing and cytogenomic arrays	• HiSeq 3000/HiSeq 4000 Systems for production-scale genomics • HiSeq 2500 Systemsfor large-scale genomics	• HiSeq X Five Systemfor production-scale whole-genome sequencing • HiSeq X Ten Systemfor population-scale whole-genome sequencing

Table 1: table of the properties of Illumina sequencing technologies. Data obtained from [121].

There are problems inherent in Illumina sequencing technologies [27–30], which have been addressed to a greater or lesser extent since its introduction. There are the obvious issues with the use of short reads which Illumina technology shares with its competitors. Initially Illumina reads were very short, and consequentially, of limited utility for tasks such as *de novo* genome assembly, but still useful for other applications such as resequencing and RNA profiling. However, as read lengths increased, so did the uses that Illumina data was put to. By the time the HiSeq 2000 was released, it was capable of generating 5 billion paired reads of 150 b.p. the Illumina age had firmly taken hold. Illumina sequencing incurs an error rate of approximately 0.1% [31], which although it sounds high, is counterbalanced by the depth of coverage generated. Each base position in the template DNA is sequenced many times allowing post sequencing analysis to highlight these systematic errors and lessen or eradicate their effects. There are other errors that can cause issues; library preparation often includes PCR amplification that has its own error rates and biases. However, as the technology, wet bench preparation techniques and data analysis matured, these problems were overcome. Illumina has become the world leader in sequencing technologies and regularly releases new sequencers such as the NovaSeq 6000 and the NextSeq which supersede the HiSeq.

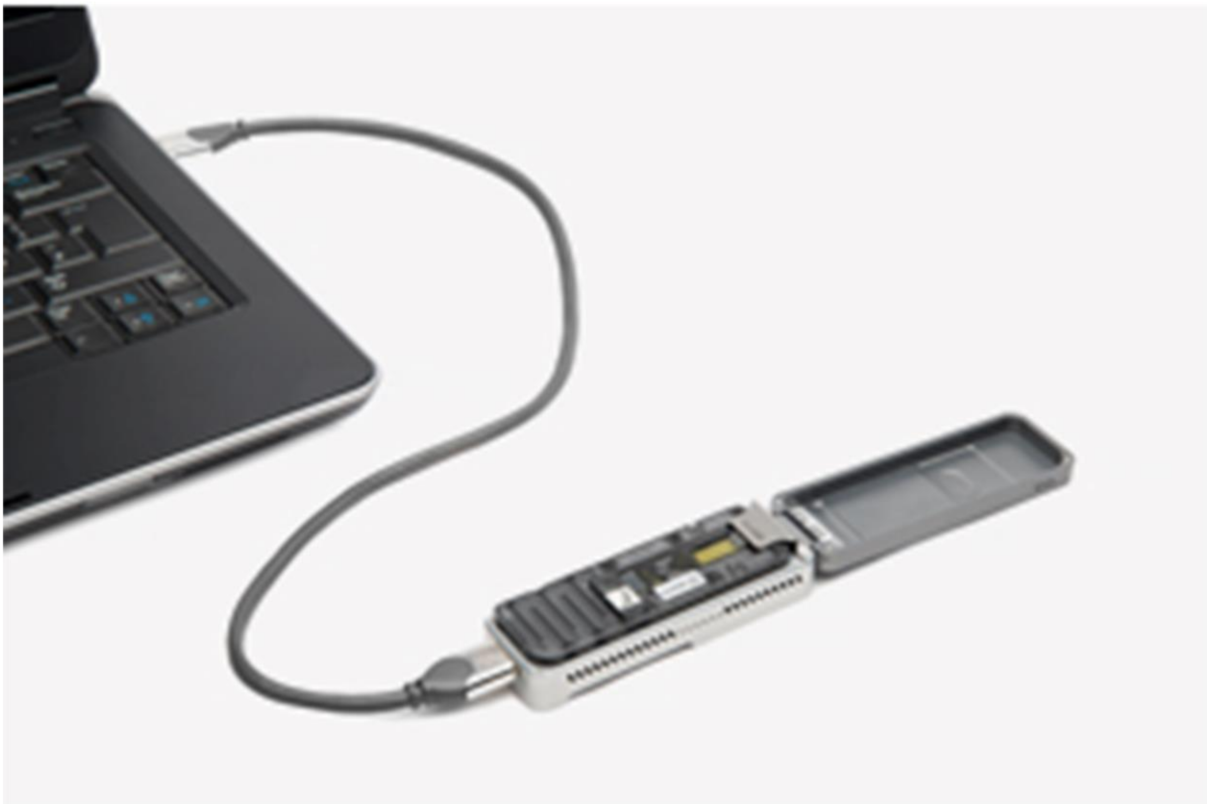
The Illumina sequencing technology is incredibly versatile and has many applications utilizing the particular characteristics of this short-read sequencer to the best effect[32]. Draft genome sequences are routinely generated, the short reads being assembled *in silico* to produce usable genomic sequence to inform studies on a wide range of diverse subjects. These subjects range from bacterial comparative genomics studies such as this one to producing the draft genome sequence of complex eukaryotic organisms such as the western lowland gorilla (*Gorilla gorilla*

gorilla) [33]. Illumina sequencing is routinely used for resequencing and alignment to reference genomes in order to identify small variations and polymorphisms between individuals or closely related isolates. This process can be used to elucidate evolutionary mechanisms and can identify genomic markers of phenotypic variation or uncover associations between genetic markers and human disease. It is even possible to detect epigenetic modifications in human DNA using Illumina sequencing and bisulphite treatment (e.g. [34]).

The development of sequencing methods certainly has not stopped with the first wave of NGS. Recently, a new wave of advances colloquially known as “third generation” sequencing have been released with the Pacific Biosciences Single Molecule Real Time technology (SMRT) probably the best known and is already installed in many sequencing centres [35]. This novel process is able to sequence long molecules of DNA a number of times, generating long accurate consensus reads, which can be used to assemble genomic DNA *de novo* or scaffold contigs generated by other sequence technologies. Pacific Biosciences claim that it will be possible to detect epigenetic modifications during the sequencing reaction, pushing forward that field of study. A notable competitor, producing excitement in the field is the Oxford Nanopore Technologies MinION - heralding the advent of the portable, real-time sequencer [36] (Figure 2). The MinION uses nanopore technology to sequence long single molecules of DNA. The DNA molecules pass through nanopores in a micro sheet and the electrical potential across the pore can be measured. The MinION measures the signal not from each single base but from 5- or 6-mers as they move thorough the nanopore each one of these 5- or 6-mers resulting in a different reading. The analysis of these signals allowing the identification of each base in the sequence [36]. This technology has the potential to generate very long reads of up to

100 kb and beyond [36,37]. These new developments are still very much in their infancy and are dealing with similar teething problems as NGS experienced in its early days with high error rates, library preparation challenges and cost still issues to be addressed [38]. However, their potential is enormous and with fourth generation sequencing already being discussed, the rate of advancement is only picking up pace [38]. The release of more efficient and less error prone technologies from companies such as Pacific Biosciences (Sequel II) and Oxford Nanopore (PromethION) are already beginning to produce results [38]. Due to the novel nature of these new approaches this project focuses mainly on data generated using NGS, namely the Illumina Hi-Seq with the addition of an assessment of the Oxford Nanopore Minion. Therefore from this point forward in the introduction discussion of techniques will relate to those data used in this project.

Figure 2: The Oxford Nanopore MinION.
Image courtesy of Nanoporetech.com [122].



Obviously, the generation of such vast amounts of sequence data creates its own challenges. Computational capacity has risen dramatically and necessarily so, the cost of sequencing has dropped dramatically (Figure 3 A + B) Figure 3A shows the costs per megabase of sequence data and Figure 3B shows the costs per human sized genome. A steep drop in cost can be observed in 2008 when NGS technologies began to replace traditional Sanger technology in sequencing centres. A line representing Moore's law is included for comparison. Technologies which keep up with Moore's law are viewed as successful, and as can be seen, sequencing has vastly outpaced it [39]. New computational approaches had to be developed to answer the questions posed by the amount and nature of the data generated. Fast efficient sequence alignment software such as BWA [40] and BOWTIE [41] were developed to align short reads generated from NGS to reference sequences. Genome assemblers such as VELVET [42] and SPAdes [43] use De Brujin graphs to efficiently produce reliable assemblies from short reads alone. The annotation of genomic sequence which was once a laborious time consuming process has now been automated with online services such as RAST [44] and the NCBI's Prokaryotic Genome Annotation Pipeline [45] making the accurate annotation of prokaryotic sequence fast and simple. Even eukaryotic sequences can now be annotated in an automated way with some degree of confidence with tools such as MAKER [46] leading the way.

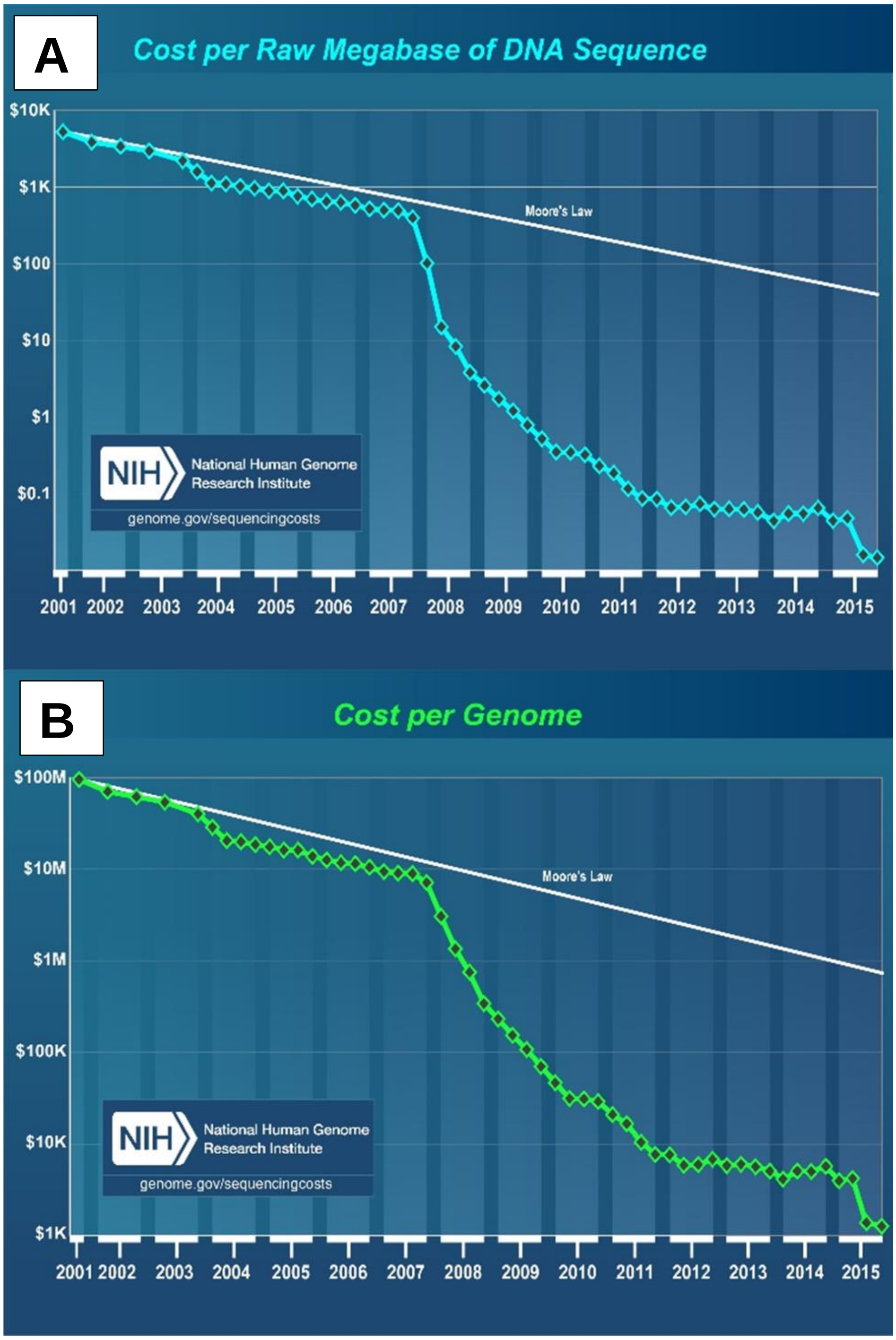


Figure 3A: Graph showing the reduction in cost of sequencing per megabase of sequence when compared to Moores law.
Figure 3B: Graph showing the reduction in cost of sequencing the human genome when compared to Moores law.
Courtesy of [39].

As mentioned previously, the generation of such vast amounts of NGS data has necessitated the expansion of the field of bioinformatics. The impact of this fast evolving field has been immense, with almost every bioscience field now influenced by its discoveries and methodologies, from clinical genetics to ecology. Studies from disparate fields of biology are now using some facet of DNA sequencing technology to answer questions specifically related their fields. Marine ecologists are undertaking vast sequencing efforts to attempt to increase the understanding of marine microbiomes using metagenomic analysis [47,48]. Clinical disease outbreaks are being studied in minute detail using the latest technologies to inform researchers and clinicians in the outbreak routes of these emerging diseases. Novel genetic traits can quickly and reliably be identified which make these novel strains more successful and therefore, worrying and the methods to control outbreaks can be quickly developed [49–51].

To facilitate these advances and discoveries there are some core methods which are used throughout the bioinformatics sphere.

Data analysis: Overcoming the bottleneck and introduction to methods used

The amount of data generated by NGS platforms is huge and without a sensible standardised way of presenting and storing this data the analysis process would be all but impossible. Fortunately, early in the development of this new technology the Wellcome Trust Sanger Institute created the fastq file format[52] (figure 4). This was an important development, as previous formats did not store the necessary information in one place. The fastq format stores each read as four separate lines with the quality information of each base sequenced stored as a separate line to enable the quality assessment of the sequencing from information held in one file.

Figure 4 . A single sequence read represented in FASTQ format

Unique header line often containing information on the sequencing run and the position on flowcells
Sequence line
Optional further information line always starting with a +
ASCII encoded quality line with a matching quality character for each base of sequence in line 2

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Typically, data generated from the Illumina platform is in the form of short reads of between 32 (for older technologies) and 300 b.p. in length. As previously mentioned they can come in the form of paired end or mate pair reads. The technological limitations of NGS technologies result in a drop of in the quality therefore the reliability of the sequence data as the read gets longer. Software such as FASTQC [53] can give excellent actionable metrics to assess the quality of NGS data and allow informed trimming and filtering by packages such as TRIMMOMATIC [54] and FASTQ-MCF [55] which eradicate low quality reads and portions of reads to give the best hope of gaining accurate results from downstream analysis.

Alignment of short sequencing reads to a reference

The alignment of short NGS reads to a reference genome is a process common to many applications of these data[56]. This process is used extensively to identify patterns of variation between individuals, isolates or sequences. However, there are problems inherent in both the data and the application. The reads are short sub-sequences of a much larger template and are likely to contain errors. Given the repetitive nature of certain genomic regions there are possibly several places which the short read can align[57]. The whole process presents a massive computational challenge as alignment must be performed for many thousands or millions of short reads. This is made even more challenging as one of the most useful applications of this process is to identify regions of difference between two closely related sequences, be that individual polymorphic bases, short inserted or deleted regions (indels) with respect to the reference and missing genes or whole regions. This means that the alignment process must take into account the possibility of these variations and report the best alignment comparison possible.

These issues have been dealt with using complex computational algorithms in such software as BOWTIE [58], BOWTIE2 [59] and BWA [40] which use the Burrows Wheeler transform to create a permanent reusable index for the reference genome. The short reads are then aligned to this reference at a colossal rate, for BOWTIE this can reach speeds of up to 25 million reads per CPU hour. For a comprehensive explanation of protocols see [41].

The amounts of data and results generated by this process present challenges for analysis, presentation and storage. There are several widely used packages that have various tools designed to standardise and facilitate the analysis of NGS data. The most widely used are SAMtools [60] and the Genome Analysis Tool Kit (GATK) produced by the Broad institute [61]. These sequence analysis toolboxes allow the efficient analysis of large NGS datasets and have had a great impact on the field by attempting to standardise formats and protocols used throughout the field in an effort to unify global research techniques.

De novo assembly of short sequencing reads

The assembly of short NGS reads into larger contiguous lengths of sequence representative of the template from which they were based upon has had may approaches [62,63]. The most popular of which has been the use of De Bruijn graphs to reduce the computational burden of dealing with such massive amounts of short and error prone reads. Examples of software which utilise this method include VELVET [64], SPAdes [43] and SOAPdenovo [65]. This method essentially breaks the reads down into k-mers – substrings of sequence. A graph is then created with each portion of each k-mer as a distinct node and the k-mers as a directed edge in the graph. Algorithms are then used to pick the most efficient path through the graph and therefore the most likely sequence [66].

The success of genome assembly from short reads depends a great deal on the pre-assembly quality control of the data generated by the sequencer. As mentioned previously, sequencing does have inherent error rates (for Illumina it is approximately 1 error in 1000 bases sequenced [22] and these errors can obviously cause great problems when assembling into a representative contig. Therefore, eradicating as many of the unreliable portions of the sequence reads prior to sequencing improves the accuracy of the finished result.

The genome assembly process is more normally a pipeline consisting of pre assembly filtering and trimming, repeated iterations of assembly with various parameters including k-mer length, followed by post assembly verification. Post assembly verification is performed using the original short read data and aligning this back to the assembly generated, in an attempt to identify breakpoints where the assembly process has become confused. The assembly is then broken at these points and an attempt can be made to reassemble the contigs using paired read data to inform the process.

There are issues inherent with the nature of the data produced by NGS and the templates being sequenced. Genomes tend to have repeated regions, mobile elements, pseudogenes and any number of other sequence anomalies that make assembly using short reads challenging. For example, if a motif repeat is longer than the reads generated then the assembly process is unable to ascertain the length of the region and the assembly will most likely break at this point. If there are several duplications within the template, the assembly may become confused leading to artefacts being included in the final assemblies or a large amount of (sometimes short) contigs. For example long repeated regions can be collapsed into a single element of the repeat in the final assembly, misrepresenting the true sequence. Any error in the sequence data will also contribute to unreliable, fragmented final assemblies.

Assembly using short NGS data has developed into a field in its own right with the NCBI genomes database containing over 70,000 bacterial and over 3000 eukaryotic genomes (as of 2016), not to mention viral, plasmid and organellar sequences. There has even been a number of assembly competitions to assess the performance of different assembly groups and software pipelines [67,68].

Comparative genomics approaches

Comparative genomics, as mentioned earlier, has been one of the fields of study made possible by the invention and proliferation of DNA sequencing and has gained in momentum vastly since the advent of NGS lowered the cost and increased the throughput and versatility of the process. DNA forms the blueprint of all life on earth. Genomic sequences code the information necessary for each organism to perform all of its biological processes. These genomic features are many and varied and come in the form of DNA sequence, genes, gene order, structural variation and RNA to name but a few and the list is ever growing as new elements are added as their functions are discovered. Even minute variation in these elements can have a profound effect on the phenotype of the organism concerned and is responsible for the biological diversity still being uncovered. The sequencing of many closely related individuals or isolates and comparing the sequence information generated allows insight into the genomic nature of phenotypic characteristics. A huge amount of research effort has been dedicated to this field of study, and the basic methodology has been applied to a vast array of disparate taxa from bacteria [69] to eukaryotes [70] ,humans [71] and even viruses [72]. It is by examining the minute differences between closely related organisms and the similarities between evolutionarily diverse but phenotypically convergent organisms that we have been able to begin to unpick the puzzle that genomics has unlocked.

It could be said that the first comparative genomics study was carried out with the publishing of the second genome sequence in the mid-1970s but it was not until the mid-1990s that the first comparisons between whole genomes of cellular organisms became a possibility. The Tatusov *et al* paper in 1996 compared the recently published genome sequence of *H. influenzae* with that of *E. coli* [73], marking the beginning of what we know today as comparative genomics. Several landmark studies followed using comparative genomics to uncover unprecedented amounts of genomic diversity within bacterial species. *Campylobacter jejuni* was found to have large numbers of homopolymeric repeats in genes responsible for surface structure biosynthesis or modification [74] and the *Bacteroides fragilis* genome was found to contain many inverted DNA repeats [75]. There are other sequencing approaches that can be used and there are many questions that lend themselves to slightly different forms of genomic investigation such as microarray studies. Willenbrock *et al* designed a microarray useful to characterise the pan and core genome of *E. coli* [76] with Dickinson *et al* and Han *et al* [77,78] characterising structural variation in canine tumour cell DNA and the chicken genome respectively.

The study of prokaryote genomics benefitted from the application of comparative genomics approaches. Given the diversity of phenotypes displayed by prokaryotes and the relatively simplistic genomic organisation when compared to eukaryotes, it is not surprising that these fascinating organisms have been the focus of much scientific research. One of the earliest studies to utilise NGS for comparative genomics was Baker *et al* 2008 [79]. This study compared 140 isolates of *Salmonella enteric* serovar Typhi from Indonesia. The field has grown steadily however along with the resources afforded by advancements in both sequencing technologies and computing power.

Comparative genomics studies have also revealed insight into bacterial evolution, showing the high incidence of horizontal gene transfer in species such as *E. coli* [80]. Comparative genomic studies have also facilitated the identification, cataloguing and comparison of bacterial virulence and avirulence factors. There are many examples of this particularly relevant are studies focusing on type three secretion system (T3SS) effectors such as Bart *et al* [81] which compared 65 strains of *Xanthomonas axonopodis* pv. *manihotis* identifying the conserved and variable effector profile for this pathovar.

Large-scale sequencing projects are becoming more and more common and with the reduction in cost and improvements in technology and library preparation the number of genomes sequenced in each study is rising (Figure 5). This along with the improvements in sequencing technology, computing resources and analysis tools only increase the potential for comparative genomics studies.

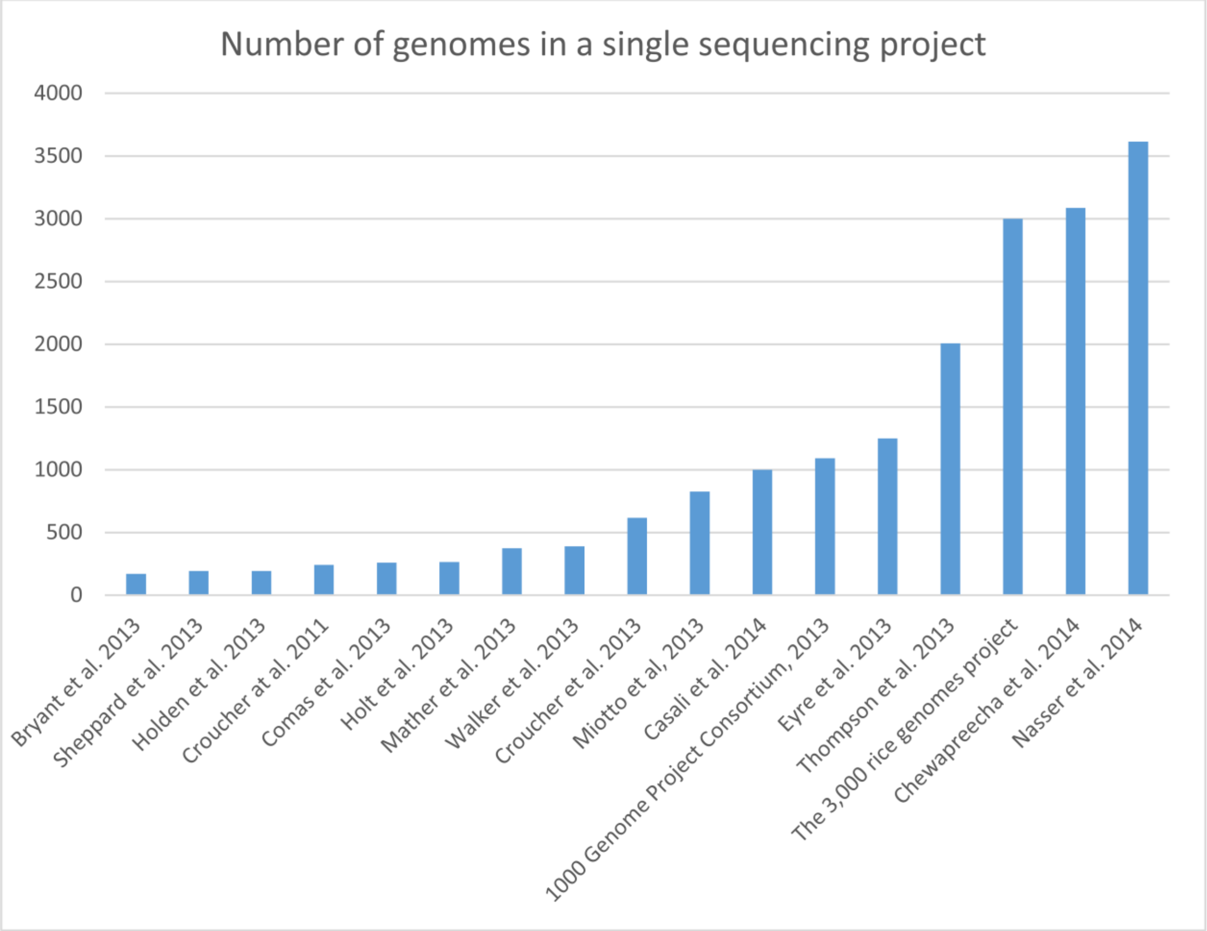


Figure 5: Graph showing the increase in numbers of genomes sequenced in a single project until 2014.
Data courtesy of [123].

Bacteria

Bacteria are the most numerous biological organisms on the planet. They can be found in every environment imaginable and have evolved to not only survive but flourish in some of the most inhospitable environs known. Bacteria have been isolated from deep sea hydrothermal vents [82,83], hot springs [84], glacial ice sheets thousands of years old [85], arctic soil [86] and even outer space [87]! The symbiotic nature of certain bacterial taxa is well known [88], there are examples of prokaryotic symbionts existing within eukaryotes and even other bacteria [89]. It has been recently revealed how important the fauna of the gut is to human health and the bacteria present in the digestive systems of ruminants has long been known to facilitate the digestion of cellulose [90]. Further to this, the endosymbiotic theory has been suggested. It is now widely accepted that ancient symbioses between early eukaryotes and internal prokaryotes gave rise to certain eukaryotic organelles such as the plastid and, more anciently and perhaps contentiously, the mitochondrion and the basal flagellar body [91]. This is supported by much evidence including the presence of discrete genomic material in both the plastid and the mitochondria which has much more in common with prokaryotic material than their host eukaryote, for example the unusual small subunit rDNA genes encoded by the plastid and mitochondrial genomes. However, it is research into the genomics of pathogenic bacteria which have garnered perhaps the most intense research effort as these bacterial species have the most direct and significant effect on the human population. There are many bacteria directly impacting human health such as the now famous Methicillin resistant *Staphylococcus aureus* or highly infectious pathogens such as *Clostridium botulinum*, *Vibrio cholera*, *Mycobacterium tuberculosis* and *Salmonella enteriditis* serovar *Typhi* which have caused untold human suffering and mortality. However, there are also bacterial pathogens with an indirect effect on human wellbeing for example plant pathogens. There are many examples of these bacterial pathogens affecting crops such as *Xanthomonas* species, *Pseudomonas* species and *Ralstonia* species. These pathogens have been the cause of crop loss and economic devastation causing untold human misery over the centuries.

There are some elements of bacterial physiology that make them ideal candidates for genomic study. In comparison to eukaryotic organisms, even unicellular eukaryotes and protists, prokaryotes are far less complex. That being said they are still fantastically intricate biological machines that we are still far from fully appreciating. Important for this study are the genomic features which set bacteria apart. Bacteria have no nucleus; therefore, their genomic material is much easier to access. Bacteria have relatively small genomes, ranging from less than 240 kb in length (e.g. *Candidatus tremblaya* princeps [92]) up to the modest (comparatively speaking) ~13 mb genome of *Sorangium Cellulosum* [93] . The bacterial genomic DNA is organised into chromosomes, normally singular but can be more for example *Vibrio cholera* is known to possess two [94] . These chromosomes are divided into non-intronic genes, with very little non-coding DNA.

As well as their genomic DNA, many bacterial species also possess extra chromosomal DNA in the form of plasmids [95]. These can be circular or linear in nature and can range in size from tiny plasmids of less than 1kb in length such as the *Candidatus tremblaya phenacola* PAVE plasmid to several mb which are equitable in size to bacterial chromosomes such as the pSCL4 mega-plasmid found in *Streptomyces clavuligerus* [96]. These plasmid sequences can code for many important functions; there are often genes held on plasmids which have an influence on important evolutionary processes such as niche adaptation and pathogenicity the importance of which will be covered later. Plasmids can also be exchanged between bacteria of different species, genera or even more distantly related taxa through conjugation. Plasmid conjugation is not the only method of accelerated evolutionary change; bacteria have also been shown to exchange genomic sequence.

Horizontal gene transfer (HGT) events have been shown to be numerous throughout the long history of bacterial evolution [80]. HGT has been an extremely important driver of the evolutionary process both micro and macro, allowing the transfer of genes coding for certain traits between distantly related taxa, thus potentially sharing genomic information important to survive in certain environments or hosts. These methods are vital elements facilitating the spread of bacterial species to disparate environmental niches and lifestyles [97,98].

Bacterial taxonomy has always proved to be a difficult subject, due in no small part to their microscopic unicellular nature. Morphologically there are many specific structures which can be used to differentiate between taxa and metabolic and biochemical characteristics can also be used. However, due to the enormous variety of prokaryotes and their unicellular nature, identification, differentiation and grouping is an issue for traditional taxonomic strategies. In the age of almost universal access to affordable sequencing technologies however, the use of genomics for classification is far more successful. Sanger sequencing was, and to certain extent still is used for the purpose of attempting to assess the topology of the bacterial tree of life. There were numerous studies focusing on single gene or a small number of genes for use in molecular phylogenetics [99]. Many bacterial genes were sampled to use as phylogenetic markers [99], the success of each being dependent on the question being asked. Highly conserved housekeeping genes are excellent candidate for single gene phylogenies trying to assess the relationships amongst quite disparate taxa as they are slowly evolving therefore evolutionarily distant taxa can be compared. The most widely used example is the 16S small subunit rRNA [100] gene which is ubiquitous throughout the bacterial kingdom. However, being ubiquitous and slowly evolving are exactly the characteristics that make it less suitable for more high resolution questions focusing on more closely related species. It has been suggested that the 16S gene is of little use as a taxonomic marker for higher than genus level perhaps species level studies [100]. For these finer resolution questions, it is often more appropriate to use genes that are perhaps less wide spread and faster evolving. For example Gyrase B(GyrB) [101] the gene which ecodes the B subunit of the gyrase enzyme vital for DNA replication is a well-accepted taxonomic marker for sub-genera level taxonomy in *Xanthomonas* species. A further option for molecular bacterial taxonomy and classification is multi locus sequence analysis (MLSA) in which analysis is performed on concatenated markers from a number of genes, or indeed their full lengths [99].

Bacterial secretion systems

One group of bacterial cellular mechanisms that has been highlighted as important for both niche adaptation and pathogenicity is the secretion systems. These miniature biological machines are utilised by Gram-negative bacteria to translocate a wide variety of substrates from the intra-cellular space to either the periplasm, the extracellular space or to the interior of another cell.

To date there have been seven secretion systems (Figure 6) identified in Gram-negative bacteria, numbered 1-7. Five of these systems are known to translocate substrates from the cell interior to the cell exterior bridging both the inner and outer cell membranes: type 1, type 2, type 3 type 4 and type 6 secretion systems. Although for the purposes of this project the focus will be on the type 3 and type 6 secretion systems.

Type three secretion system

The bacterial T3SS is analogous to a molecular needle which is assembled by the bacterial cell. These nano-syringes are used by a wide variety of bacteria, mainly proteobacteria [102] with a variety of lifestyles including pathogens and symbionts. The T3SS machinery is coded for by more than twenty genes [103] including *hrp* which forms the subunits that build the needle apparatus. The structure of the T3SS was first identified in *Salmonella typhimurium* [104] has been shown to be closely related to that of the bacterial flagellum and it has been suggested that the T3SS has either evolved directly from this system [105] or that the two systems share a common ancestor [106] . These double membrane embedded nanomachines deliver effector proteins into the cytoplasm of target cells. These effector proteins have been shown to have several functions: targeting and subverting specific cellular processes within the target cells in order to facilitate the survival, colonisation and reproduction of the pathogen. These proteins, known as effectors, have been shown to have essential functions in the bacterial cells pathogenic repertoire. The effector complement of [107] each bacterial strain can differ widely depending on host or lifestyle in which it needs to survive and the adaptations host organisms have evolved to combat infestation.

The zigzag model of plant pathogen interaction suggested by Jones and Dangl [108] posits a two stage immune response to bacterial invasion in plants. The first stage is based on the recognition of pathogen associated molecular patterns (PAMPS) by pattern recognition receptors (PRRs) in the plant cell. These PRRs recognise PAMPS triggering an immune response known as PAMP Triggered Immunity (PTI). PAMPs are slowly evolving molecular signatures of bacterial pathogens such as flagellin [109] and lipopolysacharides [110] which are an integral element of the outer membrane of gram negative bacteria. These PAMPs are important to the bacterial cell, and not easily lost or changed so make excellent targets for host cell PRRs. As a reaction to this immune response bacteria have evolved secretion systems such as the T3SS which secrete effector proteins which interfere with and subvert the PTI in the host. However, the host organism uses polymorphic ND-LRR proteins to

recognise these effectors. These ND-LRR proteins then trigger the second more severe effector triggered immune response (ETI) thus develops an arms race

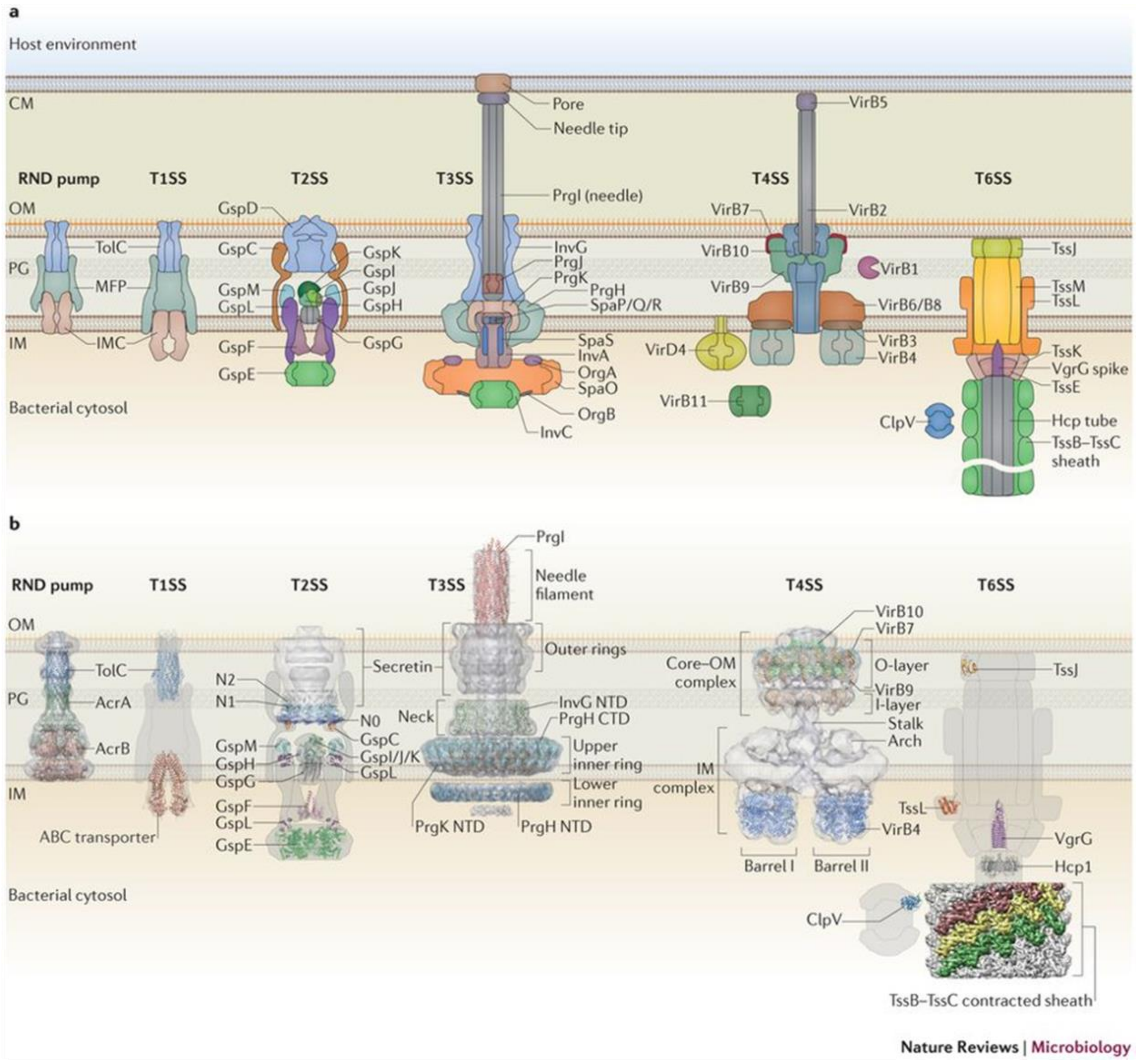


Figure 6: Schematic of the different types of bacterial secretion systems. Courtesy of [107].

between host and pathogen, with effectors variously being both virulence and avirulence factors depending on the time in the host pathogen interaction cycle. Pathogens gain novel effector proteins through, for example horizontal gene transfer, which give them the edge over the host, which in turn evolves novel ND-LRR proteins to recognise these threats and evoke a response. Modifications to the zigzag model have been suggested [111] such as the inclusion of damage associated molecular patterns (DAMPS) which is the plant recognising bacterial damage and triggering immunity based on this [112].

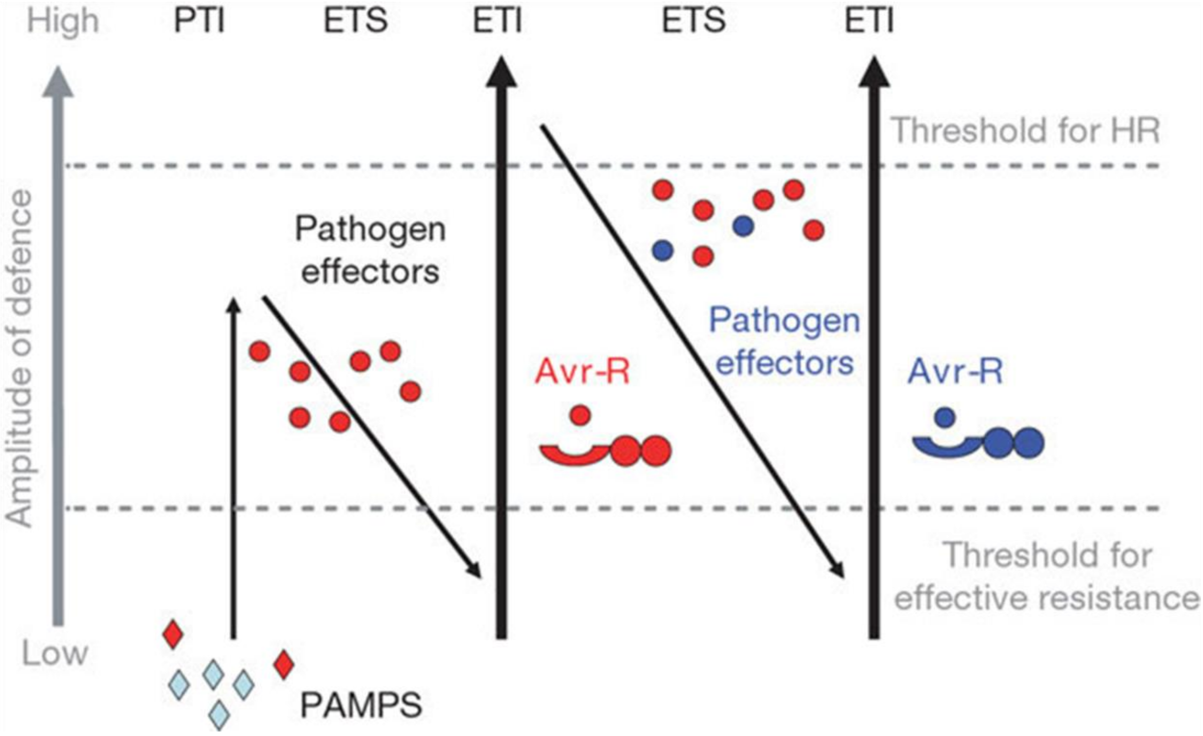


Figure 7; schematic of the zigzag model of plant-pathogen interaction. Courtesy of [108].

Type 6 secretion system

The Type 6 Secretion System (T6SS) was initially discovered in *Vibrio cholerae* in 2006 [113] and is not as well characterised as the T3SS. Since its discovery, bioinformatics studies have identified the T6SS in 25% of gram negative bacterial taxa, including *Escherichia* species [114], *Pseudomonas* species [115] and *Campylobacter* species [116]. Initially the T6SS was suggested to have a wide range of cellular functions including virulence [117] and host immunomodulation [118]. Recent studies have also suggested further inter-bacterial functions for this secretion system such as bactericidal activity [119]. The T6SS has been shown to be present in a wide range of pathogenic bacteria and there are known to be several copies in some species. The T6SS is known to be able to translocate effector proteins into both prokaryotes and eukaryotes [120]. Two important signature proteins of the T6SS machinery are Hcp (haemolysin co-regulated protein) VgrG (valine-glycine repeat G) which have been suggested as both structural elements and translocated proteins [115].

The T6SS genes are usually located on pathogenicity islands, for example the *Vibrio* pathogenicity island 1 island in *Vibrio cholera*. These genetic islands are now known to possess a variable number of genes, however the core components present in most known, functional T6SS are *IcmF*, *IcmH*, *ClpV*, *Hcp* and *VgrG*.

Thesis aims and objectives

This thesis describes the exploitation of recent technological developments in genome sequencing to investigate the genomics and evolution of pathogenicity in bacteria.

In section 1 we aim to use next generation sequencing to identify and characterise novel genomic features of bacterial pathogens which influence host specificity and virulence exemplars being the T3SS and its effector complement. We characterise and compared these genomic features identifying effector profiles and characterise the role of these in virulence. This will help address several related questions regarding the biology and the mechanisms of evolution of bacterial pathogenesis. The analysis we present here will identify novel signatures of increased virulence, phenotypic convergence and adaptation to new ecological niches in bacteria. This work has wide ranging potential to inform surveillance of these pathogens and the development of targeted resistance in crop management.

Whilst multi-genome sequencing is commonplace in studies of human pathogens, at the outset of this project (in 2010), phytopathology lagged a little behind; yet the potential benefits were apparent. We focussed on a group of bacterial phytopathogens with huge impact for food security and economics from the *Xanthomonas* species. We aimed to use next generation sequencing to survey the genomic components of *Xanthomonas* pathogens of sugar cane and bean species to identify signatures of recent evolution of pathogenicity and adaptation to new hosts. We aimed to classify strains using molecular taxonomy and to use comparative genomics to identify genomic features facilitating phenotypic convergence and the colonisation of a host by distantly related pathogens

In section two we focus on the use of next generation sequencing and genomics to contribute to the understanding of emerging pathogens. Bacterial pathogens pose a huge threat to both human health and food security. Understanding the genomics of newly emerging threats can help track the spread of emerging bacterial threats and inform response; next generation sequencing provides an ideal tool to assist this analysis.

We aimed to analyse two emerging bacterial threats. The first is a virulent form of *Campylobacter* infection found in the Far East. We use next generation sequencing and comparative genomics to identify genomic features characteristic of this worrying infection. The study aims to contribute to knowledge of this potential threat to human health and to identify molecular markers to track possible outbreaks.

The second example of the use of NGS to investigate a newly emerging bacterial threat is a recently identified *Xanthomonas* bean pathogen from a disease outbreak in Rwanda. We aimed to generate NGS data from this emerging pathogen, assemble the genome and use these data to classify this novel strain. We then survey and characterize genomic elements, contributing to the knowledge of this potentially serious emerging pathogen of beans.

Finally, in section three of this thesis we present an early evaluation of the ONT MinION, one of the flagship examples of third generation sequencing technology. This technology has the potential to revolutionise the study of the genomics of bacterial pathogens. We will present a comprehensive assessment of the utility of this novel technology for bacterial sequencing and metagenomics. In order to explore several characteristics including data quality and read length and assess the implications of any limitations of this exciting technology three strains with varying genome size and G + C content were used. The ONT MinION has since been optimised and field tested extensively and is now a key tool used in the field of bacterial genomics.

References

1. Nasser, W. *et al.* Evolutionary pathway to increased virulence and epidemic group A *Streptococcus* disease derived from 3,615 genome sequences. *Proc. Natl. Acad. Sci.* 1403138111- (2014).
2. Chewapreecha, C. *et al.* Dense genomic sampling identifies highways of *pneumococcal* recombination. *Nat. Genet.* **46**, 305–9 (2014).
3. Gamez, R. M. *et al.* Genome Sequence of the Banana Plant Growth-Promoting *Rhizobacterium Pseudomonas fluorescens* PS006. *Genome Announc.* **4**, (2016).
4. Jia, B., Jin, H. M., Lee, H. J. & Jeon, C. O. Draft Genome Sequence of *Zhouia amylolytica* AD3, Isolated from Tidal Flat Sediment. *Genome Announc.* **4**, (2016).
5. Zhang, X., Zhao, C., Hong, X., Chen, S. & Yang, S. Genome Sequence of *Marichromatium gracile* YL-28, a Purple Sulfur Bacterium with Bioremediation Potential. *Genome Announc.* **4**, (2016).
6. Lancaster, W. A. *et al.* Near-Complete Genome Sequence of *Clostridium paradoxum* Strain JW-YL-7. *Genome Announc.* **4**, (2016).
7. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* **37**, D26–D31 (2009).
8. Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. & Petersen, G. B. Nucleotide sequence of bacteriophage Lambda DNA. *J. Mol. Biol.* **162**, 729–773 (1982).
9. Smith, L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674–679 (1986).

10. Oliver, S. G. *et al.* The complete DNA sequence of yeast chromosome III. *Nature* **357**, 38–46 (1992).

11. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science (80-.).* **269**, 496–512 (1995).

12. Blattner, F. R. *et al.* The Complete Genome Sequence *of Escherichia coli* K-12 . *Science (80-.).* **277**, 1453–1462 (1997).

13. Consortium, the *C. elegans* sequencing. Genome Sequence of the Nematode *C . elegans* : A Platform for Investigating Biology. **2012**, 2012–2019 (2012).

14. Giovannoni, S. J., Britschgi, T. B., Moyer, C. L. & Field, K. G. Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345**, 60–63 (1990).

15. Morris, R. M. *et al.* SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**, 806–810 (2002).

16. Walker, G. J. *et al.* Association of Genetic Variants in NUDT15 with Thiopurine-Induced Myelosuppression in Patients with Inflammatory Bowel Disease. *JAMA - J. Am. Med. Assoc.* **321**, 753–761 (2019).

17. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

18. Venter, J. C. *et al.* The sequence of the human genome. *Science (80-.).* **291**, 1304–1351 (2001).

19. Margulies, M. *et al.* Genome Sequencing in Open Microfabricated High Density Picoliter Reactors. *Nat. Biotechnol.* **437**, 376–380 (2006).

20. Pyrosequencing Technology and Platform Overview. Available at: https://www.qiagen.com/gb/service-and-support/learning-hub/technologies-and-research-topics/pyrosequencing-resource-center/technology-overview/.

21. Rothberg, J. M. & Leamon, J. H. The development and impact of 454

sequencing. *Nat. Biotechnol.* **26**, 1117–1124 (2008).

22. Liu, L. *et al.* Comparison of next-generation sequencing systems. *Role Bioinforma. Agric.* **2012**, 1–25 (2014).

23. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).

24. Nakazato, T., Ohta, T. & Bono, H. Experimental design-based functional mining and characterization of high-throughput sequencing data in the sequence read archive. *PLoS One* **8**, e77910–e77910 (2013).

25. Ergüner, B., Üstek, D. & Sağıroğlu, M. Ş. Performance comparison of Next Generation sequencing platforms. in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 6453–6456 (2015

26. Mardis, E. R. Next-Generation Sequencing Platforms. *Annu. Rev. Anal. Chem.* **6**, 287–303 (2013).

27. Guo, Y. *et al.* The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics* **13**, 666 (2012).

28. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40**, e72–e72 (2012).

29. Meacham, F. *et al.* Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics* **12**, 451 (2011).

30. Hansen, K. D., Brenner, S. E. & Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* **38**, e131–e131 (2010).

31. Ross, M. G. *et al.* Characterizing and measuring bias in sequence data.

Genome Biol. **14**, R51 (2013).

32. Holt, R. A. & Jones, S. J. M. The new paradigm of flow cell sequencing. *Genome Res.* **18**, 839–846 (2008).

33. Scally, A. *et al.* Europe PMC Funders Group Insights into hominid evolution from the gorilla genome sequence. **483**, 169–175 (2012).

34. Deng, J. *et al.* Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat. Biotechnol.* **27**, 353–60 (2009).

35. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics, Proteomics Bioinforma.* **13**, 278–289 (2015).

36. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: Delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17**, 1–11 (2016).

37. Lu, H., Giordano, F. & Ning, Z. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics Bioinforma.* **14**, 265–279 (2016).

38. Judge, K., Harris, S. R., Reuter, S., Parkhill, J. & Peacock, S. J. Early insights into the potential of the Oxford Nanopore MinION for the detection of antimicrobial resistance genes. *J. Antimicrob. Chemother.* **70**, 2775–2778 (2015).

39. Wetterstrand, K. DNA sequencing costs - data. Available at: https://www.genome.gov/27541954/dna-sequencing-costs-data/.

40. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

41. Langmead, B. Alignment with Bowtie. 1–24 (2011).

42. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly

using de Bruijn graphs. *Genome Res.* **18**, 821–9 (2008).

43. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

44. Aziz, R. K. *et al.* The RAST server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).

45. Tatusova, T. *et al.* Prokaryotic Genome Annotation Pipeline. (2013).

46. Cantarel, B. L. *et al.* MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).

47. Gilbert, J. A. *et al.* The seasonal structure of microbial communities in the Western English Channel. *Environ. Microbiol.* **11**, 3132–3139 (2009).

48. Welch, D. B. M. & Huse, S. M. Microbial Diversity in the Deep Sea and the Underexplored 'Rare Biosphere'. *Handb. Mol. Microb. Ecol. II Metagenomics Differ. Habitats* 243–252 (2011).

49. Cheung, M. K., Li, L., Nong, W. & Kwan, H. S. 2011 German Escherichia coli O104:H4 outbreak: whole-genome phylogeny without alignment. *BMC Res. Notes* **4**, 533 (2011).

50. Scheutz, F. Eurosurveillance , Volume 16 , Issue 24 , 16 June 2011 Characteristics of the enteroaggregative shiga toxin / verotoxin-producing Escherichia coli o104 : h4 strain causing the outbreak of haemolytic uraemic syndrome in Germany , may to june 2011. *Strain* **16**, 1–8 (2011).

51. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–32 (2016).

52. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* **38**, 1767–1771 (2010).

53. Andrews S. FastQC: a quality control tool for high throughput sequence data.

54. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinforma.* (2014).

55. Aronesty, E. Comparison of Sequencing Utility Programs. *Open Bioinforma. J.* **7**, 1–8 (2013).

56. MacLean, D., Jones, J. D. G. & Studholme, D. J. Application of &#39;next-generation&#39; sequencing technologies to microbial genetics. *Nat. Rev. Microbiol.* **7**, 287 (2009).

57. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **13**, 36–46 (2011).

58. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).

59. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).

60. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

61. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *GENOME BIOL* **20**, 1297–1303 (2010).

62. Wajid, B. & Serpedin, E. Do it yourself guide to genome assembly. *Brief. Funct. Genomics* **15**, 1–9 (2014).

63. Sohn, J. & Nam, J.-W. The present and future of de novo whole-genome assembly. *Brief. Bioinform.* **19**, 23–40 (2016).

64. Zerbino, D. R. & Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).

65. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).

66. Compeau, P. E. C., Pevzner, P. A. & Tesler, G. How to apply de Bruijn graphs to genome assembly. *Nat Biotech* **29**, 987–991 (2011).

67. Earl, D. *et al.* Assemblathon 1: a competitive assessment ofde novo short read assembly methods. *Genome Res* **21**, 2224–2241 (2011).

68. Bradnam, K. R. *et al.* Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2**, 10 (2013).

69. Prentice, M. B. Bacterial comparative genomics. *Genome Biol.* **5**, 338 (2004).

70. Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287**, 2204–2215 (2000).

71. Clark, M. S. Comparative genomics: the key to understanding the human genome project. *BioEssays* **21**, 121–130 (1999).

72. Goz, E. *et al.* Generation and comparative genomics of synthetic dengue viruses. *BMC Bioinformatics* **19**, 140 (2018).

73. Tatusov, R. L. *et al.* Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with Escherichia coli. *Curr. Biol.* **6**, 279–291 (1996).

74. Parkhill, J. *et al.* The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* **403**, 665–8 (2000).

75. Loman, N. J. & Pallen, M. J. Twenty years of bacterial genome sequencing. *Nat. Rev. Microbiol.* **13**, 1–9 (2015).

76. Willenbrock, H., Hallin, P. F., Wassenaar, T. M. & Ussery, D. W. Characterization of probiotic *Escherichia coli* isolates with a novel pan-genome microarray. *Genome Biol.* **8**, R267 (2007).

77. Dickinson, P. J. *et al.* Chromosomal Aberrations in Canine Gliomas Define Candidate Genes and Common Pathways in Dogs and Humans. *J. Neuropathol. Exp. Neurol.* **0**, nlw042 (2016).

78. Han, R. *et al.* Identification and functional characterization of copy number variations in diverse chicken breeds. *BMC Genomics* **15**, 934 (2014).

79. Baker, S. *et al.* High-throughput genotyping of *Salmonella enterica* serovar *Typhi* allowing geographical assignment of haplotypes and pathotypes within an urban district of Jakarta, Indonesia. *J. Clin. Microbiol.* **46**, 1741–1746 (2008).

80. Hayashi, T. *et al.* Complete genome sequence of enterohemorrhagic *Escherichia coli* O157 : H7 and genomic comparison with a laboratory strain K-12. *DNA Res.* **8**, 11–22 (2001).

81. Bart, R. *et al.* PNAS Plus: High-throughput genomic sequencing of cassava bacterial blight strains identifies conserved effectors to target for durable resistance. *Proc. Natl. Acad. Sci.* **109**, E1972–E1979 (2012).

82. Campanaro, S. *et al.* Laterally transferred elements and high pressure adaptation in Photobacterium profundum strains. *BMC Genomics* **6**, 122 (2005).

83. Takai, K., Nealson, K. H. & Horikoshi, K. Hydrogenimonas thermophila gen. nov., sp. nov., a novel thermophilic, hydrogen-oxidizing chemolithoautotroph within the E-Proteobacteria, isolated from a black smoker in a Central Indian Ridge hydrothermal field. *Int. J. Syst. Evol. Microbiol.* **54**, 25–32 (2004).

84. Mehetre, G. T., Paranjpe, A. S., Dastager, S. G. & Dharne, M. S. Complete metagenome sequencing based bacterial diversity and functional insights from basaltic hot spring of Unkeshwar, Maharashtra, India. *Genomics Data* **7**, 140–143 (2016).

85. Knowlton, C., Veerapaneni, R., D'Elia, T. & Rogers, S. O. Microbial analyses of ancient ice core sections from greenland and antarctica. *Biology (Basel).* **2**, 206–32 (2013).

86. Johnston, E. R. *et al.* Metagenomics Reveals Pervasive Bacterial Populations and Reduced Community Diversity across the Alaska Tundra Ecosystem. *Front. Microbiol.* **7**, 1–16 (2016).

87. Novikova, N. *et al.* Survey of environmental biocontamination on board the International Space Station. *Res. Microbiol.* **157**, 5–12 (2006).

88. Sachs, J. L., Skophammer, R. G. & Regus, J. U. Evolutionary transitions in bacterial symbiosis. *Proc. Natl. Acad. Sci.* **108**, 10800 LP – 10807 (2011).

89. von Dohlen, C. D., Kohler, S., Alsop, S. T. & McManus, W. R. Mealybug beta-proteobacterial endosymbionts contain gamma-proteobacterial symbionts. *Nature* **412**, 433–6 (2001).

90. Hiltner, P. & Dehority, B. A. Effect of soluble carbohydrates on digestion of cellulose by pure cultures of rumen bacteria. *Appl. Environ. Microbiol.* **46**, 642–648 (1983).

91. Margulis, L. Symbiotic theory of the origin of eukaryotic organelles; criteria for proof. *Symp. Soc. Exp. Biol.* 21–38 (1975).

92. López-Madrigal, S., Latorre, A., Porcar, M., Moya, A. & Gil, R. Complete genome sequence of 'Candidatus Tremblaya princeps' strain PCVAL, an intriguing translational machine below the living-cell status. *J. Bacteriol.* **193**,

5587–5588 (2011).

93. Schneiker, S. *et al.* Complete genome sequence of the myxobacterium Sorangium cellulosum. *Nat. Biotechnol.* **25**, 1281–1289 (2007).

94. Heidelberg, J. F. *et al.* DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae. *Nature* **406**, 477–483 (2000).

95. Shintani, M., Sanchez, Z. K. & Kimbara, K. Genomics of microbial plasmids: Classification and identification based on replication and transfer systems and host taxonomy. *Front. Microbiol.* **6**, 1–16 (2015).

96. Medema, M. H. *et al.* The Sequence of a 1.8-Mb Bacterial Linear Plasmid Reveals a Rich Evolutionary Reservoir of Secondary Metabolic Pathways. *Genome Biol. Evol.* **2**, 212–224 (2010).

97. Awadalla, P. The evolutionary genomics of pathogen recombination. *Nat. Rev. Genet.* **4**, 50—60 (2003).

98. Ahmed, N., Dobrindt, U., Hacker, J. & Hasnain, S. E. Genomic fluidity and pathogenic bacteria: applications in diagnostics, epidemiology and intervention. *Nat. Rev. Microbiol.* **6**, 387 (2008).

99. Glaeser, S. P. & Kämpfer, P. Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. *Syst. Appl. Microbiol.* **38**, 237–245 (2015).

100. Case, R. J. *et al.* Use of 16S rRNA and rpoB Genes as Molecular Markers for Microbial Ecology Studies. *Appl. Environ. Microbiol.* **73**, 278–288 (2007).

101. Parkinson, N. *et al.* Phylogenetic analysis of Xanthomonas species by comparison of partial gyrase B gene sequences. *Int. J. Syst. Evol. Microbiol.* **57**, 2881–2887 (2007).

102. McCann, H. C. & Guttman, D. S. Evolution of the type III secretion system and its effectors in plant–microbe interactions. *New Phytol.* **177**, 33–47 (2008).

103. Galan, J. E. & Wolf-watz, H. Protein delivery into eukaryotic cells by type III secretion machines. *Nature* **444**, 567–573 (2006).

104. Kubori, T. *et al.* Supramolecular structure of the Salmonella typhimurium type III protein secretion system. *Science* **280**, 602–605 (1998).

105. Gijsegem, F. *et al.* The hrp gene locus of *Pseudomonas solanacearum,* which controls the production of a type III secretion system, encodes eight proteins related to components of the bacterial flagellar biogenesis complex. *Mol. Microbiol.* **15**, 1095–1114 (1995).

106. Gophna, U., Ron, E. Z. & Graur, D. Bacterial type III secretion systems are ancient and evolved by multiple horizontal-transfer events. *Gene* **312**, 151–163 (2003).

107. Costa, T. R. D. *et al.* Secretion systems in Gram-negative bacteria: structural and mechanistic insights. *Nat. Rev. Microbiol.* **13**, 343–359 (2015).

108. Jones, J. D. G. & Dangl, J. L. The plant immune system. *Nature* **444**, 323–329 (2006).

109. Zipfel, C. & Felix, G. Plants and animals: A different taste for microbes? *Curr. Opin. Plant Biol.* **8**, 353–360 (2005).

110. Ariki, S. *et al.* A serine protease zymogen functions as a pattern-recognition receptor for lipopolysaccharides. *Proc. Natl. Acad. Sci.* **101**, 953–958 (2004).

111. Pritchard, L. & Birch, P. R. J. The zigzag model of plant-microbe interactions: is it time to move on? *Mol. Plant Pathol.* **15**, 865–870 (2014).

112. Boller, T. & Felix, G. A renaissance of elicitors: perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors. *Annu. Rev. Plant Biol.* **60**, 379–406 (2009).

113. Pukatzk, S. *et al.* Identification of a conserved bacterial protein secretion

system in *Vibrio cholerae* using the *Dictyostelium* host model system. *Proc Natl Acad Sci USA* **103**, 1528–1533 (2006).

114. Dudley, E. G., Thomson, N. R., Parkhill, J., Morin, N. P. & Nataro, J. P. Proteomic and microarray characterization of the AggR regulon identifies a pheU pathogenicity island in enteroaggregative *Escherichia coli*. *Mol. Microbiol.* **61**, 1267–82 (2006).

115. Mougous, J. D. *et al.* A Virulence Locus of *Pseudomonas aeruginosa* Encodes a Protein Secretion Apparatus. *Science (80-.).* **312**, 1526–1530 (2006).

116. Lertpiriyapong, K. *et al. Campylobacter jejuni* Type VI Secretion System: Roles in Adaptation to Deoxycholic Acid, Host Cell Adherence, Invasion, and In Vivo Colonization. *PLoS One* **7**, e42842 (2012).

117. Ma, A. T., McAuley, S., Pukatzki, S. & Mekalanos, J. J. Translocation of a Vibrio cholerae Type VI Secretion Effector Requires Bacterial Endocytosis by Host Cells. **5**, 234–243 (2009).

118. Chow, J. & Mazmanian, S. K. A pathobiont of the microbiota balances host colonization and intestinal inflammation. **7**, 265–276 (2010).

119. Hood, R. D. *et al.* A Type VI Secretion System of Pseudomonas aeruginosa Targets a Toxin to Bacteria. **7**, 25–37 (2010).

120. Ho, B. T., Dong, T. G. & Mekalanos, J. J. A view to a kill: The bacterial type VI secretion system. *Cell Host Microbe* **15**, 9–21 (2014).

121. Illumina NGS product comparison. Available at: http://www.illumina.com/systems/sequencing.html. (Accessed: 3rd April 2016)

122. Oxpord Nanopore MinIOn.

123. lab.loman.net. Available at: http://lab.loman.net/page9/. (Accessed: 19th April 2016)

124. GenBank and WGS Statistics. Available at: https://www.ncbi.nlm.nih.gov/genbank/statistics/. (Accessed: 4th July 2019)

Chapter 2:

Draft genome sequence of *Xanthomonas axonopodis* pathovar *vasculorum* NCPPB 900

Work from this chapter was published in:

Harrison J, Studholme DJ (2014) Draft genome sequence of *Xanthomonas axonopodis* pathovar *vasculorum* NCPPB 900. FEMS Microbiol Lett 360(2):113–6.

This paper was cited by:

1. Lang, J. M. *et al.* Detection and Characterization of *Xanthomonas vasicola* pv. *vasculorum* (Cobb 1894) comb. nov. Causing Bacterial Leaf Streak of Corn in the United States. *Phytopathology* **107**, 1312–1321 (2017).

2. Studholme, D. J. *et al.* Transfer of *Xanthomonas campestris* pv. *arecae*, and *Xanthomonas campestris* pv. *musacearum* to *Xanthomonas vasicola* (Vauterin) as *Xanthomonas vasicola* pv. *arecae* comb. nov., and *Xanthomonas vasicola* pv. *musacearum* comb. nov. and description of *Xanthomonas* va. *bioRxiv* 571166 (2019).

Introduction

Bacterial species of the genus *Xanthomonas* are gram negative, rod shaped gamma proteobacteria. The xanthomonads are typically pathogens of plants and the genus contains a multitude of species groups and sub-species or pathovars. Each pathovar has evolved to have a narrow host range, often limited to just one or a very small number of host plants. *Xanthomonas axonopodis* pv. *vasculorum* (*Xav*) is one of several species of *Xanthomonas* which have evolved to infect sugarcane (*Saccharum officinarum*) including *Xanthomonas vasicola* pv. *vasculorum* (*Xvv*) and *Xanthomonas sachchari*. In this economically important crop plant, *Xav* has been shown to cause the destructive infection “gumming disease” which is one of the oldest recorded diseases of sugar cane in the world [1]. Gumming disease has been described as a two stage infection. It begins with a foliar phase continuing on to a systemic phase later in the disease cycle. Among the early symptoms are yellowish, longitudinal stripes on the margins of the leaves which later develop into red or straw coloured stripes and the leaves can develop a bacterial sheen of exuded pathogen. The systemic infection follows with the chlorosis of younger leaves caused by the infection of the vascular bundles. The internal tissues of the stalk then develop pockets of gum-like bacterial exudate from which the disease gained its common name.

Xav has afflicted global sugar cane agriculture for many years, causing massive damage to crops and economic loss. *Xav* displays similarities in both host range and phenotypic properties to *Xvv* and consequently there has been misidentification and taxonomic confusion when attempting to identify bacterial infections of sugar cane. *Xav* has historically been misidentified as *Xanthomonas campestris* pv*. vasculorum* and *Xvv at various times*. However, using molecular sequence analysis methods such as [2–4] it has been possible to differentiate *Xanthomonas* species to a greater degree. The species previously known as *Xanthomonas campestris* pv. *vasculorum* has been shown to divide into two phylogenetically distinct pathovars grouping in separate species level clades within the Xanthomonas genus, *Xanthomonas vasicola* and *Xanthomonas axonopodis*. These sequence-based taxonomic distinctions can be further verified using SDS-PAGE and fatty acid profiling [1,3,4]. According to the classification scheme detailed in their 2000 paper, Dookun *et al* showed that *Xav* has a fatty acid type D which includes isolates of race 2 and 3. This classification grouped *Xav* with other pathovars collected from Reunion Island and Mauritius over a wide time period (1960 - 1992) from Sugar cane,

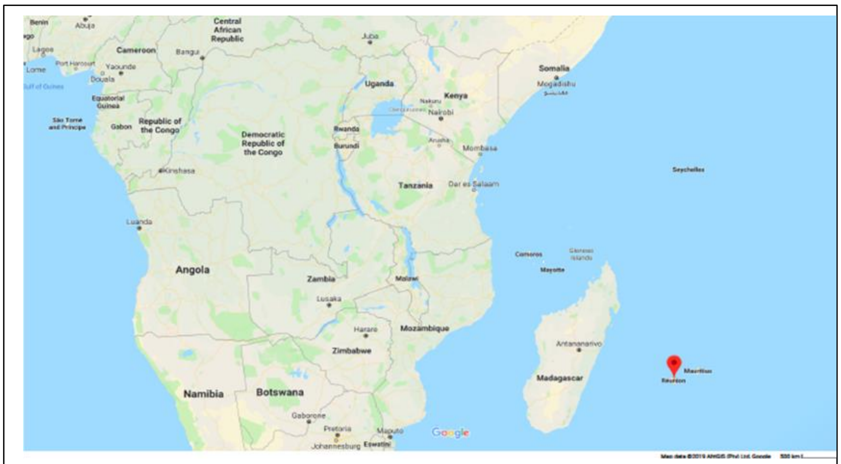


Figure 6: Map showing location of Reunion Island [5]

Roystonea regia (Cuban royal palm) and *Zea mays* (maize) [1]. This group phylogenetically branched along with type B (race 1) with in the *X. axonopodis* clade

The pathovar now identified as *Xav* NCPPB 900 was first collected from Ravine Creuse Reunion island (Figure 6) in 1960 following an epidemic in the region. It was isolated from sugar cane by A. C. Hayward and deposited as B386. This pathovar was at the time known to cause serious problems to the sugar cane industry in the region [1]. Although successful breeding strategies and other methods helped to control the disease there have been resurgences in the intervening time and has been suggested to be an emerging problem for sugar cane agriculture as recently as 2012 when it was identified in South America. Understanding the genomics of this pathogen would be of vital importance both from a food and economic security standpoint and for the wider scientific interest.

The isolate identified as *Xav* NCPPB900 was sequenced using next generation sequencing technology and assembled into a draft genome sequence. The hope was to investigate these data to identify novel genomic components of this pathogen which have facilitated adaptation to this host. Further, by comparison to other *Xanthomonas* species known to infect the same host, to identify incidents of horizontal gene transfer which may have been responsible for this phenotypic convergence. The aim of this work was also to investigate the recent evolutionary history of this bacterial pathogen. This would provide a resource to enable genomic comparison with other xanthomonads; particularly pathovars of *X. axonopodis* and other more distantly related species such as *X. vasicola* and *X. sachari* which have evolved independently to colonise sugar cane

Author Contribution

The author conducted all bioinformatic analysis for this project. This included the production of bespoke scripts and pipeline code to prepare for and conduct the MLSA, quality control and trimming of sequencing reads, multiple rounds of *de-novo* assembly, including gapfilling and scaffolding steps. The author also carried out the alignment of the assembled contigs against the closely related *X. axonopodis citri* genome and a further round of gapfilling to produce the most contiguous version of the genome possible at the time. The author also carried out all post assembly sequence analysis and comparisons to known databases.

The author also contributed significantly to the pre-project research, concept design and planning for the project along with the writing, editing and submission of manuscript and the production and editing of all figures and tables.

Draft genome sequence of *Xanthomonas axonopodis* pathovar *vasculorum* NCPPB 900

James Harrison & David J. Studholme

Biosciences, University of Exeter, Exeter, Devon, UK

Correspondence: David J. Studholme, Biosciences, University of Exeter, Exeter EX4 4QD, Devon, UK. Tel.: +44 (0) 1392 724678; fax: +44 (0)1392 263434; e-mail: d.j.studholme@exeter.ac.uk

Received 19 September 2014; accepted 22 September 2014. Final version published online 13 October 2014.

DOI: 10.1111/1574-6968.12607

Editor: Hermann Bothe

Keywords
sugarcane; vascular pathogen; phosphonate biosynthesis; type-three secretion system; effectors.

Abstract

Xanthomonas axonopodis pathovar *vasculorum* strain NCPPB 900 was isolated from sugarcane on Reunion island in 1960. Consistent with its belonging to fatty-acid type D, multi-locus sequence analysis confirmed that NCPPB 900 falls within the species *X. axonopodis*. This genome harbours sequences similar to plasmids pXCV183 from *X. campestris* pv. *vesicatoria* 85-10 and pPHB194 from *Burkholderia pseudomallei*. Its repertoire of predicted effectors includes homologues of XopAA, XopAD, XopAE, XopB, XopD, XopV, XopZ, XopC and XopI and transcriptional activator-like effectors and it is predicted to encode a novel phosphonate natural product also encoded by the genome of the phylogenetically distant *X. vasicola* pv. *vasculorum*. Availability of this novel genome sequence may facilitate the study of interactions between xanthomonads and sugarcane, a host-pathogen system that appears to have evolved several times independently within the genus *Xanthomonas* and may also provide a source of target sequences for molecular detection and diagnostics.

Xanthomonas is a genus of *Gamma proteobacteria* that are predominantly pathogens of plants (Bradbury, 1986; Hayward, 1993). Many *Xanthomonas* species are comprised of several pathovars (pv.), each of which is highly specialised to infect a narrow host range, often a single plant species (Hayward, 1993). *Xanthomonas axonopodis* pv. *vasculorum* is an agent of gumming disease (Bradbury, 1986; Vauterin *et al.*, 1995; Dookun *et al.*, 2000), a vascular disease of sugarcane. In its host range and other phenotypic properties, *X. axonopodis* pv. *vasculorum* is similar to *X. vasicola* pv. *vasculorum* and there is potential for taxonomic confusion, with some strains having been classified at various times as *X. campestris* pv. *vasculorum*, *X. axonopodis* pv. *vasculorum* and/or *X. vasicola* pv. *vasculorum*. Molecular sequence analyses have recently confirmed that strains formerly classified as *X. campestris* pv. *vasculorum* fall within two major phylogenetic groupings that correspond to two distinct species: *X. axonopodis* and *X. vasicola* (Vauterin *et al.*, 1992, 1995; Rademaker *et al.*, 2005). Therefore, multi-locus sequence data can be used to unambiguously assign isolates to either *X. axonopodis* pv. *vasculorum* or *X. vasicola* pv. *vasculorum*. These sequence-based groupings also correlate with SDS-PAGE and fatty-acid profiles (Vauterin *et al.*, 1992; Dookun *et al.*, 2000). In the current study, we sequenced an isolate of *X. axonopodis* pv. *vasculorum* to enable genomic comparison with related pathovars of *X. axonopodis* and the much more distantly related xanthomonads such as *X. vasicola* pv. *vasculorum* and *X. sacchari* (Vauterin *et al.*, 1995) that have independently evolved the ability to colonise sugarcane.

The sequenced isolate of *X. axonopodis* pv. *vasculorum* is available from the National Collection of Plant Pathogenic Bacteria under accession NCPPB 900. It was collected from sugarcane (*Saccharum officinarum*) as strain B386 by A. C. Hayward at Ravine Creuse on Reunion island in 1960 (Hayward, 1962) at which time and place this pathogen was known to cause serious problems to the sugar cane industry (Dookun *et al.*, 2000). According to the classification scheme of Dookun and colleagues (Dookun *et al.*, 2000), it has fatty-acid type D, which along with type B (race 1) phylogenetically falls within the species *X. axonopodis*. Fatty-acid type D includes race 2 and race 3 isolates from Reunion island and Mauritius collected from sugarcane but also isolates collected from palms and broom bamboo (Dookun *et al.*, 2000). We performed multi-locus sequence analysis which confirmed that NCPPB 900 falls within the *X. axonopodis* clade (Fig. 1a) as was previously shown (Ah-You *et al.*, 2009) for strain LMG 8716, which is synonymous with this strain.

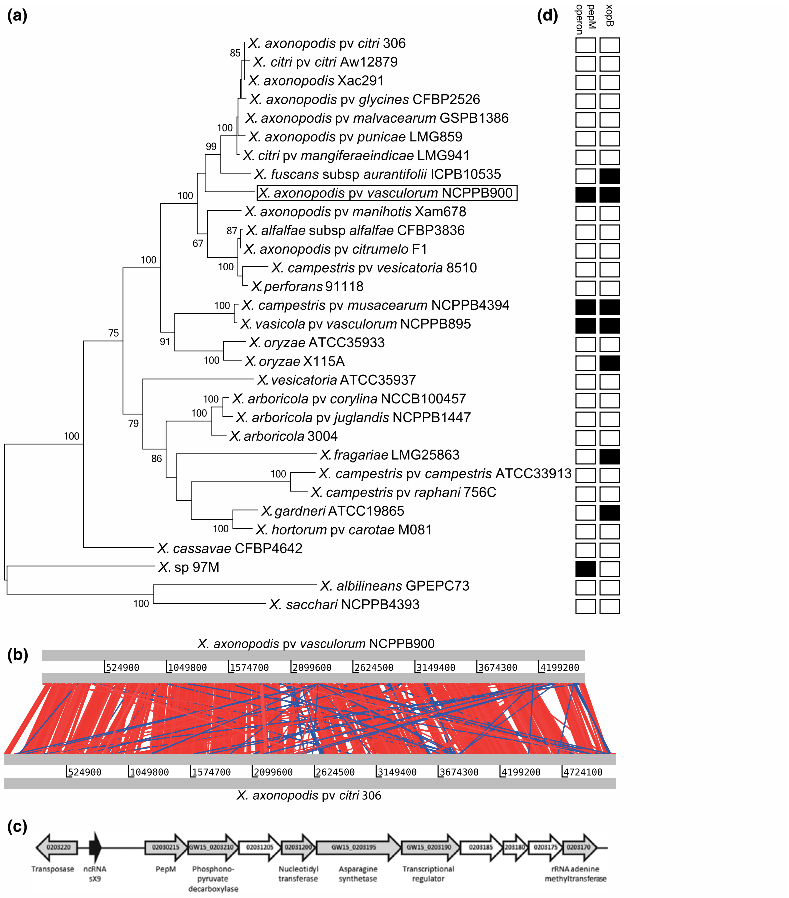

(a)
(d)
pepM operon
xopB
X. axonopodis pv citri 306
X. citri pv citri Aw12879
X. axonopodis Xac291
X. axonopodis pv glycines CFBP2526
X. axonopodis pv malvacearum GSPB1386
X. axonopodis pv punicae LMG859
X. citri pv mangiferaeindicae LMG941
X. fuscans subsp aurantifolii ICPB10535
X. axonopodis pv vasculorum NCPPB900
X. axonopodis pv manihotis Xam678
X. alfalfae subsp alfalfae CFBP3836
X. axonopodis pv citrumelo F1
X. campestris pv vesicatoria 8510
X.perforans 91118
X. campestris pv musacearum NCPPB4394
X. vasicola pv vasculorum NCPPB895
X. oryzae ATCC35933
X.oryzae X115A
X. vesicatoria ATCC35937
X. arboricola pv corylina NCCB100457
X. arboricola pv juglandis NCPPB1447
X. arboricola 3004
X.fragariae LMG25863
X. campestris pv campestris ATCC33913
X. campestris pv raphani 756C
X.gardneri ATCC19865
X. hortorum pv carotae M081
X. cassavae CFBP4642
X. sp 97M
X. albilineans GPEPC73
X. sacchari NCPPB4393
85
100
99
100
87
67
100
100
100
100
91
100
75
100
100
79
100
86
100
100
100
100
(b)
X. axonopodis pv vasculorum NCPPB900
524900
1049800
1574700
2099600
2624500
3149400
3674300
4199200
524900
1049800
1574700
2099600
2624500
3149400
3674300
4199200
4724100
X. axonopodis pv citri 306
(c)
0203220
02030215
GW15_0203210
02031205
02031200
GW15_0203195
GW15_0203190
0203185
203180
0203175
0203170
Transposase
ncRNA sX9
PepM
Phosphono-pyruvate decarboxylase
Nucleotidyl transferase
Asparagine synthetase
Transcriptional regulator
rRNA adenine methyltransferase

Fig. 1. The genome sequence assembly *of X. axonopodis* pv. *vasculorum* NCPPB 900. (a) shows the phylogenetic position of *X. axonopodis* pv. *vasculorum* based on multi-locus sequence analysis (MLSA) of six housekeeping genes (*atpD*, *dnaK*, *efP*, *fyuA*, *glnA*, *gyrB*). This maximum-likelihood tree was generated using the MEGA6 software and selecting the general time reversible model (Tamura *et al.*, 2013). Bootstrap values are shown as percentages; values below 60% are omitted. (b) indicates the presence (black) or absence (white) of *xopB* and *pepM* genes in each of the genome assemblies used in (a), based on BLASTN searches. (c) shows an alignment of the largest scaffold in the NCPPB 900 assembly aligned against the chromosome of *X. axonopodis* pv. *citri* 306 (Da Silva *et al.*, 2002) using BLASTN and visualised using the Artemis Comparison Tool (Carver *et al.*, 2005). (d) shows a putative operon in the genome of NCPPB 900 (contig scf_31858_1.1.contig_31) that is predicted to encode a phosphonate biosynthesis pathway. The operon comprises locus tags GW15_0203215 to GW15_0203270. Genes with predicted functions are shaded in grey while genes encoding hypothetical genes are shown in white. A nearby noncoding RNA gene (sX9) is indicated in black and a transposase gene in dark grey.

We generated 9 342 464 pairs of 100-bp sequence reads using the Illumina HiSeq 2500 and assembled them using VELVET version 1.2.10 (Zerbino & Birney, 2008). This yielded 322 contigs, which were scaffolded (using VELVET) into 150 scaffolds with a total length of 4 793 837 bp. The N_{50} for the contigs was 41 141 bp and for the scaffolds was 108 286 bp. We performed gap filling on the assembly using GAPCLOSER version 1.12-r6 (Luo *et al.*, 2012), which closed 116 gaps within the scaffolds resulting in an improved assembly containing only 206 contigs with an N_{50} of 64 290 bp. We performed an additional round of scaffolding by reordering the scaffolds (Rissman *et al.*, 2009) against the closed chromosomal sequence of the *X. axonopodis* pv. *citri* 306 chromosome (Da Silva *et al.*, 2002) and the sequence of the plasmid pXCV183 (Thieme *et al.*, 2005) generating a final set of 52 scaffolds, the longest one of which was 4.5 Mb in length and shared a high degree of co-linearity with the reference chromosome (Fig. 1b). These whole-genome shotgun data have been deposited at DDBJ/EMBL/GenBank under the accession JPHD00000000, both the original VELVET assembly and the updated version that underwent gap filling and further scaffolding. Contigs were annotated by the NCBI Prokaryotic Genome Annotation Pipeline 2.7 rev. 445095, predicting a total of 4109 genes including 3545 protein-coding sequences, 457 pseudogenes, 2 CRISPR arrays, 3 rRNAs, 49 tRNAs and 55 other non-coding RNAs.

One 140-kb scaffold (scf_31858_1.2) in the NCPPB 900 assembly shares 94% nucleotide sequence identity with the 182-kbp plasmid pXCV183 from the pepper-pathogen *X. campestris* pv. *vesicatoria* 85-10 (Thieme *et al.*, 2005), suggesting that NCPPB 900 harbours a plasmid similar to pXCV183. It also appears to contain at least one further plasmid: a 13.7-kbp contig (scf_31858_2.1.contig_1) shares 95% nucleotide sequence identity with the 13.4-kbp plasmid pPHB194 from a clinical isolate of the broad host-range pathogen *Burkholderia pseudomallei* (GenBank accession GQ401131).

The genome sequence of NCPPB 900 contains a further 11.5-kbp region that shares no detectable nucleotide sequence similarity with other *X. axonopodis* genomes (and therefore was likely acquired through horizontal transfer [scf_31858_1.1.contig_31]) and encodes a homologue of phosphoenolpyruvate mutase (PepM) that may be the first gene in an operon (see Fig. 1c) encoding a pathway for biosynthesis of a phosphonate, a structurally diverse class of natural product molecules containing carbon–phosphorus bonds that have interesting and useful biological properties (Yu *et al.*, 2013). The second protein encoded in the putative operon shares 42% amino acid sequence identity with the functionally characterised phosphonopyruvate decarboxylase from *Streptomyces hygroscopicus* (SwissProt: Q54271) (Nakashita *et al.*, 2000). Although this putative operon is not found in most other sequenced *Xanthomonas* genomes (Fig. 1d), it is conserved (90% nucleotide sequence identity) in all sequenced strains of *X. vasicola* pv. *vasculorum* and the closely related *X. vasicola* pv. *musacearum*. The operon might also present in the draft genome assembly of *Xanthomonas* species 97M, but the assembly (GenBank: AQPR01000000) is highly fragmented, and BLASTN hits to the NCPPB 900 sequence are scattered over several 97M contigs. The identity of the putative phosphonate product of this pathway is unknown, and the complement of genes is distinct from any of the operons catalogued by Yu and colleagues (Yu *et al.*, 2013) in their extensive survey of sequenced genomes.

In common with most other *Xanthomonas* species, NCPPB 900 encodes a type-III secretion system (TTSS) (White *et al.*, 2009). Its repertoire of predicted effectors includes homologues of XopAA, XopAD, XopAE, XopB, XopD, XopV, XopZ, XopC and XopI. This effector profile is distinct from other members of the species; for example, XopB is not encoded in any of the other 75 available *X. axonopodis* genome sequences (Fig. 1d). In common with other strains of *X. axonopodis* and *X. oryzae*, it also encodes homologues of the transcriptional activator-like (TAL) effectors (Scholze & Boch, 2011) (locus tags GW15_0222170, GW15_0222435, GW15_0222440) although these highly repetitive sequences are not fully resolved in the NCPPB 900 assembly.

In conclusion, this study presents the first genome sequence of a strain of *X. axonopodis* pv. *vasculorum*, which is phylogenetically distinct from other xanthomonads isolated from sugarcane (e.g. *X. vasicola* pv. *vasculorum*, *X. sacchari* and *X. albilineans*). Comparisons of these genomes might reveal insights into adaptation to this host plant or to patterns of horizontal gene transfer among the sugarcane microbiome. Availability of this sequence may facilitate studies on interactions between xanthomonads and sugarcane, a host-pathogen system that has apparently evolved several times independently within the genus *Xanthomonas* and may also provide a source of sequences for molecular detection and diagnostics.

Acknowledgements

J.H. was supported by a BBSRC PhD studentship. The authors are grateful to Audrey Fabros, Karen Moore and Konrad Paszkiewicz of the Exeter Sequencing Service facility, which is supported by Wellcome Trust Institutional Strategic Support Fund (WT097835MF), Wellcome Trust Multi User Equipment Award (WT101650MA) and BBSRC LOLA award (BB/K003240/1). The authors also thank Richard Thwaites and Melanie Sapp of the Food and Environment Research Agency (Fera) for assistance with growing the bacteria and preparation of genomic DNA and to Julian Smith (Fera) for helpful discussions.

References

Ah-You N, Gagnevin L, Grimont PA *et al.* (2009) Polyphasic characterization of xanthomonads pathogenic to members of the *Anacardiaceae* and their relatedness to species of *Xanthomonas*. *Int J Syst Evol Microbiol* **59**: 306–318.

Bradbury JF (1986) *Guide to Plant Pathogenic Bacteria*. CAB International, Slough.

Carver TJ, Rutherford KM, Berriman M, Rajandream M-A, Barrell BG & Parkhill J (2005) ACT: the artemis comparison tool. *Bioinformatics* **21**: 3422–3423.

Da Silva AC, Ferro JA, Reinach FC *et al.* (2002) Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. *Nature* **417**: 459–463.

Dookun A, Stead DE & Autrey LJ (2000) Variation among strains of *Xanthomonas campestris* pv. *vasculorum* from Mauritius and other countries based on fatty acid analysis. *Syst Appl Microbiol* **23**: 148–155.

Hayward AC (1962) Studies on Bacterial Pathogens of Sugar Cane. Part I. Differentiation of isolates of *Xanthomonas vasculorum*, with notes on an undescribed *Xanthomonas* sp. from sugar cane in Natal and Trinidad. *Mauritius Sugar Industry Research Institute Occasional Paper* **13**: 13–27.

Hayward AC (1993) The hosts of *Xanthomonas*. *Xanthomonas* 1993 1119 (Swings JG & Civerolo EL, eds), 25 pp. ref 1–19.

Luo R *et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* **1**: 18.

Nakashita H, Kozuka K, Hidaka T, Hara O & Seto H (2000) Identification and expression of the gene encoding phosphonopyruvate decarboxylase of *Streptomyces hygroscopicus*. *Biochim Biophys Acta* **1490**: 159–162.

Rademaker JLW, Louws FJ, Schultz MH, Rossbach U, Vauterin L, Swings J & de Bruijn FJ (2005) A comprehensive species to strain taxonomic framework for *Xanthomonas*. *Phytopathology* **95**: 1098–1111.

Rissman AI, Mau B, Biehl BS, Darling AE, Glasner JD & Perna NT (2009) Reordering contigs of draft genomes using the Mauve aligner. *Bioinformatics* **25**: 2071–2073.

Scholze H & Boch J (2011) TAL effectors are remote controls for gene activation. *Curr Opin Microbiol* **14**: 47–53.

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

Thieme F, Koebnik R, Bekel T *et al.* (2005) Insights into genome plasticity and pathogenicity of the plant pathogenic bacterium *Xanthomonas campestris* pv. *vesicatoria* revealed by the complete genome sequence. *J Bacteriol* **187**: 7254–7266.

Vauterin L, Yang P, Hoste B, Pot B, Swings J & Kersters K (1992) Taxonomy of xanthomonads from cereals and grasses based on SDS-PAGE of proteins, fatty acid analysis and DNA hybridization. *J Gen Microbiol* **138**: 1467–1477.

Vauterin L, Hoste B, Kersters K & Swings J (1995) Reclassification of *Xanthomonas*. *Int J Syst Bacteriol* **45**: 472–489.

White FF, Potnis N, Jones JB & Koebnik R (2009) The type III effectors of *Xanthomonas*. *Mol Plant Pathol* **10**: 749–766.

Yu X, Doroghazi JR, Janga SC, Zhang JK, Circello B, Griffin BM, Labeda DP & Metcalf WW (2013) Diversity and abundance of phosphonate biosynthetic genes in nature. *P Natl Acad Sci USA* **110**: 20759–20764.

Zerbino DR & Birney E (2008) VELVET: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res* **18**: 821–829.

References

1. Dookun, A, Stead, D. E. & Autrey, L. J. Variation among strains of *Xanthomonas campestris* pv. *vasculorum* from Mauritius and other countries based on fatty acid analysis. *Syst. Appl. Microbiol.* **23**, 148–155 (2000).

2. Vauterin, L., Hoste, B., Kersters, K. & Swings, J. Reclassification of *Xanthomonas*. *Int. J. Syst. Evol. Microbiol.* **45**, 472 (1995).

3. Vauterin, L. *et al.* Taxonomy of xanthomonads from cereals and grasses based on SDS-PAGE of proteins, fatty acid analysis and DNA hybridization. *J. Gen. Microbiol.* **138**, 1467–1477 (1992).

4. Rademaker, J. L. W. *et al.* A Comprehensive Species to Strain Taxonomic Framework for *Xanthomonas*. *Phytopathology* **95**, 1098–1111 (2005).

5. google maps - reunion. Available at: https://goo.gl/maps/VoPerPq33snNNC4v8. (Accessed: 6th July 2019)

Chapter 3:

Genome sequencing reveals a new lineage associated with lablab bean and genetic exchange between *Xanthomonas axonopodis* pv. *phaseoli* and *Xanthomonas fuscans* subsp. *fuscans*

Work from this chapter was published in:

Aritua V, Harrison J, Sapp M, Buruchara R, Smith J, Studholme DJ (2015) Genome sequencing reveals a new lineage associated with lablab bean and genetic exchange between *Xanthomonas axonopodis* pv. *phaseoli* and *Xanthomonas fuscans* subsp. *fuscans*. Front Microbiol 6(OCT):1–18.

This paper was cited by:

1. Tugume, J. K., Tusiime, G., Sekamate, A. M., Buruchara, R. & Mukankusi, C. M. Diversity and interaction of common bacterial blight disease-causing bacteria (*Xanthomonas* spp.) *with Phaseolus vulgaris* L. *Crop J.* **7**, 1–7 (2019).
2. Long, J.-Y. *et al.* Mutagenesis of PhaR, a Regulator Gene of Polyhydroxyalkanoate Biosynthesis of *Xanthomonas oryzae* pv. *oryzae* Caused Pleiotropic Phenotype Changes. *Front. Microbiol.* **9**, 1–10 (2018).
3. Ruh, M., Briand, M., Bonneau, S., Jacques, M. A. & Chen, N. W. G. *Xanthomonas* adaptation to common bean is associated with horizontal transfers of genes encoding TAL effectors. *BMC Genomics* **18**, 1–18 (2017).
4. Rai, K. K., Rai, N. & Rai, S. P. Recent advancement in modern genomic tools for adaptation of *Lablab purpureus L* to biotic and abiotic stresses: present mechanisms and future adaptations. *Acta Physiol. Plant.* **40**, 1–

29 (2018).

5. Jacques, M.-A. *et al.* Using Ecology, Physiology, and Genomics to Understand Host Specificity in *Xanthomonas*. *Annu. Rev. Phytopathol.* **54**, 163–187 (2016).

6. Midha, S. *et al.* Population genomic insights into variation and evolution of *Xanthomonas oryzae* pv. *oryzae*. *Sci. Rep.* **7**, 1–13 (2017).

Introduction.

This study focuses on the investigation of two pathovars of *Xanthomonas* which are known to be pathogenic on bean species. These pathovars cause common bacterial blight (CBB)[1], a devastating seed borne infection which presents a serious challenge to bean production in many African countries. *Xanthomonas axonopodis* pv. *phaseoli* (*Xap*) and *Xanthomonas fuscans* subsp. *fuscans* (*Xff*) both infect the common bean (*phaseolus vulgaris*) as their main host. However, these pathovars have also been known to infect the closely related species' lima bean (*Phaseolus lunatus*) and lablab bean (*Lablab purpureus*) among other legumes[2].

One of the major phenotypic characteristics which separates these strains is the production of brown pigment when cultured on tyrosine-containing media. Fuscous strains which produce the pigment are classified as *Xff* and those that do not produce the pigment (non-fuscous) are classified as *Xap*[3].

There is some taxonomic confusion surrounding these strains, as non-fuscous strains are classified within the *Xanthomonas axonopodis* species. However, the non-fuscous *Xff* have also been suggested to belong to a subclade of the same species or a distinct species in their own right[4].

There is also debate as to whether the strains isolated from the various bean hosts represent cases of single bacterial populations moving between hosts, or whether they form distinct genetic lineages or taxa. An interesting question as the identification of genomic variability and conservation between strains is an important part of the fight against bacterial pathogens. Identifying genetic markers and avirulence factors such as T3SS effectors can be used to inform the deployment of genetic resistance within the host plants in order to combat the spread of the disease caused by these pathogens[5].

As discussed previously, genetic exchange between bacterial lineages is commonplace and given the convergence of hosts for these pathogen classes, it is interesting to identify incidences of genetic exchange and horizontal transfer between distinct bacterial lineages sharing the same host.

The aim of this study was firstly to improve existing genomic resources of *Xanthomonas* bean pathogens, sequencing, assembling, annotating and analysing the 26 chosen strains. This information can then be used to investigate the recent evolutionary history of these phylogenetically diverse bean pathogens. To identify shared genomic features possibly responsible for adaptation to this evolutionary niche and possible virulence factors such as T3SS effectors. These features, which have likely been exchanged by horizontal gene transfer between these pathovars likely have facilitated pathogenicity on legumes. Further, to determine a core effector complement to inform the future deployment of genetic resistance genes in bean agriculture. Finally, to investigate intra-pathovar genetic variation within these strains attempting to identify sequence variations which can be exploited as molecular markers for use in epidemiological and phylogeographic studies

Author contributions

The author conducted all bioinformatic analysis for this project; this included using bespoke scripts and pipeline code for the quality control and trimming of sequencing reads and *de novo* assembly for each of the novel strains included in this project, the MLSA, SNP calling and all sequence analysis. The author also prepared, tested and optimised the innovative novel methodology for pan genome analysis used in this project.

The author also contributed significantly to the pre-project research, concept design and planning for the project along with the writing, editing and submission of manuscript and the production and editing of all figures and tables.

Manuscript

frontiers in Microbiology

ORIGINAL RESEARCH
published: 07 October 2015
doi: 10.3389/fmicb.2015.01080

Genome sequencing reveals a new lineage associated with lablab bean and genetic exchange between *Xanthomonas axonopodis* pv. *phaseoli* and *Xanthomonas fuscans* subsp. *fuscans*

Valente Aritua[1], James Harrison[2], Melanie Sapp[3], Robin Buruchara[4], Julian Smith[3] and David J. Studholme[2]*

[1] International Center for Tropical Agriculture, Kampala, Uganda, [2] Biosciences, University of Exeter, Exeter, UK, [3] Fera Science Ltd., York, UK, [4] Africa Regional Office, International Center for Tropical Agriculture, Consultative Group for International Agricultural Research (CGIAR), Nairobi, Kenya

OPEN ACCESS

Edited by:
Nicolas Denancé,
Institut National de la Recherche
Agronomique, France

Reviewed by:
Dennis Gross,
Texas A&M University, USA
Marie-Agnès Jacques,
Institut National de la Recherche
Agronomique, France

***Correspondence:**
David J. Studholme,
Biosciences, University of Exeter,
Stocker Road, Exeter EX4 4QD, UK
d.j.studholme@exeter.ac.uk

Specialty section:
This article was submitted to
Plant Biotic Interactions,
a section of the journal
Frontiers in Microbiology

Received: 30 May 2015
Accepted: 22 September 2015
Published: 07 October 2015

Citation:
Aritua V, Harrison J, Sapp M, Buruchara R, Smith J and Studholme DJ (2015) Genome sequencing reveals a new lineage associated with lablab bean and genetic exchange between Xanthomonas axonopodis pv. phaseoli and Xanthomonas fuscans subsp. fuscans. Front. Microbiol. 6:1080. doi: 10.3389/fmicb.2015.01080

Common bacterial blight is a devastating seed-borne disease of common beans that also occurs on other legume species including lablab and Lima beans. We sequenced and analyzed the genomes of 26 strains of *Xanthomonas axonopodis* pv. *phaseoli* and *X. fuscans* subsp. *fuscans*, the causative agents of this disease, collected over four decades and six continents. This revealed considerable genetic variation within both taxa, encompassing both single-nucleotide variants and differences in gene content, that could be exploited for tracking pathogen spread. The bacterial strain from Lima bean fell within the previously described Genetic Lineage 1, along with the pathovar type strain (NCPPB 3035). The strains from lablab represent a new, previously unknown genetic lineage closely related to strains of *X. axonopodis* pv. *glycines*. Finally, we identified more than 100 genes that appear to have been recently acquired by *Xanthomonas axonopodis* pv. *phaseoli* from *X. fuscans* subsp. *fuscans*.

Keywords: beans, *Phaseolus vulgaris*, *Phaseolus lunatus*, *Lablab purpureus*, *Dolichos lablab*, *Xanthomonas fuscans*, *Xanthomonas axonopodis*

Introduction

Common bacterial blight (CBB) is a devastating, widespread and seed-borne disease of common beans (*Phaseolus vulgaris*). The bacteria that cause CBB are genetically diverse (Gilbertson et al., 1991; Alavi et al., 2008; Parkinson et al., 2009; Fourie and Herselman, 2011) and include fuscous strains, which produce a brown pigment on tyrosine-containing medium, and non-fuscous strains. Currently, the non-fuscous strains are classified as *Xanthomonas axonopodis* pv. *phaseoli* (*Xap*) while the fuscous strains are classified into a different species as *X. fuscans* subsp. *fuscans* (*Xff*) (Schaad et al., 2005; Bull et al., 2012), though some authors consider the species *X. fuscans* to be a subclade within *X. axonopodis* (Rodriguez-R et al., 2012; Mhedbi-Hajri et al., 2013).

The main host of *Xap* and *Xff* is common bean (*Phaseolus vulgaris*) but they have also been isolated from the closely related Lima bean (*Phaseolus lunatus*) and lablab bean (*Lablab purpureus*, formerly *Dolichos lablab*) as well as several other legumes, including *Vigna* species (Bradbury, 1986). Lablab bean is a drought-resistant legume that stays green during the dry season and is used to improve soil and to feed livestock (Schaaffhausen, 1963). Lablab bean has been reported to be the main leguminous fodder crop used in Sudan around Khartoum, where it is known as hyacinth bean, bonavist bean or, in Arabic, lubia afin (Schaaffhausen, 1963). It is also grown in Sudan as a pulse legume (Mahdi and Atabani, 1992). Infection by *Xap* has been observed when lablab was sown during the rainy months (Tarr, 1958). It is not currently clear whether a single bacterial population moves frequently between host species or to what extent CBB agents colonizing different plant species represent distinct and genetically isolated populations or distinct taxa. For example, do the strains from Lima bean and lablab bean belong to the same genetic lineages as do strains from common bean?

A key determinant of pathogenicity in *Xanthomonas* species is the Hrp type-three secretion system (T3SS), which functions as a molecular syringe that secretes and translocates a number of bacterial effector proteins into the cytoplasm of the host cell, thereby modifying the host defenses to the advantage of the pathogen (Alfano and Collmer, 1997, 2004; Galán and Collmer, 1999; Grant et al., 2006; Kay and Bonas, 2009). Host ranges of *X. axonopodis* are significantly associated with the bacteria's repertoires of effectors and phylogenetically distinct strains' convergence on a common host plant might be at least partially explained by shared effectors (Hajri et al., 2009). The particular set of effectors expressed by a pathogen has some practical implications; many plant resistance genes trigger host defenses in response to detection of specific pathogen effectors. Effectors may act as virulence factors, enabling the pathogen to overcome host defenses. Therefore, rational deployment of available genetic resistance depends on knowledge of which effectors are likely to be present in a pathogen population. For example, it might be prudent to deploy resistance genes that recognize core effectors that are present in all strains of the pathogen that the plant will encounter rather than against rarely occurring effectors; this was the rationale for a recent study of the genome sequences of 65 strains of *Xanthomonas axonopodis* pv. *manihotis* (Bart et al., 2012).

CBB is currently a serious challenge to bean production in many African countries. In order to make optimal and rational use of limited available resources to contain and manage the impacts of this disease, it is important to understand the spread pathways of the *Xap* and *Xff* pathogens over both long and short geographical distances. Studies of spread rely on molecular markers that can be used to link strains from different times and locations based on their sharing similar genotypes.

According to multi-locus sequence analysis (MLSA), strains from *Phaseolus* species each fell into one of four genetic lineages (GL): GL 1, GL 2, GL 3, and GL *fuscans* (Mhedbi-Hajri et al., 2013) that corresponded to genetic lineages previously determined on the basis of amplified fragment length polymorphism (AFLP) (Alavi et al., 2008). The MLSA-based genetic lineages are consistent with an earlier classification of *X. axonopodis* strains into "genetic groups" based on conserved repetitive sequences BOX, enterobacterial repetitive intergenic consensus (ERIC), and repetitive extragenic palindromic (REP) (rep-PCR) (Rademaker et al., 2005), though the MLSA-based classification provides higher resolution. Rademaker's genetic group 9.4 includes GL 1, while genetic group 9.6 includes both GL 2 and GL 3 and GL *fuscans* (Mhedbi-Hajri et al., 2013), implying that GL 2 and GL 3 are more closely related to GL *fuscans* than to GL 1.

Whole-genome sequencing is relatively cheap, easy and quick and readily discovers genetic variation that can be utilized as neutral molecular markers to track specific genotypes (Vinatzer et al., 2014; Goss, 2015). It can also reveal biologically interesting variation and the incidence and distribution of avirulence factors (e.g., T3SS effectors) across the pathogen population allowing for rational deployment of genetic resistance in host crop plants, as was recently proposed for cassava and its pathogen *X. axonopodis* pv. *manihotis* (Bart et al., 2012). Other authors have pointed out that deployment of resistance without an awareness of pathogenic variation within the pathogen population could result in costly failure (Taylor et al., 1996; Fourie and Herselman, 2011). At the time of writing (July 2015) sequence assemblies are publicly available for 379 *Xanthomonas* genomes. No genome sequences were currently available for *Xap*, but two *Xff* genome sequences have been published: a finished genome for strain 4834-R (Darrasse et al., 2013b) and a draft assembly for strain 4844 (Indiana et al., 2014). A previous review (Ryan et al., 2011) presented some of the insights into *Xanthomonas* biology revealed by genome sequencing.

In the current study, we aimed to exploit whole-genome sequencing to catalog genetic diversity of CBB pathogens within each of the MLSA-based genetic lineages from common bean and strains from lablab and Lima beans. We also hypothesized that there might be some genetic features that are shared between phylogenetically distant lineages of CBB pathogens that reflect genetic exchange or adaptation to a common host. Therefore, we sequenced and bioinformatically analyzed the genomes of 26 strains deposited in the strain collections as *Xap* or *Xff* spanning six continents and more than four decades.

The specific objectives of this study were:

- To determine the phylogenomic relationships between *Xap* and *Xff* bacterial strains from different host species: common, lablab and Lima beans.
- To identify genetic variation within *Xap* or within *Xff*. These sequence variations could be exploited as molecular markers for use in epidemiological and phylogeographic studies.
- To determine patterns of conservation and variation in the complement of T3SS effectors between and within *Xap* and *Xff*. This knowledge can inform future rational deployment of disease resistance genes in beans.
- To identify genes or alleles that have been recently transmitted between phylogenetically divergent CBB bacteria.
- Identify candidate genetic determinants of the fuscous genotype, i.e., production of the brown pigment on tyrosine-containing medium.

Materials and Methods

Genome Sequencing

Genomic DNA was prepared from overnight liquid cultures of bacteria revived from the NCPPB grown on Yeast extract-Dextrose-Calcium Carbonate solid medium (i.e., agar plates) for 2 days at 28°C. DNA extraction was performed using the QIAamp DNA Mini kit (Qiagen, Hilden, Germany) applying proteinase K incubation for 30 min. We used the Nextera XT kit (Illumina, San Diego, USA) for library preparation following manufacturer's instructions. Purification was carried after tagmentation using AMPure XP beads (Beckman Coulter, High Wycombe, UK) prior to pooling. The 15pM library was then sequenced on an Illumina MiSeq using reagent kit chemistry v3 with 600 cycles.

Bioinformatics

Quality Control on Genomic Sequence Data

The quality of sequence data was checked using FastQC (Andrews)[1] Poor-quality and adaptor-containing reads were filtered and trimmed using FastQ-MCF (Aronesty, 2011).

Alignment of Sequence Reads vs. a Reference Genome Sequence

For alignment of genomic sequence reads against reference genome sequences of *Xff* 4834-R (Darrasse et al., 2013b) and *X. axonopodis* pv. *citri* 306 (Da Silva et al., 2002), we used BWA-MEM (Li, 2014). Resulting alignments were visualized using IGV (Thorvaldsdóttir et al., 2013).

Phylogenetic Analysis and Calling Single-nucleotide Variations

Phylogenetic analysis of the multi-locus sequence data was conducted in MEGA6 (Tamura et al., 2013). Multiple sequence alignments were performed using Muscle (Edgar, 2004). Evolutionary history was inferred using the maximum likelihood method based on the general time reversible model (Nei and Kumar, 2000). Initial tree(s) for the heuristic search were obtained by applying the Neighbor-Joining method to a matrix of pairwise distances estimated using the maximum composite likelihood (MCL) approach.

For phylogenetic analysis of whole-genome assemblies, we used the Parsnp program from the Harvest suite (Treangen et al., 2014). Phylogenetic trees generated from Parsnp in Newick format were imported into MEGA6 for preparation of the final figures. Parsnp uses FastTree2 to generate approximately maximum likelihood trees (Price et al., 2010). Distributions of single-nucleotide variations, calculated by Parsnp, were visualized using Gingr from the Harvest suite.

To check the reliability of the SNPs called by Parsnp, we further checked them using our previously described method (Mazzaglia et al., 2012; Wasukira et al., 2012; Clarke et al., 2015). For this method, we aligned the sequence reads against the reference genome sequence using BWA-mem version 0.7.5a-r405 (Li, 2013, 2014) with default parameter values and excluding any reads that did not map uniquely to a single site on the reference genome. From the resulting alignments, we generated pileup files using SAMtools version 0.1.19-96b5f2294a (Li et al., 2009). We then parsed the pileup-formatted alignments to examine the polymorphism status of each single-nucleotide site in the entire *Xff* 4834-R reference genome. For each single-nucleotide site we categorized it as either ambiguous or unambiguous. A site was considered to be un- ambiguous only if there was at least 5× coverage by genomic sequence reads from each and every bacterial strain and only if for each and every bacterial strain, at least 95% of the aligned reads were in agreement. Any sites that did not satisfy these criteria were considered to be ambiguous and excluded from further analysis. Over the remaining unambiguous sites, we could be very confident in the genotype for all the sequenced strains.

De Novo Assembly

Prior to assembly, we combined overlapping reads using FLASH (Magoč and Salzberg, 2011). Genomes were assembled using SPAdes version 3.5.0 (Bankevich et al., 2012) with read error correction and with the "- - careful" switch. We assessed the quality of the assemblies and generated summary statistics using Quast (Gurevich et al., 2013) and REAPR (Hunt et al., 2013).

Automated Annotation of Genome Assemblies

Genome assemblies were annotated via the Prokaryotic Genomes Automatic Annotation Pipeline (PGAAP) at the NCBI.

Comparison of Gene Content

To determine the presence or absence of genes in the newly sequenced genomes, we used alignment of genomic sequence reads against a reference pan-genome rather than comparison between genome assemblies. The reference pan-genome consisted of a set of gene sequences, each being a sole representative of a cluster of orthologous genes from all *Xanthomonas* genomes whose sequences were currently available; clustering of orthologous gene sequences was performed using UCLUST (Edgar, 2010). The reason for taking this approach (i.e., alignment of raw reads rather than alignment of assemblies) was to avoid potential errors arising from gaps in the genome assemblies. We aligned sequence reads against the reference genome sequence using BWA-MEM (Li and Durbin, 2009; Li, 2014) and used coverageBed from the BEDtools package (Quinlan and Hall, 2010) to determine the breadth of coverage of each gene in the resulting alignment. These breadths of coverage were visualized as heatmaps using the pheatmap module in R (R Development Core Team, 2013). We also compared genome assemblies using BRIG (Alikhan et al., 2011).

Results

Overview of Sequencing Results

We performed genomic re-sequencing on a collection of 26 *Xap* and *Xff* strains from the strain collections at NCPPB and CIAT as summarized in **Table 1**. For most of the strains, we obtained a depth of coverage of at least 40 x and thus were able

[1] Andrews, S. FastQC A Quality Control tool for High Throughput Sequence Data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ [Accessed May 13, 2015].

TABLE 1 | Sequenced strains.

Strain	Country and date of isolation	Host	Depth of coverage	Clade	Accession numbers
Xap NCPPB 556 (LMG 829)	Sudan (Shambat) 1957	Lablab bean	15 x	"Lablab"	JTJF00000000 SRX1048889
Xap NCPPB 557 (LMG 830)	Sudan (Wad-Medani) 1957	Lablab bean	57 x	"Lablab"	JWTE00000000 SRX1048890
Xap NCPPB 2064 (LMG 8015)	Sudan (Wad-Medani) 1965	Lablab bean	108 x	"Lablab"	JSEZ00000000 SRX1048891
Xap NCPPB 1713 (LMG 8013,)	Zimbabwe 1962	Lablab bean	30 x	"Lablab"	JWTD00000000 SRX1048892
"*Xap*" NCPPB 3660	Brazil 1975	Common bean	63 x	"Fuscans"	JSEX00000000 SRX1050058
Xff NCPPB 381* (LMG 826, CFBP 6165)	Canada 1957	Common bean	40 x	"Fuscans"	JTKK00000000 SRX1050059
Xff CIAT X621	South Africa (Cedan) 1995	Common bean	68 x	"Fuscans"	JXHS00000000 SRX1050082
Xff CIAT XCP631	Colombia 2004	Unknown	33 x	"Fuscans"	JXLW00000000 SRX1050252
Xff NCPPB 1056* (LMG 7457)	Ethiopia 1961	Common bean	44 x	"Fuscans"	JSEV00000000 SRX1049856
"*Xap*" NCPPB 1058**	Ethiopia 1961	Common bean	150 x	"Fuscans"	JSEY00000000 SRX1049857
Xff NCPPB 1433* (LMG 8016)	Hungary 1956	Common bean	51 x	"Fuscans"	JSBT00000000 SRX1049858
Xff NCPPB 2665* (LMG 841)	Italy 1973	Common bean	70 x	"Fuscans"	JSBQ00000000 SRX1049859
Xff NCPPB 1654* (LMG 837)	South Africa 1963	Common bean	93 x	"Fuscans"	JSBR00000000 SRX1049860
"Xap" NCPPB 670** (LMG 832)	Uganda 1958	Common bean	32 x	"Fuscans"	JRRE00000000 SRX1049872
Xff NCPPB 1402* (LMG 7459)	Uganda 1962	Common bean	17 x	"Fuscans"	JSEW00000000 SRX1049873
Xff NCPPB 1158* (LMG 7458)	UK 1961	Common bean	40 x	"Fuscans"	JSBS00000000 SRX1049874
Xff NCPPB 1495* (LMG 8017)	UK 1963	Common bean	30 x	"Fuscans"	JSEU00000000 SRX1049875
Xap NCPPB1646 (LMG 8011)	Australia 1964	Common bean	48 x	GL1	JTCT00000000 SRX1050299
Xap NCPPB 301	Canada pre-1951	Not known	50 x	GL1	JTCU00000000 SRX1050300
Xap CIAT XCP123	Colombia 1974	Lima bean	30 x	GL1	JXLV00000000 SRX1050292
Xap NCPPB 1420 (LMG 836)	Hungary 1956	Common bean	92 x	GL1	JTCV00000000 SRX1050301
Xap NCPPB 1811 (LMG 8014, CFBP 6164)	Romania 1966	Common bean	54 x	GL1	JWTF00000000 SRX1050302
Xap NCPPB 1680 (LMG 8012)	Tanzania 1964	Common bean	66 x	GL1	JWTG00000000 SRX1050303
Xap NCPPB 3035 (T) (LMG 7455, CFBP 6546)	USA pre-1978	Common bean	27 x	GL1	JSFA00000000 SRX1050305
Xap NCPPB 220 (NCTC 4331)	USA pre-1948	Not known	62 x	GL1	JWTH00000000 SRX1050306
Xap NCPPB 1138 (LMG 834)	Zambia 1961	Common bean	45 x	GL1	JWTI00000000 SRX1050323
"*Xap*" NCPPB 1128	Jamaica 1961	Common bean	80 x	Unknown *Xanthomonas* species	LFME00000000.1 SRX1090401

All strains had been deposited as Xap, except for those marked with an asterisk (*), which had been deposited as "X. axonopodis pv. phaseoli variant fuscans." Strains marked with two asterisks (**) were deposited as Xap but are reported to produce brown pigment, according to the accession cards that were submitted along with the strains into the NCPPB. Depth of coverage was estimated from alignments of raw sequence reads against the reference genome of X. axonopodis pv. citri 306 (Da Silva et al., 2002) using BWA-MEM (Li and Durbin, 2009; Li, 2014). GenBank accession numbers are given for the genome assemblies and SRA accession numbers are given for the raw sequence reads. Accession numbers are given for synonymous strains from the Belgian Coordinated Collections of Micro-Organisms (LMG), the Collection Française de Bactéries associées aux Plantes (CFBP) and National Collection of Type Cultures (NCTC).

to generate *de novo* genome assemblies. However, for seven of the genomes, there was less than 40 x coverage. We investigated the relationship between coverage depth and assembly quality by assembling subsets of the sequence reads from NCPPB 1058. We found that contig N_{50} length peaked at around 40 x coverage, with further increases in depth yielding little or no increase in contig lengths.

Figure 1 shows an overview of the *de novo* assemblies of each sequenced *Xap* and *Xff* genome aligned against that of the *X. axonopodis* pv. *citri* 306 (Da Silva et al., 2002). See also the Supplementary Figures for genome-wide alignments of the assemblies using Mauve (Darling et al., 2004). Note that the 26 *Xap* and *Xff* genomes were assembled *de novo*, using SPAdes (Bankevich et al., 2012), without use of a reference sequence.

The contiguities of the assemblies were comparable to those of previously sequenced *Xanthomonas* genomes. This is illustrated by the distribution of N_{50} contig lengths, which ranged from 39.4 to 123.6 kb. The range for a recent study of 65 *X. axonopodis* pv. *manihotis* was 7.4–111.0 kb (Bart et al., 2012). A full summary of assembly statistics, calculated using Quast (Gurevich et al., 2013), is provided in the Supplementary Table S1.

Contiguity of an assembly does not necessarily correlate with accuracy. Therefore, in addition to the Quast analysis of assembly contiguity, we also assessed the accuracies of the assemblies using REAPR (Hunt et al., 2013). This method is based on aligning to the assembly the sequence reads from which it was generated. This allows detection of anomalies in coverage of the assembly by reads and flags two classes of potential errors: fragment coverage distribution (FCD) errors and low fragment coverage errors. We compared the frequencies of these two classes of potential error for each of our genome assemblies and also for each of the 65 previously published *X. axonopodis* pv. *manihotis* assemblies

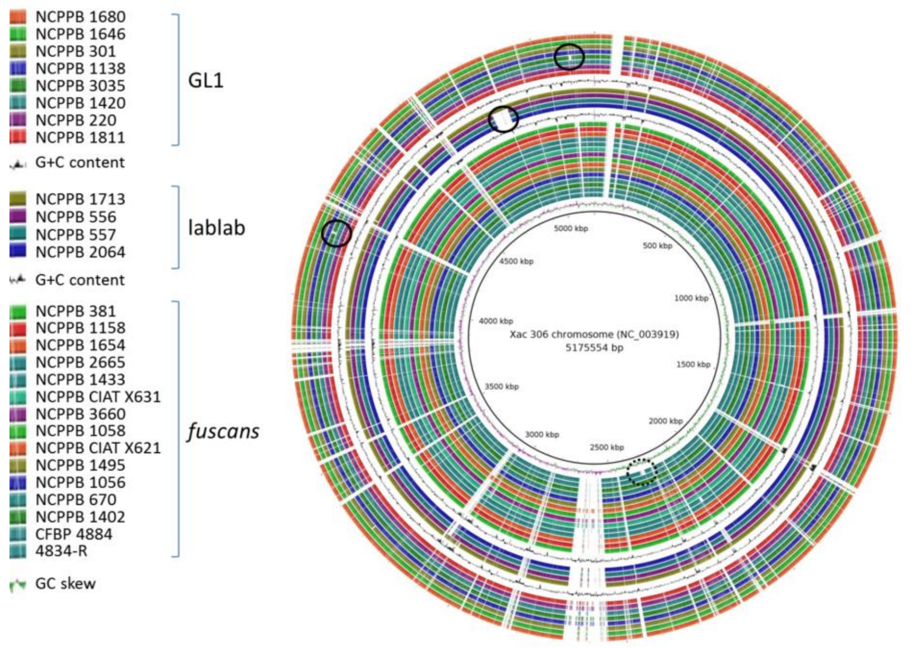


FIGURE 1 | Overview of genomic conservation among *Xap* and *Xff*. The newly sequenced genome sequences and those of previously sequenced Xff strains 4834-R (Darrasse et al., 2013b) and CFBP 4884 (Indiana et al., 2014) were aligned against the *Xac* 306 chromosome reference sequence (Da Silva et al., 2002) using BLASTN with an E-value threshold of 1 × 10^{-6}. The alignments are visualized using BLAST Ring Image Generator (BRIG) (Alikhan et al., 2011). The innermost ring indicates the position on the reference chromosome. Positions covered by BLASTN alignments are indicated with a solid color; whitespace gaps represent genomic regions not covered by the BLASTN alignments. For clarity, the three genetic lineages (GL 1, lablab-associated strains and GL *fuscans*) are separated by repetitions of the plot of G+C content. The black circle indicates the previously reported (Darrasse et al., 2013b) absence of the flagellar gene cluster in *Xff* 4834-R. The solid-lined black circles indicate strain-specific genomic deletions observed in the present study.

(Bart et al., 2012); see Supplementary Figure S1. The genome assemblies generated in the present study were of comparable quality to those from the previously published study. However, there is a general trend toward our genome assemblies having more "low fragment coverage" errors and fewer "FCD" errors.

To ascertain the phylogenetic positions of each sequenced strain, we initially used a multi-locus sequence analysis (MLSA) approach, using concatenated sequences from six genes that had been used in previous MLSA studies (Young et al., 2008; Almeida et al., 2010; Hajri et al., 2012; Hamza et al., 2012). This approach had the advantage that we could include in the analysis many *Xap* strains and other xanthomonads whose genomes had not been sequenced but for which MLSA data were available. Nucleotide sequences are available for these six genes from a large number of xanthomonads, either from whole-genome sequence assemblies or from the MLSA studies. We combined the publicly available sequences with homologous sequences extracted from the genomes newly sequenced for this study. The results of the MLSA revealed that the newly sequenced *Xap* and *Xff* genomes each fell into one of three distinct clades: GL 1, GL *fuscans* and a previously undescribed lineage associated with lablab bean (**Figure 2**).

The newly sequenced strains from lablab bean comprised a third clade, quite distinct from both *Xap* GL1 and from GL *fuscans* and indeed all previously described lineages of bean pathogens. The lablab-associated strains are closely related to members of Rademaker's genetic group 9.5, along with strains of pathovars *bilvae*, *citri*, *malvacearum,* and *mangiferaeindicae* that are pathogens of diverse plants including Bengal quince, *Citrus* spp., cotton and mango respectively (Bradbury, 1986; Rademaker et al., 2005). Also falling within this MLSA-based clade are strains of *X. axonopodis* pv. *glycines*, causative agent of bacterial pustule in soybean (Jones, 1987).

Genome-wide SNP Analysis Elucidates Phylogeny at Greater Resolution

Based on six-gene MLSA alone, strains could be ascribed to one of the three genetic lineages (GL 1, GL *fuscans,* and GL lablab). However, genome-wide sequence comparisons provided additional resolution and revealed distinct clades

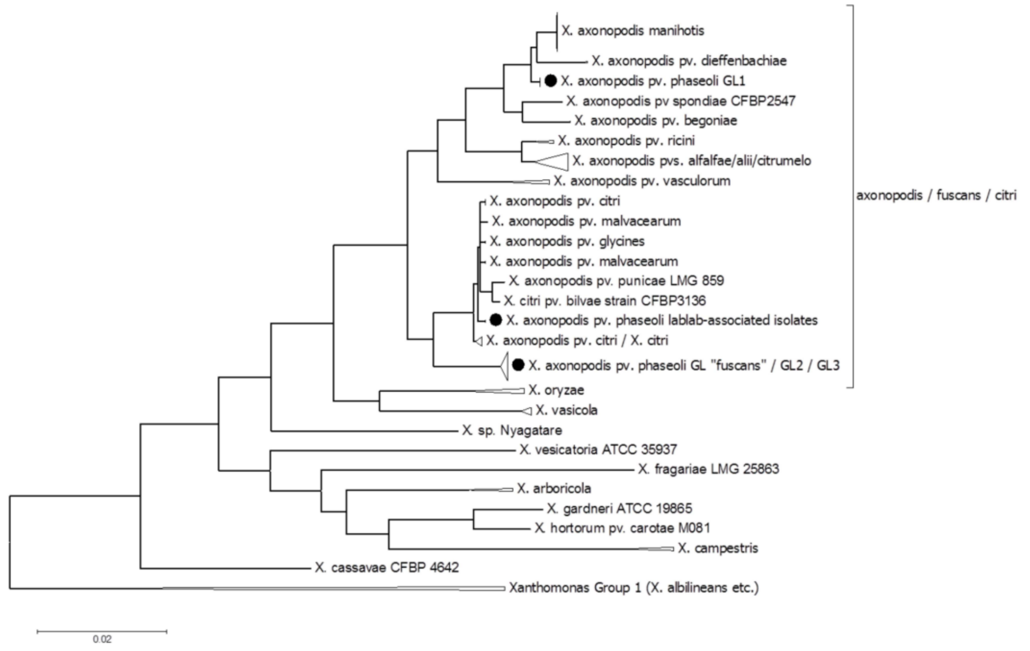

X. axonopodis manihotis
X. axonopodis pv. dieffenbachiae
X. axonopodis pv. phaseoli GL1
X. axonopodis pv spondiae CFBP2547
X. axonopodis pv. begoniae
X. axonopodis pv. ricini
X. axonopodis pvs. alfalfae/alii/citrumelo
X. axonopodis pv. vasculorum
X. axonopodis pv. citri
X. axonopodis pv. malvacearum
X. axonopodis pv. glycines
X. axonopodis pv. malvacearum
X. axonopodis pv. punicae LMG 859
X. citri pv. bilvae strain CFBP3136
X. axonopodis pv. phaseoli lablab-associated isolates
X. axonopodis pv. citri / X. citri
X. axonopodis pv. phaseoli GL "fuscans" / GL2 / GL3
X. oryzae
X. vasicola
X. sp. Nyagatare
X. vesicatoria ATCC 35937
X. fragariae LMG 25863
X. arboricola
X. gardneri ATCC 19865
X. hortorum pv. carotae M081
X. campestris
X. cassavae CFBP 4642
Xanthomonas Group 1 (X. albilineans etc.)
axonopodis / fuscans / citri
0.02

FIGURE 2 | Multi-locus sequence analysis to determine the phylogenetic positions of the sequenced strains within the species *X. axonopodis* and *X. fuscans*. The phylogenetic tree is based on alignment of six concatenated gene sequences (*atpD*, *dnaK*, *efp*, *fyuA*, *glnA*, and *gyrB*). The sequences for NCPPB 220, 301, 1138, 1420, 1646, 1680, 1811, 3035, and CIAT XCP123 were identical to those of CFBP 412, 6164, 6546, 6982, 6983, 6984, and 6985, which are classified as belonging to "pv. *phaseoli* GL1" (Alavi et al., 2008; Mhedbi-Hajri et al., 2013) and genetic group 9.4 (Rademaker et al., 2005). The sequences for NCPPB 381, 670, 1056, 1058, 1158, 1402, 1433, 1495, 1654, 2665, 3660, and CIAT X621 were identical to those of CFBP 1845, and 4834-R, which are classified as "pv. *phaseoli* GL *fuscans*" (Alavi et al., 2008; Mhedbi-Hajri et al., 2013) and genetic group 9.6 (Rademaker et al., 2005). The evolutionary history was inferred by using the Maximum Likelihood method based on the General Time Reversible model (Nei and Kumar, 2000). The tree with the highest log likelihood (−17100.2449) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree (s) for the heuristic search were obtained by applying the Neighbor-Joining method to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 284 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 2697 positions in the final dataset. Evolutionary analyses were conducted in MEGA6 (Tamura et al., 2013).

(or sub-lineages) within each lineage. **Figure 3A** illustrates the distribution of single-nucleotide variations across the reference sequence of the chromosome of *Xff* 4834-R for 160 publicly available related genome sequences. **Figure 3B** shows a phylogenetic reconstruction of those 160 genomes based on those variants. Consistent with the MLSA results, the strains sequenced in the present study similarly fell into three clades.

Within the *fuscans* lineage, the genome-wide comparison revealed at least three distinct sub-lineages, depicted in **Figure 4** in red, blue and green respectively. Each of these three lineages includes strains from diverse geographical locations and years. For example, one sub-lineage includes strains from France (1998), Hungary (1956), Italy (1963), South Africa (1963) and the UK (1962). This suggests that this sub-lineage has been circulating in Europe for nearly six decades and has spread between Europe and South Africa at least once, perhaps indirectly via another locality. This pattern is consistent with spread of the pathogen via global trade of seeds.

The genome-wide sequence analysis also reveals that multiple genetic lineages may be present within a single geographical area. For example, NCPPB 1056 and NCPPB 1058 were both isolated in the same country and the same year (Ethiopia, 1961) and fall into two distinct sub-lineages (**Figure 4**).

Similar, intra-lineage variation can be observed for strains within the lablab-associated strains (**Figure 5**) and GL 1 (**Figure 6**). Among lablab-associated strains, those collected in Sudan between 1957 and 1965 cluster together and are distinct from NCPPB 1713, which originates from Zimbabwe in 1962. Within GL 1, there are two multi-strain sub-lineages, which are indicated in blue and green in **Figure 5**. The former sub-lineage spans Australia, Canada, and Tanzania. The latter sub-lineage includes strains from Hungary, Romania, and the USA. Strain NCPPB 1138 (from Zambia, 1961) is distinct from both of these. The single GL 1 strain from Lima bean (CIAT XCP123, Colombia, 1974) is distinct from all of the

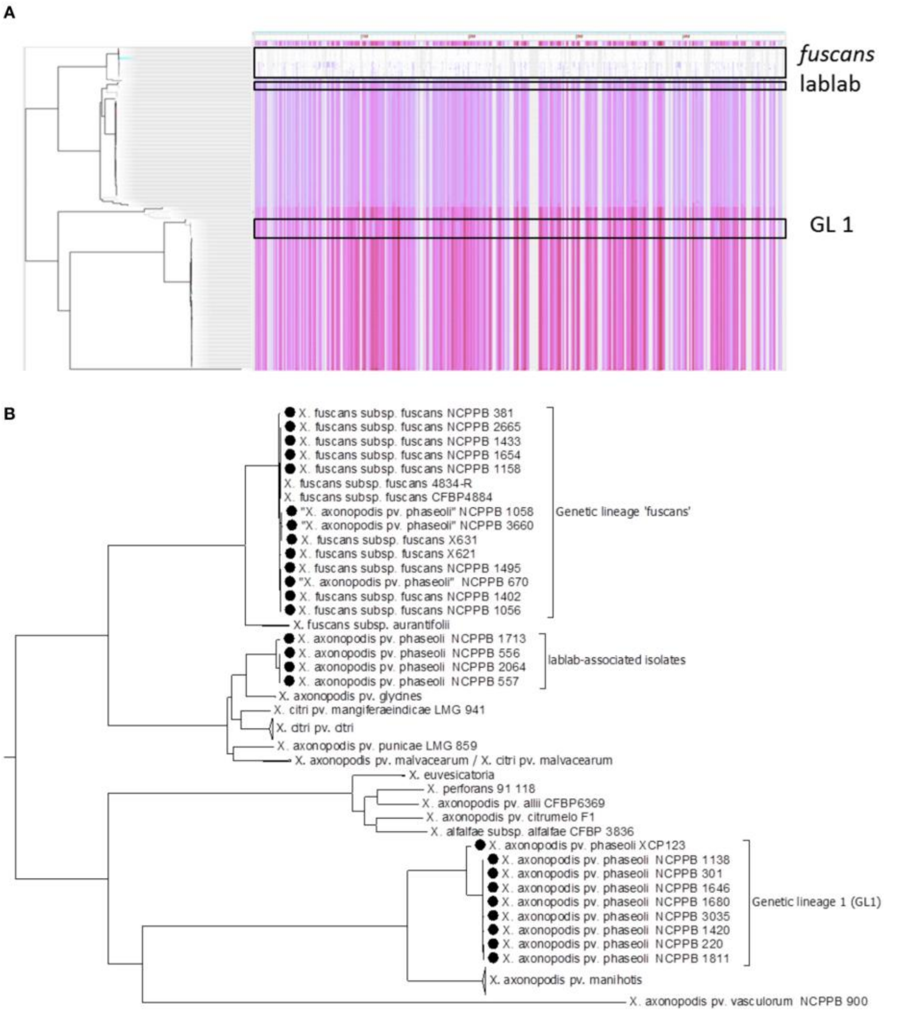


FIGURE 3 | Genome-wide identification of single-nucleotide variations. Panel **(A)** shows a density plot of single-nucleotide variations paired with a phylogenetic tree, both generated using Harvest (Treangen et al., 2014) with the chromosome of Xff 4834-R (Darrasse et al., 2013b) as the reference sequence. Panel **(B)** shows the same phylogenetic tree as in Panel **(A)** but with taxon labels and some clades collapsed for clarity. In addition to the 26 newly sequenced genomes, all 134 publicly available genome assemblies from *X. axonopodis*, *X. fuscans*, *X. citri,* and *X. euvesicatoria* were included.

strains from common bean (**Figure 5**); however, based on MLSA alone, it is indistinguishable from the other GL 1 strains.

Across the 4,981,995-bp chromosome sequence of *Xff* 4834-R, Harvest identified a total of 135,321 SNPs. This number includes all single-nucleotide sites that show variation between any of the 160 genome assemblies included in the analysis. A subset of 61,462 of those SNPs showed polymorphism among the 26 *Xap* and *Xff* genomes sequenced in the present study. The Harvest SNP calling takes as its input assembled genome sequences. Thus, substitution errors in the assemblies then could appear as false positives. Gaps in the assemblies are unlikely to generate false-positive SNP calls as Harvest only considers the core genome, i.e., those regions of the genome that are present in all of the genome assemblies and discards genomic regions present in only a subset of the assemblies. To assess the reliability of the Harvest SNP calls, we compared the results with a read-based method of SNP calling that we have used previously (Mazzaglia et al., 2012; Wasukira

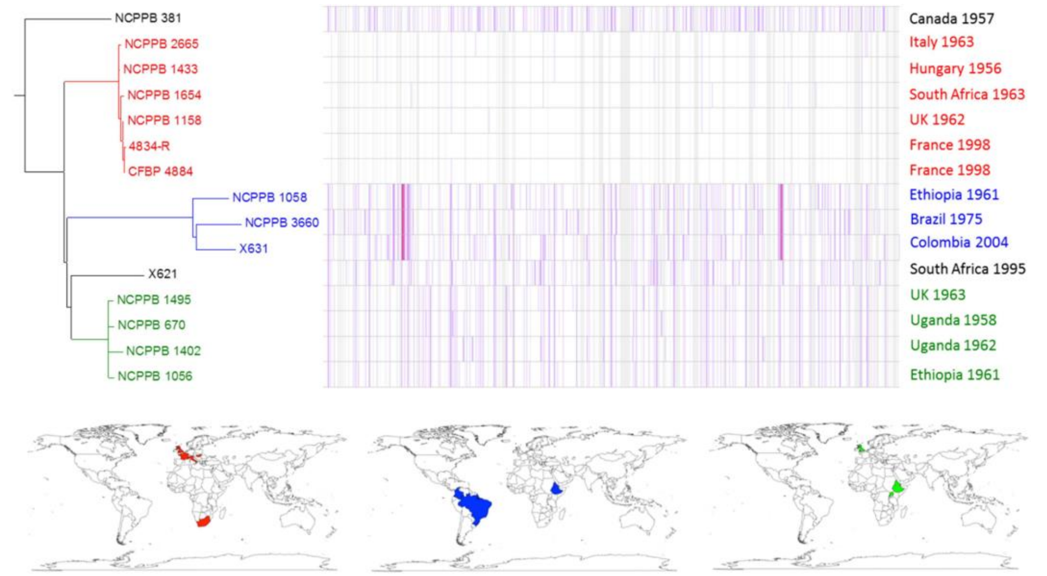


FIGURE 4 | Single-nucleotide variation within *fuscans* genetic lineage. A density plot of single-nucleotide variations is shown paired with a phylogenetic tree, both generated using Harvest (Treangen et al., 2014) with the chromosome of *Xff* 4834-R (Darrasse et al., 2013b) as the reference sequence. The country and year of isolation is indicated for each bacterial strain. Three multi-strain sub-lineages are indicated by coloring in red, blue, and green respectively. The geographical locations of the countries of isolation are indicated on the world maps for each of the three sub-lineages, again colored respectively in red, blue, or green.

et al., 2012; Clarke et al., 2015). Read-based methods have the advantage of not being reliant on assembly and they exploit the signal from multiple independent overlapping sequence reads at each site in the genome sequence. However, sequence reads are not available for the majority of *Xanthomonas* genome sequences, since for most studies only the assemblies and not the reads have been deposited in the public repositories. Of the 61,462 SNPs that Harvest called for the *Xap* and *Xff* genomes, our read-based method confirmed 53,811 (87.5%).

It is evident from **Figures 3–6** that single-nucleotide variations occur throughout the chromosome. However, the distribution is not uniform and there are several apparent "hotspots" of variation. The most likely explanation for these regions of higher-than-average sequence divergence is horizontal acquisition of genetic material from relatively distantly related strains. Such incongruent patterns of sequence similarity due to horizontal transfer have been reported previously in *Xanthomonas* species (Fargier et al., 2011; Hamza et al., 2012).

Gene-content Varies between and within Each Clade

Consistent with the indications of horizontal genetic transfer described in the previous section, we observed significant variations in gene presence and absence among strains within each of the three genetic lineages (**Figure 7**). Within the *fuscans* strains, there were 1188 clusters of orthologous genes that were present in at least one strain and absent from at least one other (**Figure 7A**). Among the lablab-associated strains, 472 orthologous gene clusters showed presence-absence polymorphism (**Figure 7B**). Among GL 1, the number was 535 (**Figure 7A**). Clustering of genomes according to gene content is broadly congruent with phylogeny. Supplementary Tables S2–S7 list genes whose presence distinguishes between Xff, Xap GL 1 and lablab-associated strains. Additionally, the four lablab-associated strains all contain six genes that have no close homologs amongst other sequenced xanthomonads. These are predicted to encode: three hypothetical proteins (KHS05433.1, KKY05378.1, and KHS05434.1), pilus assembly protein PilW (KHS05489.1), an oxidoreductase (KHS05432.1), and an epimerase (KHS05485.1).

Strain-specific Large Chromosomal Deletions

A large chromosomal deletion has been previously reported in *Xff* 4834-R in which a large part of the flagellar gene cluster is absent (Darrasse et al., 2013b). This deletion is visible in **Figure 1** at around position 2310 kb in the *Xac* 306 chromosome and indicated by a black circle with broken line. Although, similar deletions were reported in 5% of the strains tested (Darrasse et al., 2013b), this flagellar gene cluster was intact in all of the genomes sequenced in the current study as well as in the previously sequenced *Xff* 4884 (Indiana et al., 2014).

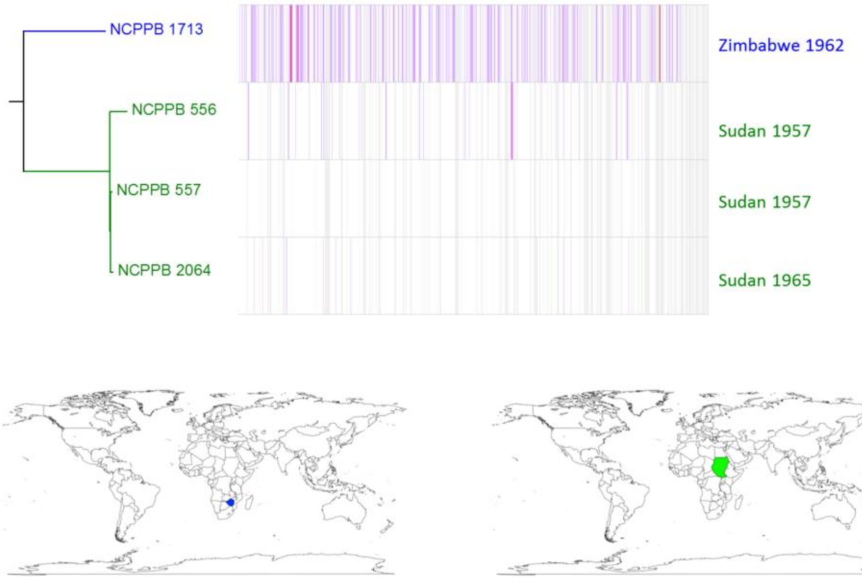


FIGURE 5 | Single-nucleotide variation among lablab-associated strains. A density plot of single-nucleotide variations is shown paired with a phylogenetic tree, both generated using Harvest (Treangen et al., 2014) with the chromosome of NCPPB 557 as the reference sequence. The country and year of isolation is indicated for each bacterial strain. Two sub-lineages are indicated by coloring in blue (a single strain) and red (three strains) respectively. The geographical locations of the countries of isolation are indicated on the world maps for each of the three sub-lineages, again colored respectively in blue or red.

In addition to the strain-specific flagellar deletion, **Figure 1** reveals several other large genomic deletions, examples of which are indicated with black circles. The largest example is a 50-kb region of the *Xac* 306 chromosome sequence that is absent from the three Sudanese lablab-associated strains but present in the Zimbabwe strain. This absence is visible in **Figure 1** at between 4.82 and 4.87 Mb on the reference chromosome sequence and indicated by a black circle. The absence of this region is supported not only by the *de novo* assemblies of NCPPB 556, 557 and 2064, but also by alignment of the raw sequence reads against the *Xac* 306 reference genome, eliminating the possibility that it merely represents an assembly artifact. This region is illustrated in Supplementary Figure S6, includes locus tags XAC4111–XAC4147 and is predicted to encode a type-6 secretion system (Darrasse et al., 2013b).

Other examples include a deletion of approximately 9 Kb that is deleted in *Xap* NCPPB 3035, resulting in loss of its ortholog of gene XAC RS17930 and parts of the two flanking genes XAC_RS17925 and XAC_RS17935 at around position 4.20 Mb on the reference genome (Supplementary Figure S7). A second example of a deletion unique to NCPPB 3035 spans approximately 10 kb at around position 5.10 Mb (Supplementary Figure S8). The deleted region contains genes XAC_RS21755 (predicted plasmid stabilization protein) to XAC_RS21815 (predicted transposase) and likely represents a mobile element.

Strains of *Xap* GL1 Encode a SPI-1-like T3SS

A previously published suppression subtractive hybridizations study comparing bean pathogens and closely related xanthomonads revealed the presence of genes encoding several protein components of a T3SS similar to that of *Salmonella* pathogenicity island 1 (SPI-1) in the genome of *Xap* CFBP 6164 (Alavi et al., 2008). This strain is synonymous with NCPPB 1811 and belongs to lineage GL 1. Subsequently, genome sequencing revealed that *X. albilineans* encodes a SPI-1-like T3SS (Pieretti et al., 2009, 2015) and targeted sequencing confirmed its presence in two further *Xap* GL 1 strains: CFBP 2534 (same as NCPPB 3035) and CFBP 6982 (Marguerettaz et al., 2011). Whole-genome sequencing in the current study indicated that this SPI-1-like T3SS was encoded in the genomes of all GL 1 strains from common bean and Lima bean (**Figure 8**) but was absent from GL fuscans and from the lablab-associated strains. All of the putative structural genes for the T3SS are conserved in *Xap* GL 1 but the *xapABCDEFGH* genes, hypothesized to encode effectors that are substrates of the T3SS in *X. albilineans* (Marguerettaz et al., 2011), are not conserved in *Xap*.

Repertoires of Hrp T3SS Effectors

Previous genome sequencing of *Xff* 4834-R revealed the presence of genes encoding 30 predicted effectors potentially secreted by the Hrp T3SS (Darrasse et al., 2013b). We searched for orthologs

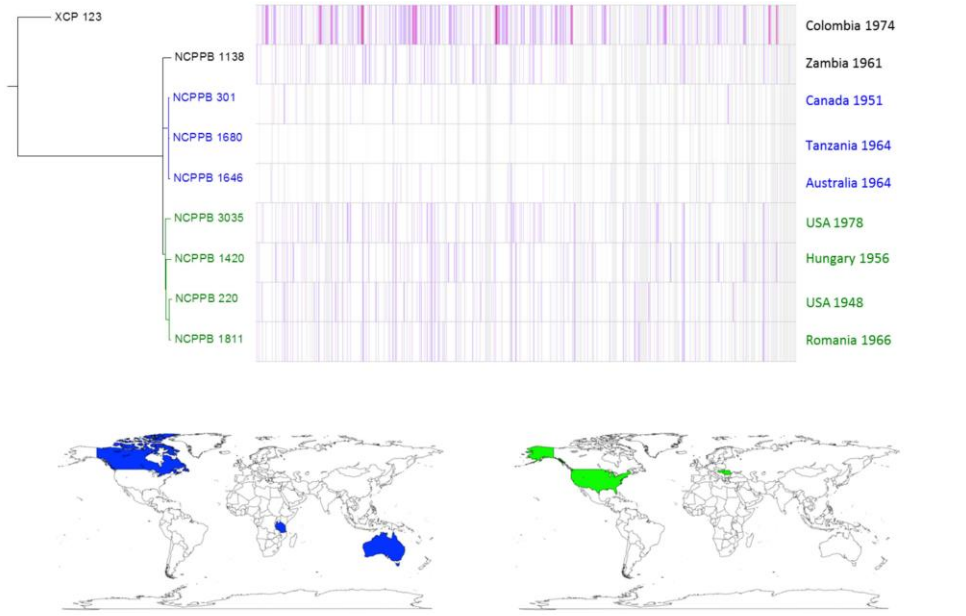


FIGURE 6 | Single-nucleotide variation within *Xap* GL 1. A density plot of single-nucleotide variations is shown paired with a phylogenetic tree, both generated using Harvest (Treangen et al., 2014) with the chromosome of NCPPB 1680 as the reference sequence. The country and year of isolation is indicated for each bacterial strain. Two multi-strain sub-lineages are indicated by coloring in blue and red respectively. The geographical locations of the countries of isolation are indicated on the world maps for each of the three sub-lineages, again colored respectively in blue or red.

of these and other *Xanthomonas* T3SS effectors in the newly sequenced *Xap* and *Xff* genomes using TBLASTN (Altschul et al., 1990) to search the genome assemblies against each protein query sequence. The results are summarized in **Figure 9**. There is a core set of 14 effectors that is encoded in all sequenced strains of *Xap* and *Xff*: XopK, XopZ, XopR, XopV, XopE1, XopN, XopQ, XopAK, XopA, XopL, AvrBs2, and XopX. Four of these are also included in the core set of effectors conserved among 65 strains of *X. axonopodis* pv. *manihotis* (Bart et al., 2012), namely XopE1, XopL, XopN, and XopV. Several others are encoded in most but not all of the newly sequenced genomes, for example: XopC1, XfuTAL2, and XopJ5. Others appear to be limited to just one of the three lineages. For example, XopF2 is limited to lineage *fuscans*, XopC2 is found only in *Xap* GL1 and XopAI is restricted to the lablab-associated strains.

The Molecular basis for Pigmentation

Some bacterial strains from CBB infections produce a brown pigment when grown in tyrosine-containing medium and are therefore described as "fuscous." The pigment is not believed to be directly associated with virulence (Gilbertson et al., 1991; Fourie, 2002) but fuscous strains tend to be very virulent on bean (Birch et al., 1997; Toth et al., 1998). The brown color arises from oxidized homogentisic acid (2,5 dihydroxyphenyl acetic acid), an intermediate in the tyrosine catabolic pathway that gets secreted and oxidized in these fuscous strains (Goodwin and Sopher, 1994). Genome sequencing of the fuscous strain *Xff* 4834-R revealed a single-nucleotide deletion in *hmgA*, the gene encoding homogentisate oxygenase (Darrasse et al., 2013b). This enzyme catalyzes a step in the tyrosine degradation pathway that converts tyrosine to fumarate and hence its inactivation likely disrupts tyrosine degradation leading to accumulation of homogentisate and its subsequent oxidation to form the brown pigment. Consistent with this hypothesis, we found that the single-nucleotide deletion was present in all of the sequenced strains belonging to GL *fuscans* resulting in a predicted protein product that is truncated, while the *hmgA* gene was intact in all of the *Xap* GL1 and lablab-associated *Xap* genomes (see **Figure 10**).

Recent Genetic Exchange between *Xap* GL 1 and GL Fuscans

Patterns of single-nucleotide variation (**Figure 3A**) revealed some regions of the genome where *Xap* GL 1 had many fewer variants with respect to the *Xff* 4384-R reference genome than did the closely related *X. axonopodis* pv. *manihotis*. Closer inspection revealed numerous genes where the *Xap* GL 1 strains shared an identical allele with *Xff*, a pattern that is incongruent with their relatively distant phylogenetic relationship.

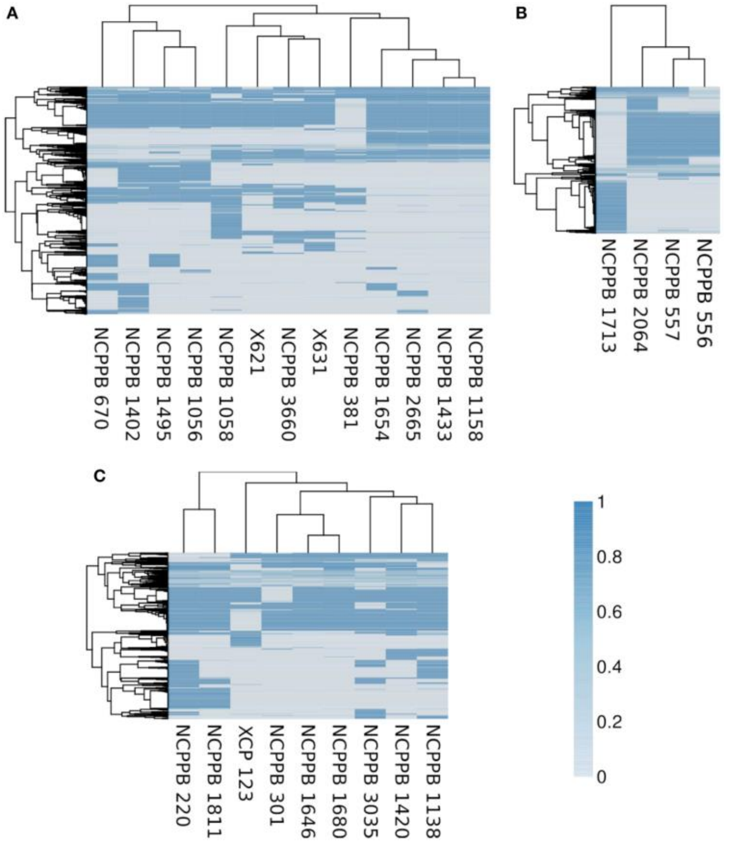


FIGURE 7 | Variation in gene content within each genetic lineage. The heatmaps essentially indicate the presence or absence of each gene in each sequenced genome. To determine the breadth of coverage of a gene, the genomic raw sequence reads are aligned against a reference pan-genome using BWA-MEM (Li, 2013, 2014) and the breadth of coverage is calculated using coverageBed (Quinlan and Hall, 2010). A coverage of one indicates complete coverage of the gene by aligned genomic sequence reads, indicating presence of the gene. A coverage of zero indicates that no genomic sequence reads matched the gene, indicating that it is absent (or at least highly divergent in sequence). In each heatmap, the genomes (columns) are clustered according to gene content and the genes (rows) are clustered according to their patterns of presence and absence across the genomes. The core genome, i.e., the subset of genes that are present in all strains, is excluded from the heatmap. **(A)** Summarizes the pattern of presence and absence for 1188 genes in the *fuscans* lineage, **(B)** for 472 genes in lablab-associated strains and **(C)** for 535 genes in *Xap* GL 1. Note that each gene is the representative of a cluster of orthologous genes.

To further investigate this phenomenon, we calculated pairwise nucleotide sequence identities for each *Xap* GL 1 gene vs. its closest homolog in other lineages within *X. axonopodis* and *X. fuscans*. The results are summarized in **Figure 11**. Pairwise sequence identities between *Xap* GL 1 and *Xff* (GL *fuscans*) followed a bimodal distribution with peaks at around 96% and at 100%. The peak at 100% was not observed for identities between *Xap* GL 1 and other lineages (*X. axonopodis* pv. *glycines*, *X. axonopodis* pv. *citri*, *X. axonopodis* pv. *manihotis*, *X. fuscans* subsp. *Aurantifolii,* and lablab-associated *Xap*). **Table 2** lists examples of genes with 100% identity between *Xap* GL 1 and *Xff*. Essentially the same set of genes is affected in all of the *Xap* GL 1 strains and the alleles are more similar to alleles from pathovars *citri* and *glycines* than to *manihotis*. Therefore, the most parsimonious explanation is that these alleles have been acquired by the ancestors of *Xap* GL 1 from the *fuscans* lineage.

Genome sequencing of *Xap* strains from lablab bean has revealed a previously unknown distinct lineage of *Xap*. This lineage is more closely related to strains of *X. axonopodis* pv. *glycines* that to any of the previously described genetic lineages of *Xap*. The existence of a separate lablab-associated lineage on lablab suggests that there may not be frequent movement of CBB bacteria between this species and common bean. However, conformation of this hypothesis will require genotyping of larger

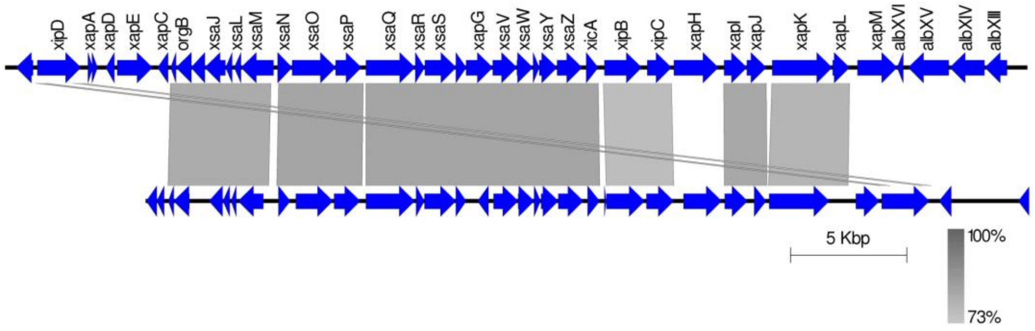


FIGURE 8 | The SPI-1-like T3SS in *Xap* GL 1. The figure shows a TBLASTN alignment of the *X. albilineans* chromosome (Pieretti et al., 2009) vs. the genome of *Xap* NCPPB 3035 (sequenced in the present study). Sequence identity is indicated by the black-gray-white color scale.

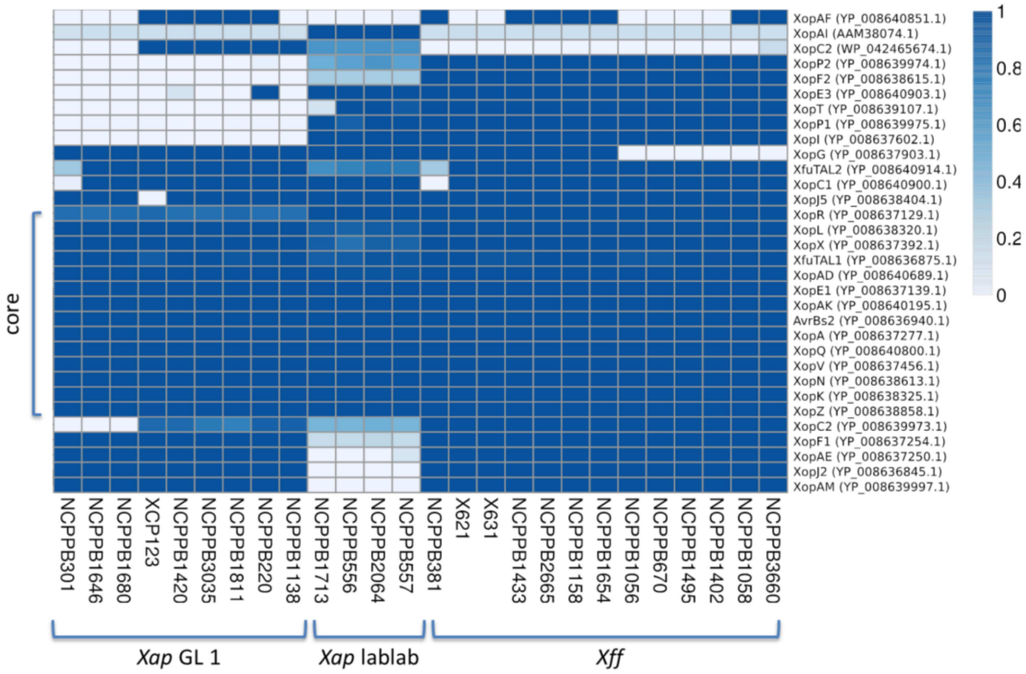


FIGURE 9 | Repertoires of T3SS effectors. The heatmap indicates the proportion covered by aligned genomic sequence reads over each T3SS effector gene DNA sequence. GenBank accession numbers are given on the right side. Genomic sequence reads were aligned against the effector gene sequences using BWA-MEM. A coverage of 1.0 represents complete coverage, indicating that the gene is present in the respective genome. A coverage of 0.0 represents a complete absence of aligned sequence reads, indicating absence of the gene from the respective genome.

numbers of strains; with the availability of these genome data it will be straightforward to design PCR-based assays to identify bacterial strains belonging to this newly discovered lineage.

It was previously observed that a *Xap* strain from common bean (NCPPB 302) was less pathogenic on lablab than bacteria isolated from naturally infected lablab (Sabet, 1959). The same

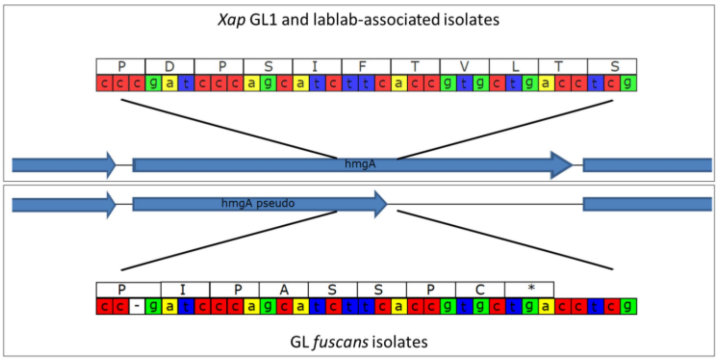

Xap GL1 and lablab-associated isolates
P D P S I F T V L T S
c c c g a t c c c a g c a t c t t c a c c g t g c t g a c c t c g
hmgA
hmgA pseudo
P I P A S S P C *
c c - g a t c c c a g c a t c t t c a c c g t g c t g a c c t c g
GL *fuscans* isolates

FIGURE 10 | Disruption of the *hmgA* gene in GL *fuscans*. The cartoon illustrates a sequence polymorphism in the hmgA gene whereby all of the sequenced GL *fuscans* strains (including previously sequenced 4834-R and CFBP4884) have a single-nucleotide deletion that results in a frame-shift and premature stop codon. All of the *Xap* GL 1 and lablab-associated strains encode a full-length protein product.

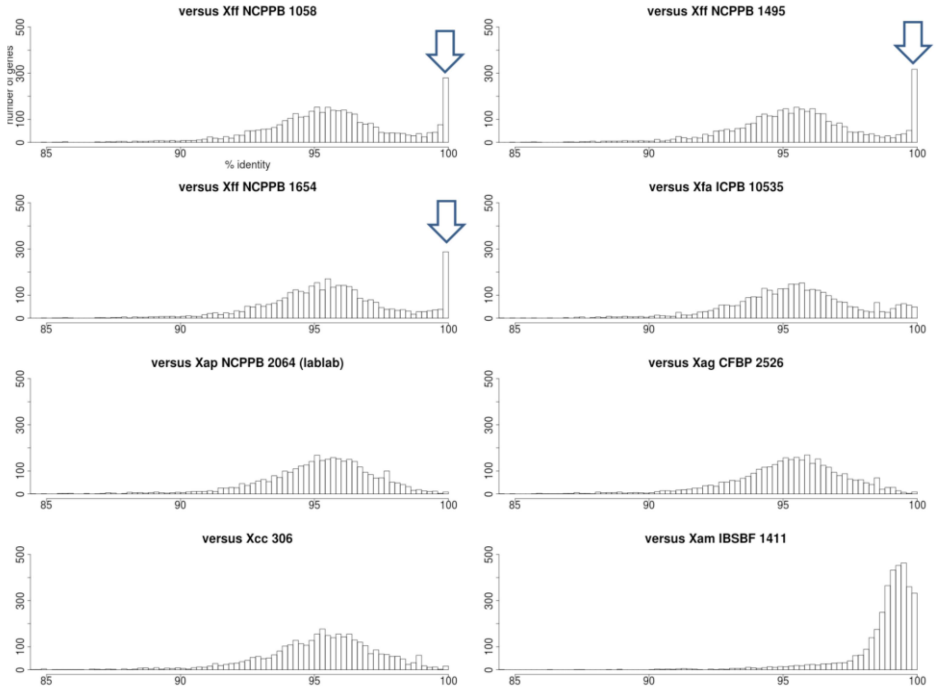


FIGURE 11 | Nucleotide sequence identities of *Xap* GL 1 genes vs. other lineages and pathovars. Each histogram shows the frequency distribution of sequence identity between genes from a *Xap* GL 1 strain (NCPPB 1680) and their closest BLASTN matches in other lineages including GL *fuscans* (NCPPB 1058, 1494 and 1654), lablab-associated *Xap* (NCPPB 2064), X. *fuscans* subsp. *aurantifolii* (ICPB 10535), *X. axonopodis* pv. *glycines* (CFBP 2526), *X. axonopodis* pv. *citri* (306) and *X. axonopodis* pv. *manihotis* (IBSBF 1411) (Da Silva et al., 2002; Moreira et al., 2010; Bart et al., 2012; Darrasse et al., 2013a). The arrowheads indicate the positions of the peaks at 100% sequence identity between GL 1 and GL *fuscans*.

TABLE 2 | Genes in *Xap* GL 1 that share 100% nucleotide sequence identity with *Xff* (GL *fuscans*).

GenBank accession number and predicted product of *Xap* NCPPB 1680	*Xag* CFBP 2526 (%)	*Xam* IBSBF 1411 (%)	*Xap* NCPPB2064 (%)	*Xfa* ICPB 10535 (%)
KHS20952 glutamine amidotransferase	97.73	99.47	97.61	98.14
KHS20955 preprotein translocase	97.28	94.42	98.88	99.20
KHS21044 phospholipase	96.36	93.73	96.49	99.75
KHS21241 ATPase AAA	97.92	93.84	97.92	98.67
KHS21242 histidine kinase	96.95	93.91	98.54	99.58
KHS21578 preprotein translocase subunit SecE	98.28	95.58	98.03	99.51
KHS21579 transcription antiterminator NusG	98.03	94.97	98.03	99.82
KHS21580 50S ribosomal protein L11	98.36	97.20	98.13	100.00
KHS21581 50S ribosomal protein L1	97.71	98.14	97.56	99.28
KHS21586 30S ribosomal protein S12	99.20	98.66	98.93	100.00
KHS22189 membrane protein	94.84	93.28	0.00	99.06
KHS22203 type VI secretion protein	97.93	96.69	0.00	99.17
KHS22241 UDP-2 3-diacylglucosamine hydrolase	98.25	95.42	98.12	99.46
KHS22843 cardiolipin synthetase	94.63	94.94	94.63	99.68
KHS23016 ribose-phosphate pyrophosphokinase	98.02	97.08	98.33	99.37
KHS23018 peptidyl-tRNA hydrolase	96.98	95.21	96.98	98.80
KHS23156 heat shock protein GrpE	99.23	95.56	99.23	99.81
KHS23251 membrane protein	98.83	97.96	98.98	99.56
KHS23252 membrane protein	97.67	94.25	97.67	99.46
KHS23254 methionine ABC transporter substrate-binding	97.78	95.64	97.78	98.75
KHS23255 metal ABC transporter permease	99.14	96.09	99.14	99.86
KHS23256 methionine ABC transporter ATP-binding protein	97.12	95.13	97.02	99.21
KHS23257 membrane protein	97.43	97.70	97.54	97.66
KHS23258 nucleotide-binding protein	98.56	96.91	98.14	99.38
KHS23269 type III secretion system effector protein XopAK	94.94	88.47	94.94	97.12
KHS23489 phosphomethylpyrimidine kinase	98.15	95.67	98.15	99.88
KHS23490 membrane protein	98.51	95.23	98.06	99.85
KHS23713 S-adenosylmethionine synthetase	99.01	98.02	98.92	99.83
KHS24197 molybdopterin-guanine dinucleotide biosynthesis	97.54	94.02	97.36	99.47
KHS24815 50S ribosomal protein L6	95.07	96.94	96.02	95.83
KHS24816 30S ribosomal protein S8	98.24	99.25	98.24	98.24
KHS24818 50S ribosomal protein L5	97.97	98.52	97.97	99.26
KHS24819 50S ribosomal protein L24	100.00	99.68	100.00	100.00
KHS24820 50S ribosomal protein L14	99.73	98.91	98.91	99.46
KHS24821 30S ribosomal protein S17	99.26	98.14	98.88	99.63
KHS24822 50S ribosomal protein L29	99.46	97.84	99.46	99.46
KHS25207 histone-like nucleoid-structuring protein	98.50	96.26	97.76	99.75
KHS25209 membrane protein	96.86	95.48	96.86	98.82
KHS25211 3-methyladenine DNA glycosylase	95.97	93.71	95.65	99.19
KHS25388 EF hand domain-containing protein	98.02	93.65	98.02	98.58
KHS25405 thioredoxin	99.12	97.36	98.83	100.00
KHS25407 ABC transporter	99.42	97.67	99.42	100.00
KHS25490 cupin	96.29	93.98	96.20	98.95
KHS25537 peptide ABC transporter permease	98.81	96.57	98.81	99.60
KHS26053 RNA-binding protein	98.35	96.69	99.17	99.59
KHS26541 3-demethylubiquinone-9 3-methyltransferase	97.47	98.61	97.47	97.61
KHS26908 lytic transglycosylase	95.53	86.24	94.99	99.11
KHS26912 membrane protein	97.87	96.17	97.77	99.36
KHS26916 malto-oligosyltrehalose trehalohydrolase	97.32	93.65	96.93	99.49
KHS27359 S-(hydroxymethyl)glutathione dehydrogenase	97.92	94.94	97.83	99.82

(Continued)

TABLE 2 | Continued

GenBank accession number and predicted product of *Xap* NCPPB 1680	*Xag* CFBP 2526 (%)	*Xam* IBSBF 1411 (%)	*Xap* NCPPB2064 (%)	*Xfa* ICPB 10535 (%)
KHS27633 Crp/Fnr family transcriptional regulator	95.80	94.35	95.66	99.47
KHS27636 ATP phosphoribosyltransferase	96.28	95.30	96.50	99.67
KHS27638 histidinol-phosphate aminotransferase	96.70	90.87	97.07	99.45
KHS27639 imidazoleglycerol-phosphate dehydratase	95.21	93.17	95.12	99.65
KHS27640 imidazole glycerol phosphate synthase	95.18	93.52	95.02	99.67
KHS28695 GntR family transcriptional regulator	97.79	97.51	97.79	99.17
KHS28696 vitamin B12 ABC transporter permease	98.20	96.41	98.38	99.28
KHS28992 NlpC-P60 family protein	94.19	91.30	94.48	97.88
KHS28993 ErfK/YbiS/YcfS/YnhG family protein	96.65	91.75	96.13	99.48
KHS29058 membrane protein	98.59	95.07	97.54	100.00
KHS29941 CDP-diacylglycerol–serine	99.22	97.02	98.97	99.74
KKY054115 membrane protein	96.92	94.62	94.62	99.23
KKY05325 membrane protein	100.00	93.75	100.00	100.00

Each of the Xap NCPPB 1680 genes shares 100% identity over at least 95% of its length with its ortholog in Xff strains NCPPB 1058, 1495, and 1654. For comparison, percentage nucleotide sequence identities are given for each gene vs. X. axonopodis pv. manihotis (Xam), lablab-associated X. axonopodis pv. phaseoli (Xap).

study also reported that the *Xap* strains (Dol1, 2 and 3) were less pathogenic on common bean than was *Xap* NCPPB 302, hinting at the presence of distinct populations of *Xap* differentially adapted to different host species. Furthermore, a subsequent study found that *Xap* strain Dol 3, isolated from lablab in Medani, Sudan, 1965, was pathogenic only on common bean and lablab bean; it was not pathogenic on any of the other leguminous plants that were tested, including several *Vigna* spp., *Rhynchosia memnonia*, mungo bean, pigeon pea, alfalfa, butterfly pea, velvet bean, pea, and white lupin.

To the best of our knowledge, no recent quantitative data are available for the extent and severity of common bacterial blight on lablab. However, in 1959, leaf blight on this crop was reported as widespread and often severe in the Gezira and central Sudan (Sabet, 1959).

The single sequenced bacterial strain from Lima bean clearly fell within *Xap* GL 1, along with strains from common bean, including the pathovar type strain (NCPPB 3035). However, genome-wide phylogenetic reconstruction revealed that the Lima-associated strain was the most early-branching within this lineage and suggests that it has been genetically isolated from the population that is geographically widely dispersed on common bean (**Figure 6**). Again, the availability of these genomic data will facilitate development of PCR-based assays to rapidly genotype larger panels of strains to elucidate the population genetics.

The newly sequenced genomes confirm and extend previous observations (Alavi et al., 2008; Marguerettaz et al., 2011; Egan et al., 2014), suggesting that a SPI-1-like T3SS is probably universal among *Xap* GL 1 but absent from *Xff* and from the newly discovered lablab-associated lineage. We also confirm that a frame-shift in the *hmgA* gene, resulting in a presumably defective homogentisate 1,2-dioxygenase, is common to all sequenced strains of *Xff* and probably explains the accumulation of brown pigment in fuscous strains (Darrasse et al., 2013b). The *hmgA* gene appeared to be intact in all the GL 1 and lablab-associated strains consistent with the absence of report of pigment in these.

Previous comparative genomics studies of *Xanthomonas* species have highlighted the presence of rearrangements of fragments of the genome (Qian et al., 2005; Darrasse et al., 2013a). We observed no evidence of such rearrangements among the *Xff*, *Xap* GL1 nor among the lablab-associated *Xap* genomes sequenced in the present study (See Supplementary Figures S3–S5). However, the lack of evidence should not be interpreted as meaning that there are no such rearrangements; draft-quality genome assemblies, such as those generated in the present other related studies (Bart et al., 2012; Indiana et al., 2014; Schwartz et al., 2015), are fragmented into multiple contigs and/or scaffolds and if the breakpoints in the genomic rearrangements coincide with gaps or breakpoints in the assembly, then they would not be detected.

A previous study reported large genomic deletions in about 5% of the examined *Xanthomonas* strains, including Xff 4834-R, resulting in loss of flagellar motility (Darrasse et al., 2013b). Although, none of the genomes sequenced in the present study displayed this deletion, there were several other strain-specific multi-kilobase deletions (see **Figure 1**) suggesting that this is a relatively common phenomenon among xanthomonads.

Discussion

In the present study, we sequenced the genomes of 26 strains of the causative agents of CBB, whose times and places of isolation spanned several decades and several continents. This resource adds to the already published genome sequences of *Xff* 4834-R and CFBP 4884 (Darrasse et al., 2013b; Indiana et al., 2014) with a further 13 sequenced genomes. We also present the first genome sequences for *Xap*, including 9 strains belonging to a previously described lineage known as GL 1. These 9 sequenced GL 1 strains include 8 from common bean and one from Lima bean. We also sequenced a further four strains from lablab bean.

Our data reveal genetic sub-lineages within *Xff* and within *Xap* GL 1, each having a widely dispersed geographical distribution. The availability of these genome sequence data will be a useful source of genetic variation for use in developing molecular markers for distinguishing individual sub-lineages or genotypes and thus aiding the study of routes of pathogen spread (Vinatzer et al., 2014; Goss, 2015). We observed considerable intra-lineage variation with respect to gene content as well as single-nucleotide variations (**Figures 4–7**).

Whole-genome sequencing revealed the repertoires of predicted T3SS effectors. Our results (**Figure 9**) were consistent with a previous survey of effector genes (Hajri et al., 2009) except for two apparent discrepancies. First, we find no evidence for presence of *avrRxo1* (*xopAJ*) in the genomes of *Xap* nor *Xff* though Hajri and colleagues found this gene in *Xap* GL 1. Second, genome-wide sequencing sequencing was able to distinguish between *xopF1* and *xopF2*. We find *xopF1* in both *Xff* and *Xap* GL 1 but find *xopF2* only in *Xff*. Hajri reported presence of *xopF2* in both *Xff* and *Xap* GL1; this might be explained by cross-hybrisisation of *xopF1* with the *xopF2* probes.

Arguably the most surprising finding to arise from the present study is the observation that Xff and Xap GL 1 share 100% identical alleles at dozens of loci even though on average most loci share only about 96% identity. This phenomenon is apparent from the bimodal distributions of sequence identities in **Figure 11**. This phenomenon is apparently restricted to sharing between GL 1 and Xff; no such bimodal distribution is seen between GL 1 and the lablab-associated strain not between GL 1 and *X. fuscans* subsp. *aurantifolii* (which is closely related to Xff). Furthermore, many of the alleles sharing 100% identity between GL 1 and *Xff* show significantly less identity between *Xap* GL 1 and *X. axonopodis* pv. *manihotis*, despite the close phylogenetic relationship between these last two. On the other hand, these shared sequences are more similar to sequences from *X. fuscans* subsp. *aurantifolii* than to sequence from *X. axonopodis* pv. *manihotis*, suggesting that they were acquired by *Xap* GL 1 from *Xff* rather than *vice versa*. Examples are listed in **Table 2**. It remains to be tested whether these alleles are adaptive for survival on a common ecological niche.

Acknowledgments

We are grateful for assistance from Jayne Hall, Erin Lewis and James Chisholm as well as the curators of NCPPB at Fera and Carlos Jara who kindly provided the CIAT strains and information about their origins. Genome sequencing was supported by the Canadian International Development Agency (Department of Foreign Affairs, Trade, and Development) grant to Pan-Africa Bean Research Alliance. VA was supported on this work by the BBSRC SCPRID Bean Grant BB/J011568/1. JH was supported by a BBSRC PhD studentship. DS was supported by BBSRC Grant BB/L012499/1.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2015.01080

References

Alavi, S. M., Sanjari, S., Durand, F., Brin, C., Manceau, C., and Poussier, S. (2008). Assessment of the genetic diversity of *Xanthomonas axonopodis* pv. phaseoli and *Xanthomonas fuscans* subsp. fuscans as a basis to identify putative pathogenicity genes and a type III secretion system of the SPI-1 family by multiple suppression subtractive h. *Appl. Environ. Microbiol.* 74, 3295–3301. doi: 10.1128/AEM.02507-07

Alfano, J. R., and Collmer, A. (1997). The type III (Hrp) secretion pathway of plant pathogenic bacteria: trafficking harpins, Avr proteins, and death. *J. Bacteriol.* 179, 5655–5662.

Alfano, J. R., and Collmer, A. (2004). Type III secretion system effector proteins: double agents in bacterial disease and plant defense. *Annu. Rev. Phytopathol.* 42, 385–414. doi: 10.1146/annurev.phyto.42.040103.110731

Alikhan, N.-F., Petty, N. K., Ben Zakour, N. L., and Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12:402. doi: 10.1186/1471-2164-12-402

Almeida, N. F., Yan, S., Cai, R., Clarke, C. R., Morris, C. E., Schaad, N. W., et al. (2010). PAMDB, a multilocus sequence typing and analysis database and website for plant-associated microbes. *Phytopathology* 100, 208–215. doi: 10.1094/PHYTO-100-3-0208

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Aronesty, E. (2011). *ea-utils?: Command-line Tools for Processing Biological Sequencing Data.* Available online at: http://code.google.com/p/ea-utils

Mahdi, A. A., and Atabani, I. M. A. (1992). Response of Bradyrhizobium-inoculated soyabean and lablab bean to inoculation with vesicular-arbuscular mycorrhizae. *Exp. Agric.* 28, 399. doi: 10.1017/S001447970002010X

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Bart, R., Cohn, M., Kassen, A., McCallum, E. J., Shybut, M., Petriello, A., et al. (2012). High-throughput genomic sequencing of cassava bacterial blight strains identifies conserved effectors to target for durable resistance. *Proc. Natl. Acad. Sci. U.S.A.* 109, E1972–E1979. doi: 10.1073/pnas.1208003109

Birch, P. R. J., Hyman, L. J., Taylor, R., Opio, A, F., Bragard, C., and Toth, I. K. (1997). RAPD PCR-based differentiation of *Xanthomonas campestris* pv. *phaseoli* and *Xanthomonas campestris* pv. *phaseoli* var. *fuscans*. *Eur. J. Plant Pathol.* 103, 809–814.

Bradbury, J. F. (1986). *Guide to Plant Pathogenic Bacteria.* CAB International Available online at: http://www.cabdirect.org/abstracts/19871324885.html; jsessionid=D5C161DFDC715BFA77A16BC4EE1BF1AE [Accessed July 19, 2014]

Bull, C. T., De Boer, S. H., Denny, T. P., Firrao, G., Fischer-Le Saux, M., Saddler, G. S., et al. (2012). List of new names of plant pathogenic bacteria (2008-2010). *J. Plant Pathol.* 94, 21–27. doi: 10.4454/JPP.V96I2.026

Clarke, C. R., Studholme, D. J., Hayes, B., Runde, B., Weisberg, A., Cai, R., et al. (2015). Genome-enabled phylogeographic investigation of the quarantine pathogen ralstonia solanacearum race 3 biovar 2 and screening for sources of resistance against its core effectors. *Phytopathology* 105, 597–607. doi: 10.1094/PHYTO-12-14-0373-R

Darling, A. C. E., Mau, B., Blattner, F. R., and Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14, 1394–1403. doi: 10.1101/gr.2289704

Darrasse, A., Bolot, S., Serres-Giardi, L., Charbit, E., Boureau, T., Fisher-Le Saux, M., et al. (2013a). High-quality draft genome sequences of *Xanthomonas* *axonopodis* pv. glycines Strains CFBP 2526 and CFBP 7119. *Genome Announc.* 1:e01036-13. doi: 10.1128/genomeA.01036-13

Darrasse, A., Carrère, S., Barbe, V., Boureau, T., Arrieta-Ortiz, M. L., Bonneau, S., et al. (2013b). Genome sequence of *Xanthomonas fuscans* subsp. fuscans strain 4834-R reveals that flagellar motility is not a general feature of xanthomonads. *BMC Genomics* 14:761. doi: 10.1186/1471-2164-14-761

Da Silva, A. C. R., Ferro, J. A., Reinach, F. C., Farah, C. S., Furlan, L. R., Quaggio, R. B., et al. (2002). Comparison of the genomes of two Xanthomonas pathogens with differing host specificities. *Nature* 417, 459–463. doi: 10.1038/417459a

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Egan, F., Barret, M., and O'Gara, F. (2014). The SPI-1-like Type III secretion system: more roles than you think. *Front. Plant Sci.* 5:34. doi: 10.3389/fpls.2014.00034

Fargier, E., Saux, M. F.-L., and Manceau, C. (2011). A multilocus sequence analysis of Xanthomonas campestris reveals a complex structure within crucifer-attacking pathovars of this species. *Syst. Appl. Microbiol.* 34, 156–165. doi: 10.1016/j.syapm.2010.09.001

Fourie, D. (2002). Distribution and severity of bacterial diseases on dry beans (Phaseolus vulgaris L.) in South Africa. *J. Phytopathol.* 150, 220–226. doi: 10.1046/j.1439-0434.2002.00745.x

Fourie, D., and Herselman, L. (2011). Pathogenic and genetic variation in *Xanthomonas axonopodis* pv. Phaseoli and its fuscans variant in Southern Africa. *African Crop Sci. J.* 19, 393–407. doi: 10.4314/acsj.v19i4

Galán, J. E., and Collmer, A. (1999). Type III secretion machines: bacterial devices for protein delivery into host cells. *Science* 284, 1322–1328.

Gilbertson, R. L., Otoya, M. M., Pastor-Corrales, M. A., and Maxwell, D. P. (1991). Genetic diversity in common blight bacteria is revealed by cloned repetitive DNA sequences. *Ann. Rep. Bean Impr. Conf.* 34, 37–38.

Goodwin, P. H., and Sopher, C. R. (1994). Brown pigmentation of Xanthomonas campestris pv. phaseoli associated with homogentisic acid. *Can. J. Microbiol.* 40, 28–34. doi: 10.1139/m94-005

Goss, E. M. (2015). Genome-enabled analysis of plant-pathogen migration. *Annu. Rev. Phytopathol.* 53, 121–135. doi: 10.1146/annurev-phyto-080614-115936

Grant, S. R., Fisher, E. J., Chang, J. H., Mole, B. M., and Dangl, J. L. (2006). Subterfuge and manipulation: type III effector proteins of phytopathogenic bacteria. *Annu. Rev. Microbiol.* 60, 425–449. doi: 10.1146/annurev.micro.60.080805.142251

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072–1075. doi: 10.1093/bioinformatics/btt086

Hajri, A., Brin, C., Hunault, G., Lardeux, F., Lemaire, C., Manceau, C., et al. (2009). A "repertoire for repertoire" hypothesis: repertoires of type three effectors are candidate determinants of host specificity in Xanthomonas. *PLoS ONE* 4:e6632. doi: 10.1371/journal.pone.0006632

Hajri, A., Brin, C., Zhao, S., David, P., Feng, J.-X., Koebnik, R., et al. (2012). Multilocus sequence analysis and type III effector repertoire mining provide new insights into the evolutionary history and virulence of Xanthomonas oryzae. *Mol. Plant Pathol.* 13, 288–302. doi: 10.1111/j.1364-3703.2011.00745.x

Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., and Otto, T. D. (2013). REAPR: a universal tool for genome assembly evaluation. *Genome Biol.* 14:R47. doi: 10.1186/gb-2013-14-5-r47

Indiana, A., Briand, M., Arlat, M., Gagnevin, L., Koebnik, R., Noël, L. D., et al. (2014). Draft Genome Sequence of the Flagellated *Xanthomonas fuscans* subsp. fuscans Strain CFBP 4884. *Genome Announc.* 2, e00966–14. doi: 10.1128/genomeA.00966-14

Jones, S. B. (1987). Bacterial pustule disease of soybean: microscopy of pustule development in a susceptible cultivar. *Phytopathology* 77:266. doi: 10.1094/Phyto-77-266

Kay, S., and Bonas, U. (2009). How Xanthomonas type III effectors manipulate the host plant. *Curr. Opin. Microbiol.* 12, 37–43. doi: 10.1016/j.mib.2008.12.006

Li, H. (2013). *Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM*. 3. Available online at: http://arxiv.org/abs/1303.3997 [Accessed July 20, 2014].

Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* 30, 1–9. doi: 10.1093/bioinformatics/btu356

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Magoč, T., and Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27, 2957–2963. doi: 10.1093/bioinformatics/btr507

Marguerettaz, M., Pieretti, I., Gayral, P., Puig, J., Brin, C., Cociancich, S., et al. (2011). Genomic and evolutionary features of the SPI-1 type III secretion system that is present in Xanthomonas albilineans but is not essential for xylem colonization and symptom development of sugarcane leaf scald. *Mol. Plant Microbe. Interact.* 24, 246–259. doi: 10.1094/MPMI-08-10-0188

Mazzaglia, A., Studholme, D. J., Taratufolo, M. C., Cai, R., Almeida, N. F., Goodman, T., et al. (2012). Pseudomonas syringae pv. actinidiae (PSA) strains from recent bacterial canker of kiwifruit outbreaks belong to the same genetic lineage. *PLoS ONE* 7:e36518. doi: 10.1371/journal.pone.0036518

Mhedbi-Hajri, N., Hajri, A., Boureau, T., Darrasse, A., Durand, K., Brin, C., et al. (2013). Evolutionary history of the plant pathogenic bacterium *Xanthomonas axonopodis*. *PLoS ONE* 8:e58474. doi: 10.1371/journal.pone.0058474

Moreira, L. M., Almeida, N. F., Potnis, N., Digiampietri, L. A., Adi, S. S., Bortolossi, J. C., et al. (2010). Novel insights into the genomic basis of citrus canker based on the genome sequences of two strains of *Xanthomonas fuscans* subsp. aurantifolii. *BMC Genomics* 11:238. doi: 10.1186/1471-2164-11-238

Nei, M., and Kumar, S. (2000). *Molecular Evolution and Phylogenetics.* Oxford University Press. Available online at: https://books.google.com/books?hl=en&lr=&id=vtWW9bmVd1IC&pgis=1 [Accessed May 13, 2015].

Parkinson, N., Cowie, C., Heeney, J., and Stead, D. (2009). Phylogenetic structure of Xanthomonas determined by comparison of gyrB sequences. *Int. J. Syst. Evol. Microbiol.* 59, 264–274. doi: 10.1099/ijs.0.65825-0

Pieretti, I., Pesic, A., Petras, D., Royer, M., Süssmuth, R. D., and Cociancich, S. (2015). What makes Xanthomonas albilineans unique amongst xanthomonads? *Front. Plant Sci.* 6:289. doi: 10.3389/fpls.2015.00289

Pieretti, I., Royer, M., Barbe, V., Carrere, S., Koebnik, R., Cociancich, S., et al. (2009). The complete genome sequence of Xanthomonas albilineans provides new insights into the reductive genome evolution of the xylem-limited Xanthomonadaceae. *BMC Genomics* 10:616. doi: 10.1186/1471-2164-10-616

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5:e9460. doi: 10.1371/journal.pone.0009490

Qian, W., Jia, Y., Ren, S. X., He, Y. Q., Feng, J., Lu, L., et al. (2005). Comparative and functional genomic analyses of the pathogenicity of phytopathogen. *Genome Res.* 15, 757–767. doi: 10.1101/gr.3378705

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq,033

Rademaker, J. L. W., Louws, F. J., Schultz, M. H., Rossbach, U., Vauterin, L., Swings, J., et al. (2005). A comprehensive species to strain taxonomic framework for xanthomonas. *Phytopathology* 95, 1098–1111. doi: 10.1094/PHYTO-95-1098

R Development Core Team, R. (2013). R: a language and environment for statistical computing. *R Found. Stat. Comput.* 1, 409. doi: 10.1007/978-3-540-74686-7

Hamza, A. A., Robene-Soustrade, I., Jouen, E., Lefeuvre, P., Chiroleu, F., Fisher-Le Saux, M., et al. (2012). MultiLocus sequence analysis- and amplified fragment length polymorphism-based characterization of xanthomonads associated with bacterial spot of tomato and pepper and their relatedness to Xanthomonas species. *Syst. Appl. Microbiol.* 35, 183–190. doi: 10.1016/j.syapm.2011.12.005

Rodriguez-R, L. M., Grajales, A., Arrieta-Ortiz, M., Salazar, C., Restrepo, S., and Bernal, A. (2012). Genomes-based phylogeny of the genus Xanthomonas. *BMC Microbiol.* 12:43. doi: 10.1186/1471-2180-12-43

Ryan, R. P., Vorhölter, F.-J., Potnis, N., Jones, J. B., Van Sluys, M.-A., Bogdanove, A. J., et al. (2011). Pathogenomics of Xanthomonas: understanding bacterium-plant interactions. *Nat. Rev. Microbiol.* 9, 344–355. doi: 10.1038/nrmicro2558

Sabet, K. A. (1959). Studies in the bacterial diseases of Sudan crops III. On the occurrence, host range and taxonomy of the bacteria causing leaf blight diseases of certain leguminous plants. *Ann. Appl. Biol.* 47, 318–331.

Schaad, N. W., Postnikova, E., Lacy, G. H., Sechler, A., Agarkova, I., Stromberg, P. E., et al. (2005). Reclassification of Xanthomonas campestris pv. citri (ex Hasse 1915) Dye 1978 forms A, B/C/D, and E as X. smithii subsp. citri (ex Hasse) sp. nov. nom. rev. comb. nov., X. fuscans subsp. aurantifolii (ex Gabriel 1989) sp. nov. nom. rev. comb. nov., and X. *Syst. Appl. Microbiol.* 28, 494–518. doi: 10.1016/j.syapm.2005.03.017

Schaaffhausen, R. V. (1963). Dolichos lablab or hyacinth bean: - Its uses for feed, food and soil improvement. *Econ. Bot.* 17, 146–153. doi: 10.1007/BF02985365

Schwartz, A. R., Potnis, N., Timilsina, S., Wilson, M., Patane, J., Martins, J. J., et al. (2015). Phylogenomics of Xanthomonas field strains infecting pepper and tomato reveals diversity in effector repertoires and identifies determinants of host specificity. *Front. Microbiol.* 6:535. doi: 10.3389/fmicb.2015.00535

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst,197

Tarr, S. A. J. (1958). Experiments in the Sudan Gezira on control of wilt of dolichos bean (Dolichos lablab) associated with attack by cockshafer grubs (Scizonya sp.). *Ann. Appl. Biol.* 46, 630–638. doi: 10.1111/j.1744-7348.1958.tb02246.x

Taylor, J. D., Teverson, D. M., Allen, D. J., and Pastor Corrales, M. A. (1996). Identification and origin of races of Pseudomonas syringae pv. phaseolicola from Africa and other bean growing areas. *Plant Pathol.* 45, 469–478. doi: 10.1046/j.1365-3059.1996.d01-147.x

Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings Bioinforma.* 14, 178–192. doi: 10.1093/bib/bbs017

Toth, I. K., Hyman, L. J., Taylor, R., and Birch, P. R. J. (1998). PCR-based detection of Xanthomonas campestris pv. phaseoli var. fuscans in plant material and its differentiation from X. c. pv. phaseoli. *J. Appl. Microbiol.* 85, 327–336. doi: 10.1046/j.1365-2672.1998.00514.x

Treangen, T. J., Ondov, B. D., Koren, S., and Phillippy, A. M. (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol.* 15:524. doi: 10.1186/s13059-014-0524-x

Vinatzer, B. A., Monteil, C. L., and Clarke, C. R. (2014). Harnessing population genomics to understand how bacterial pathogens emerge, adapt to crop hosts, and disseminate. *Annu. Rev. Phytopathol.* 52, 19–43. doi: 10.1146/annurev-phyto-102313-045907

Wasukira, A., Tayebwa, J., Thwaites, R., Paszkiewicz, K., Aritua, V., Kubiriba, J., et al. (2012). Genome-wide sequencing reveals two major sub-lineages in the genetically monomorphic pathogen xanthomonas campestris pathovar musacearum. *Genes (Basel)* 3, 361–377. doi: 10.3390/genes3030361

Young, J. M., Park, D.-C., Shearman, H. M., and Fargier, E. (2008). A multilocus sequence analysis of the genus Xanthomonas. *Syst. Appl. Microbiol.* 31, 366–377. doi: 10.1016/j.syapm.2008.06.004

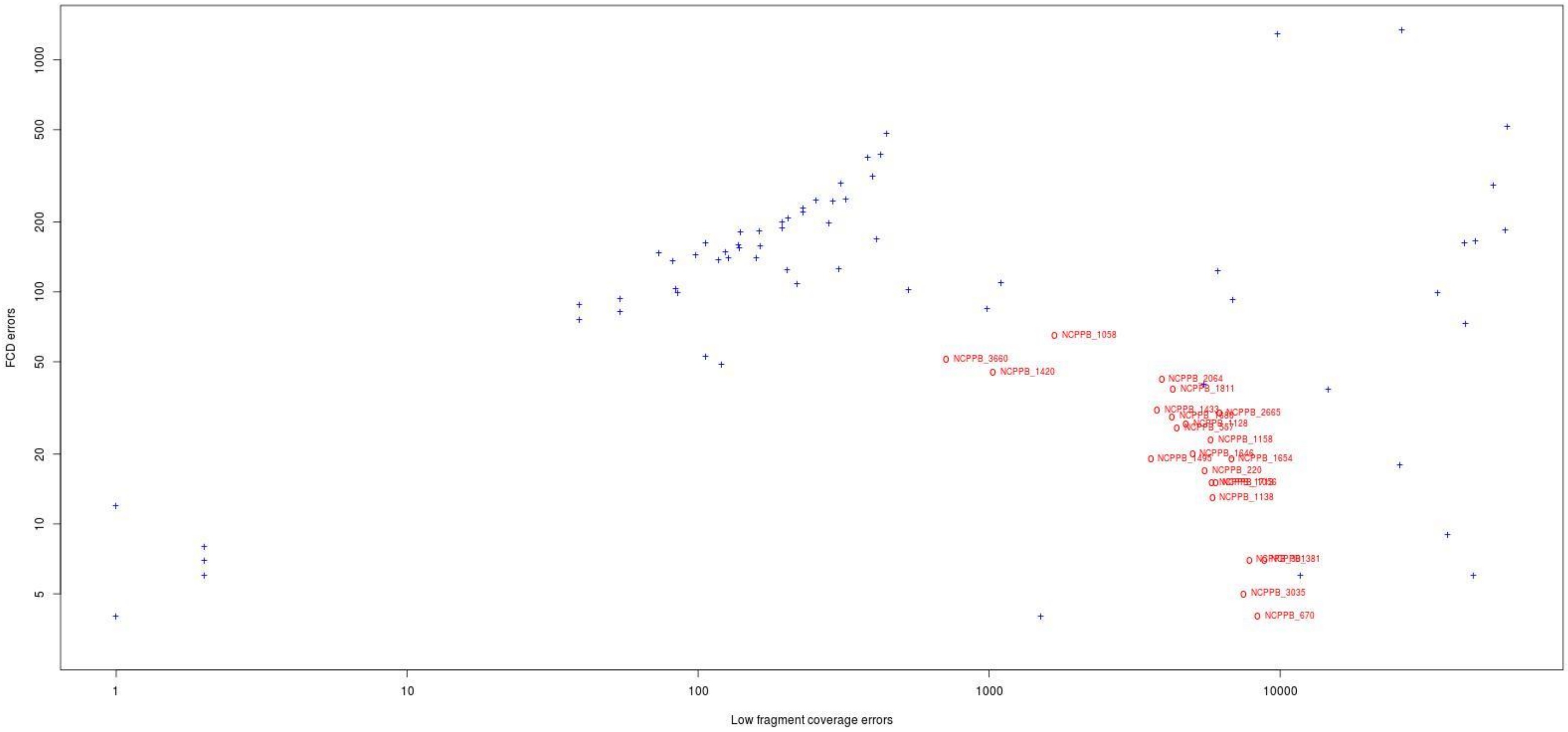
Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Supplementary Information

Figure S1 Comparison of assembly accuracy. To assess the accuracy of the newly assembled genome sequences, without a ground-truth reference genome sequence against which to compare, we used the REAPR tool (Hunt et al., 2013). REAPR detects two classes of potential errors in assemblies: Fragment coverage distribution (FCD) errors and low fragment coverage errors. We ran REAPR against each of the assemblies generated in the present study and also against each of the assemblies of *X. axonopodis* pv. *manihotis* described in a previously published study (Bart et al., 2012). A perfect error-free assembly would fall in the bottom left corner of the plot whereas a poor error-rich sequence would fall in the top right.

Red circles (o): Genome sequence assemblies from the present study.

Blue crosses (+): Genome assemblies from previously published study (Bart et al., 2012).



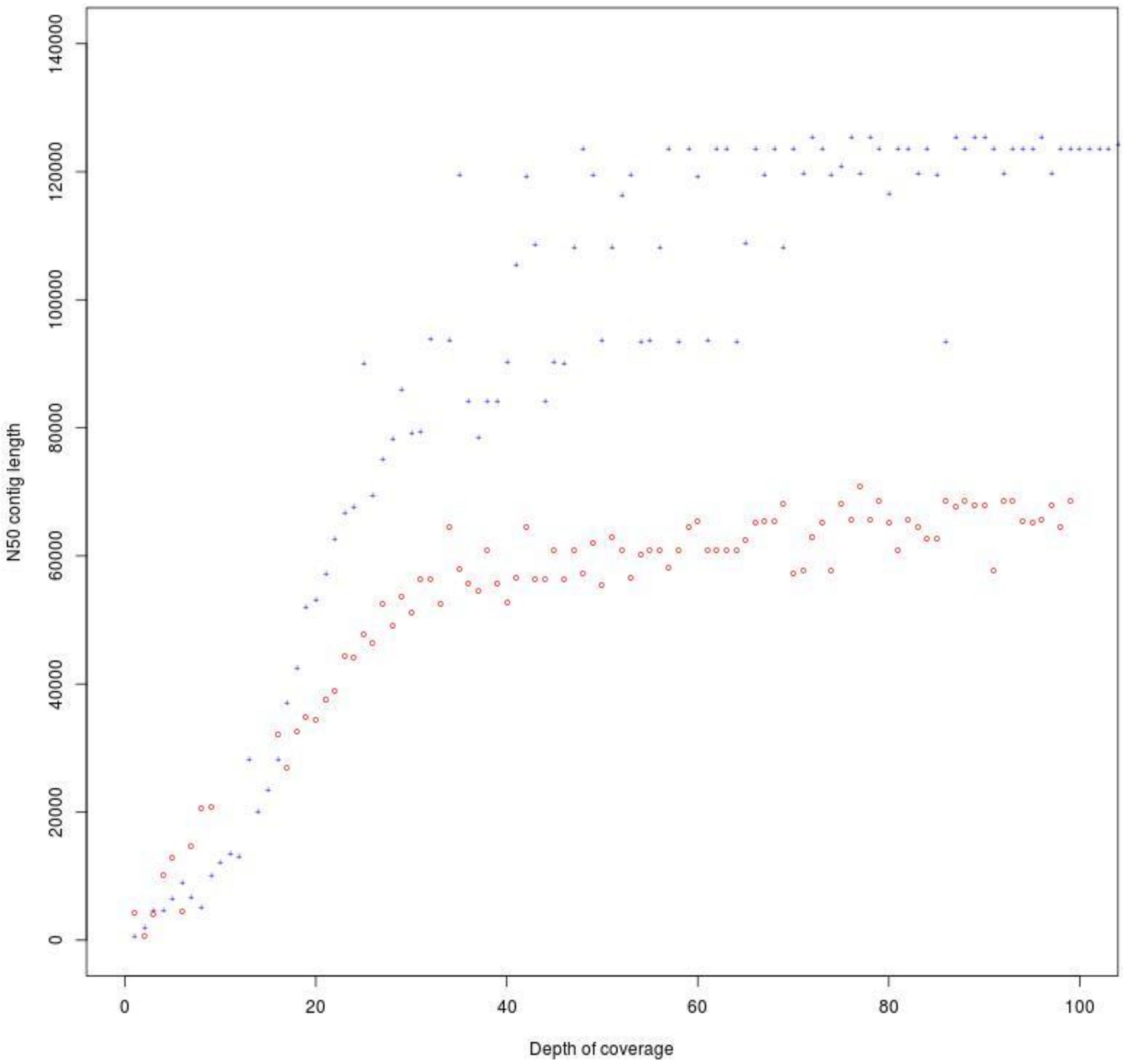


Figure S2 Effect of depth of coverage on the contiguity of genome assembly. To assess the effect of depth of read coverage on assembly contiguity, we subsampled reads from the datasets for strains NCPPB 2064 and NCPPB 1058 to give a range of depths from 1 to 100 x. Depth of coverage was calculated by dividing the total number of nucleotides in the input sequence reads by 4,700 (i.e. assuming that the genome size is 4.7 Mb). Each subsample of reads was assembled with SPAdes using the same protocol as for the assemblies described in the main text. The N_{50} contig length was calculated for each assembly using Quast (Gurevich et al., 2013).

Red circles (**o**): Genome sequence assemblies of *Xff* NCPPB 1058.

Blue crosses (**+**): Genome assemblies of *Xap* NCPPB 2064.

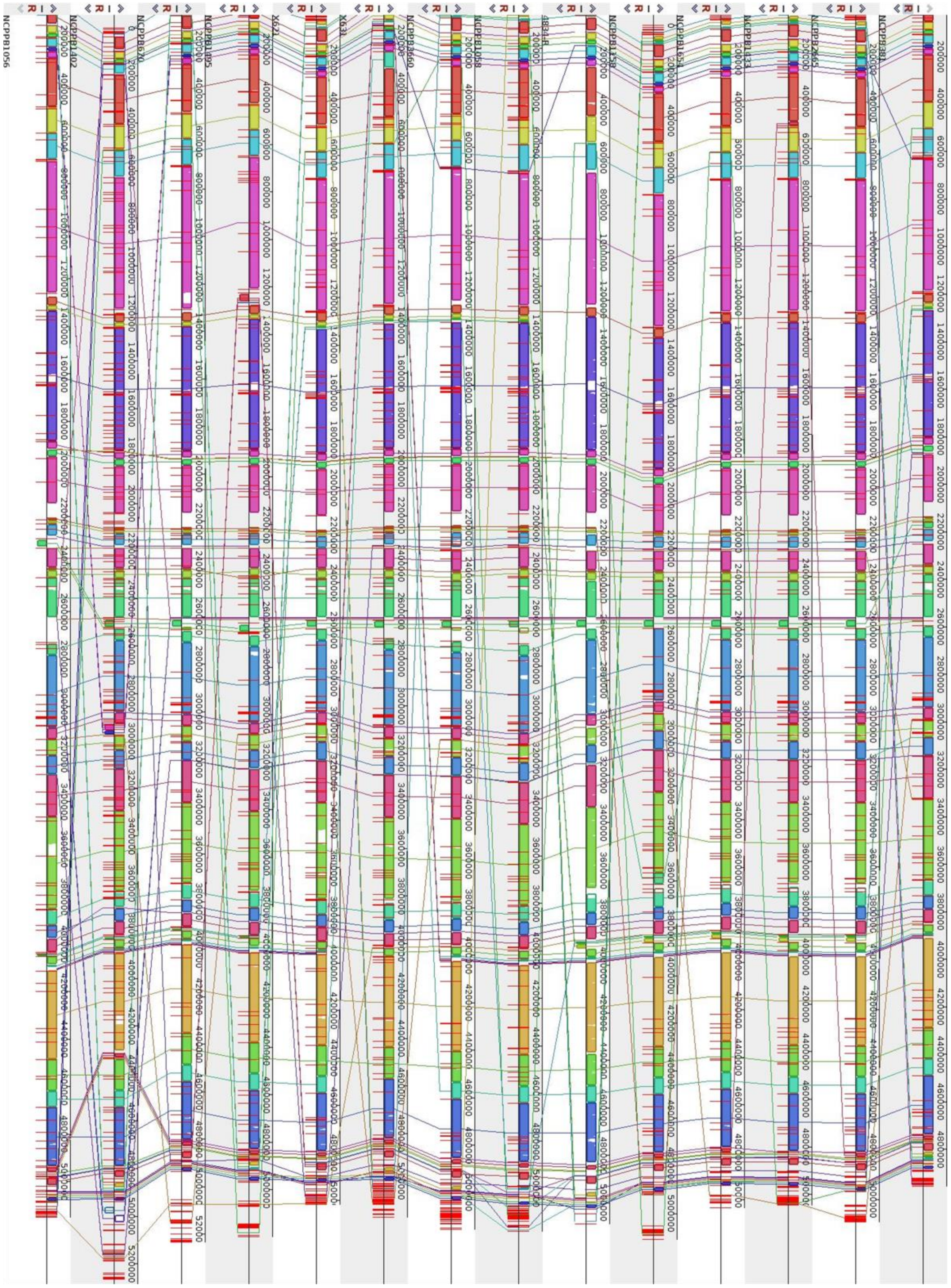


Figure S3 Whole-genome alignments of *X. fuscans* subsp. *fuscans*. The genome assemblies of each sequenced isolate were aligned, and using Mauve (Darling et al., 2004, 2010; Rissman et al., 2009).

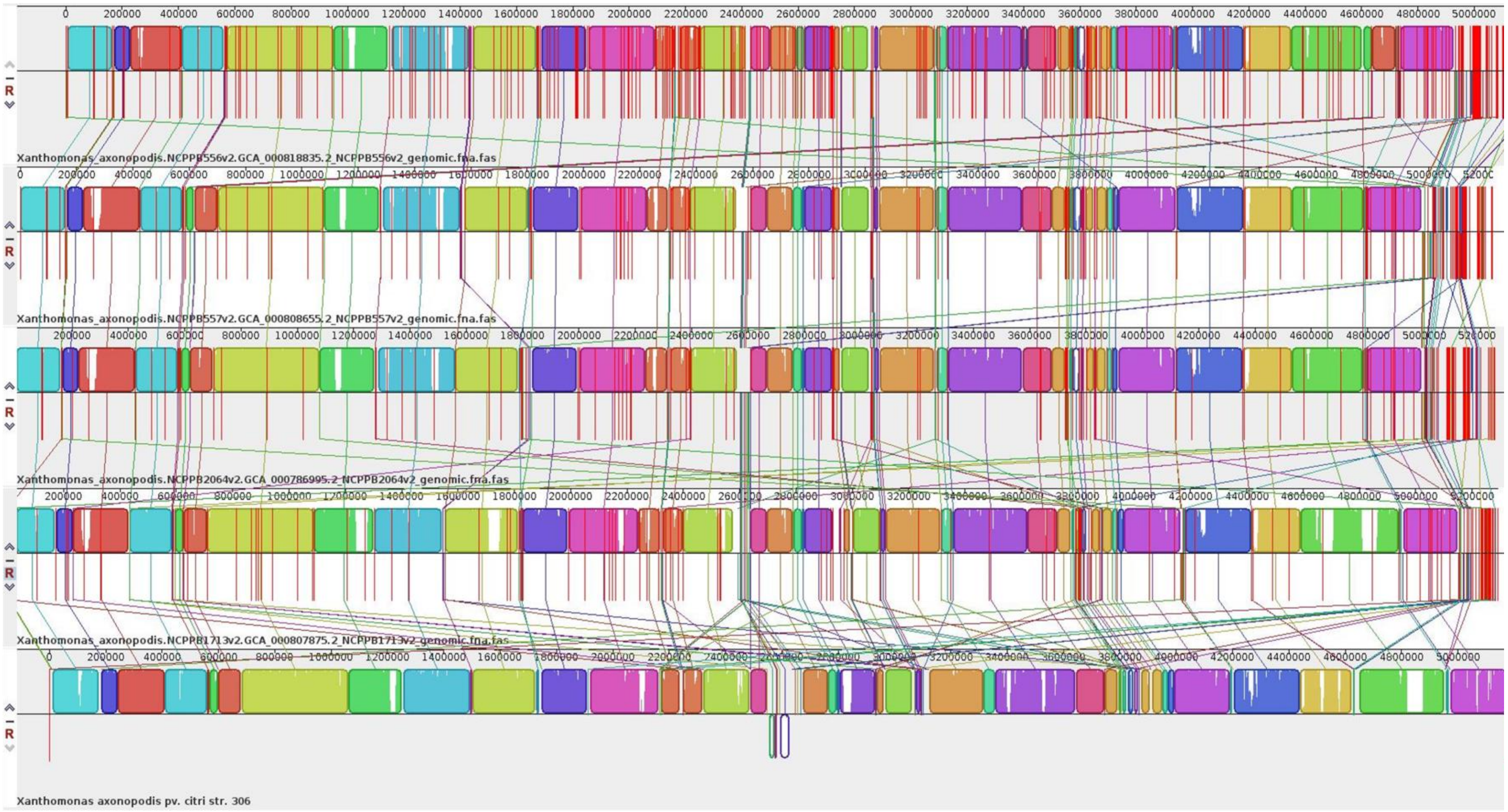


Figure S4 Whole-genome alignments of *X. phaseoli* pv. *phaseoli* lablab-associated isolates.

The genome assemblies of each sequenced isolate were aligned, and using Mauve (Darling et al., 2004, 2010; Rissman et al., 2009).

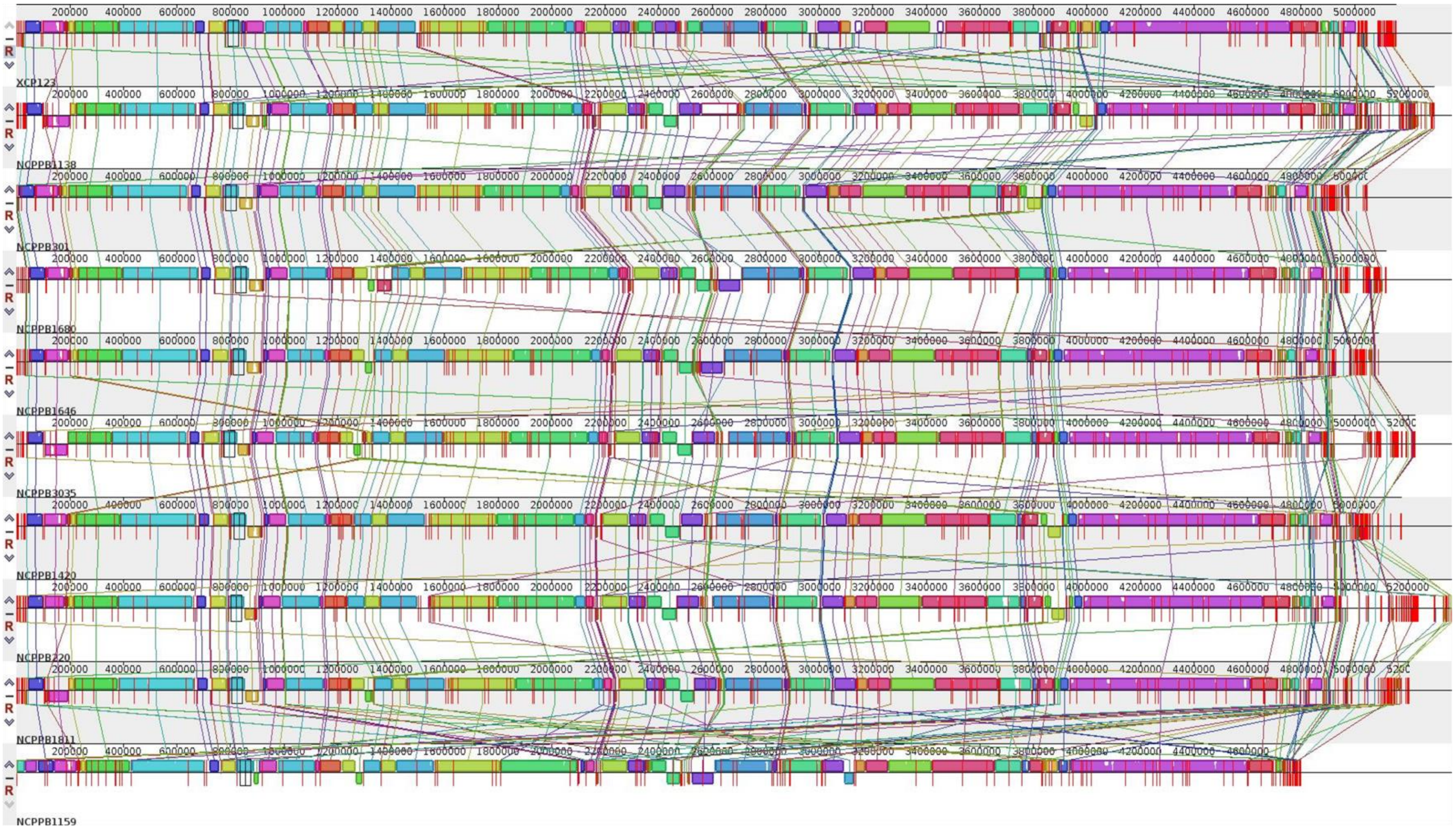


Figure S5 Whole-genome alignments of *X. phaseoli* pv. *phaseoli* genetic lineage 1 (GL1). The genome assemblies of each sequenced isolate were aligned, and using Mauve (Darling et al., 2004, 2010; Rissman et al., 2009). The genome assembly of *X. axonopodis* pv. *manihotis* NCPPB 1159 (Bart et al., 2012) is also included for comparison.

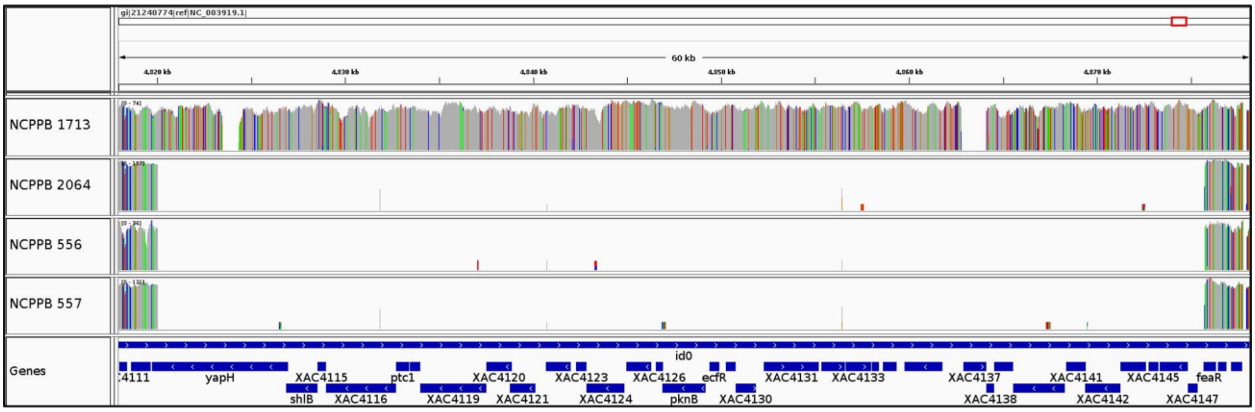


Figure S6 A 60-kbp genomic deletion found in Sudanese lablab-associated isolates NCPPB 2064, NCPPB 556 and NCPPB 557 but not in Zimbabwean isolate NCPPB 1713. The MiSeq sequence reads were aligned against the reference genome sequence of *X. axonopodis* pv. *citri* 306 (da Silva et al., 2002) using BWA-MEM (Li, 2013, 2014). The depth of coverage plots are visualised using IGV (Thorvaldsdóttir et al., 2013).

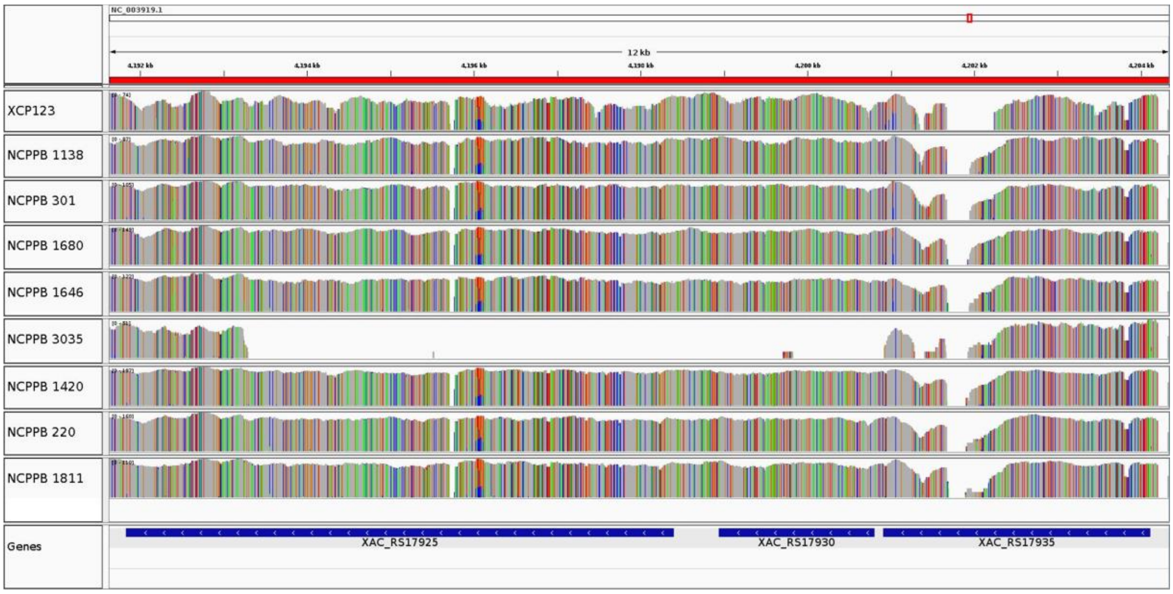


Figure S7 A 8-kbp genomic deletion found in *Xap* NCPPB 3035 but not in other sequenced *Xap* GL1 isolates. The MiSeq sequence reads were aligned against the reference genome sequence of *X. axonopodis* pv. *citri* 306 (da Silva et al., 2002) using BWA-MEM (Li, 2013, 2014). The depth of coverage plots are visualised using IGV (Thorvaldsdóttir et al., 2013).

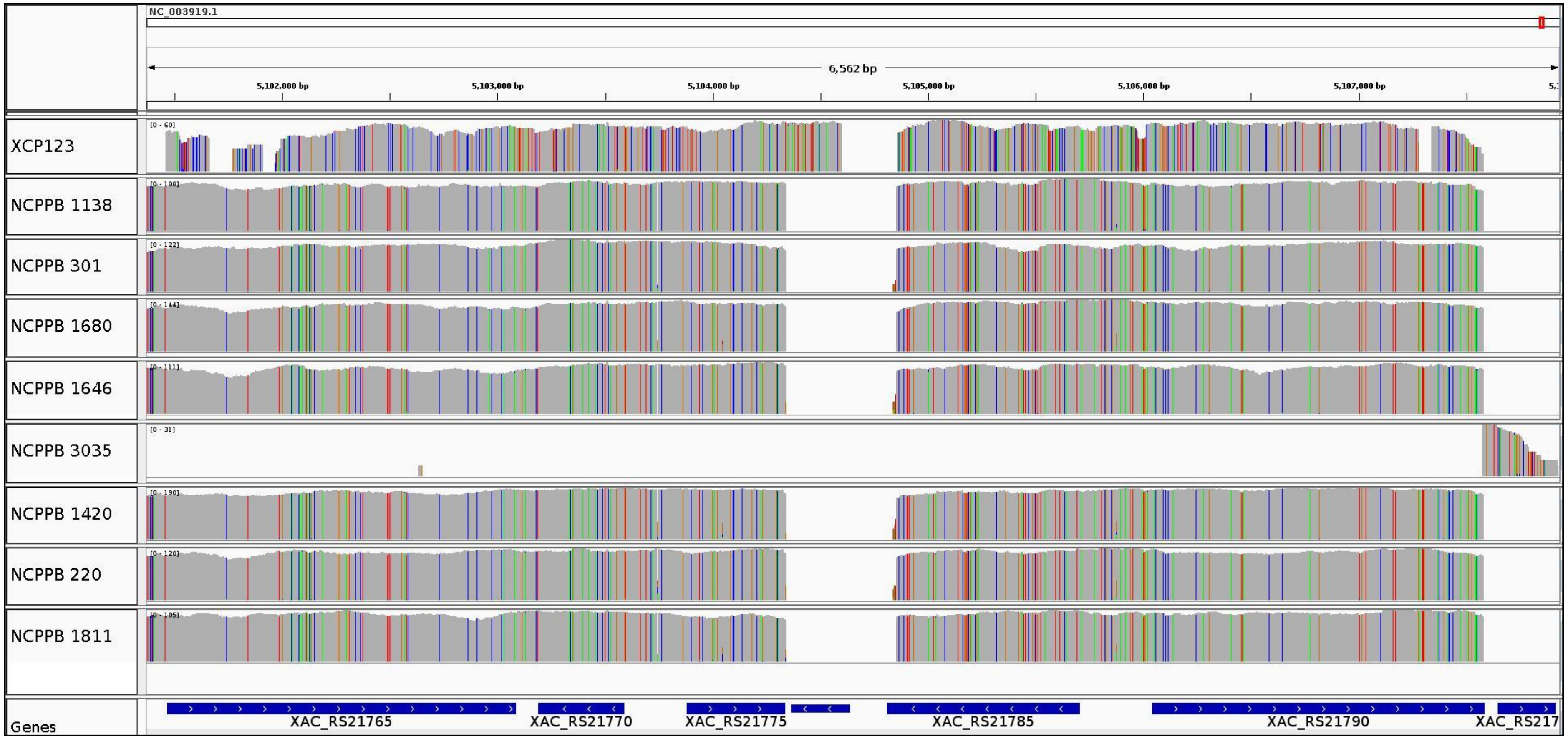


Figure S8 A 6-kbp genomic deletion found in *Xap* NCPPB 3035 but not in other sequenced Xap *GL1* isolates. The MiSeq sequence reads were aligned against the reference genome sequence of *X. axonopodis* pv. *citri* 306 (da Silva et al., 2002) using BWA-MEM (Li, 2013, 2014). The depth of coverage plots are visualised using IGV (Thorvaldsdóttir et al., 2013).

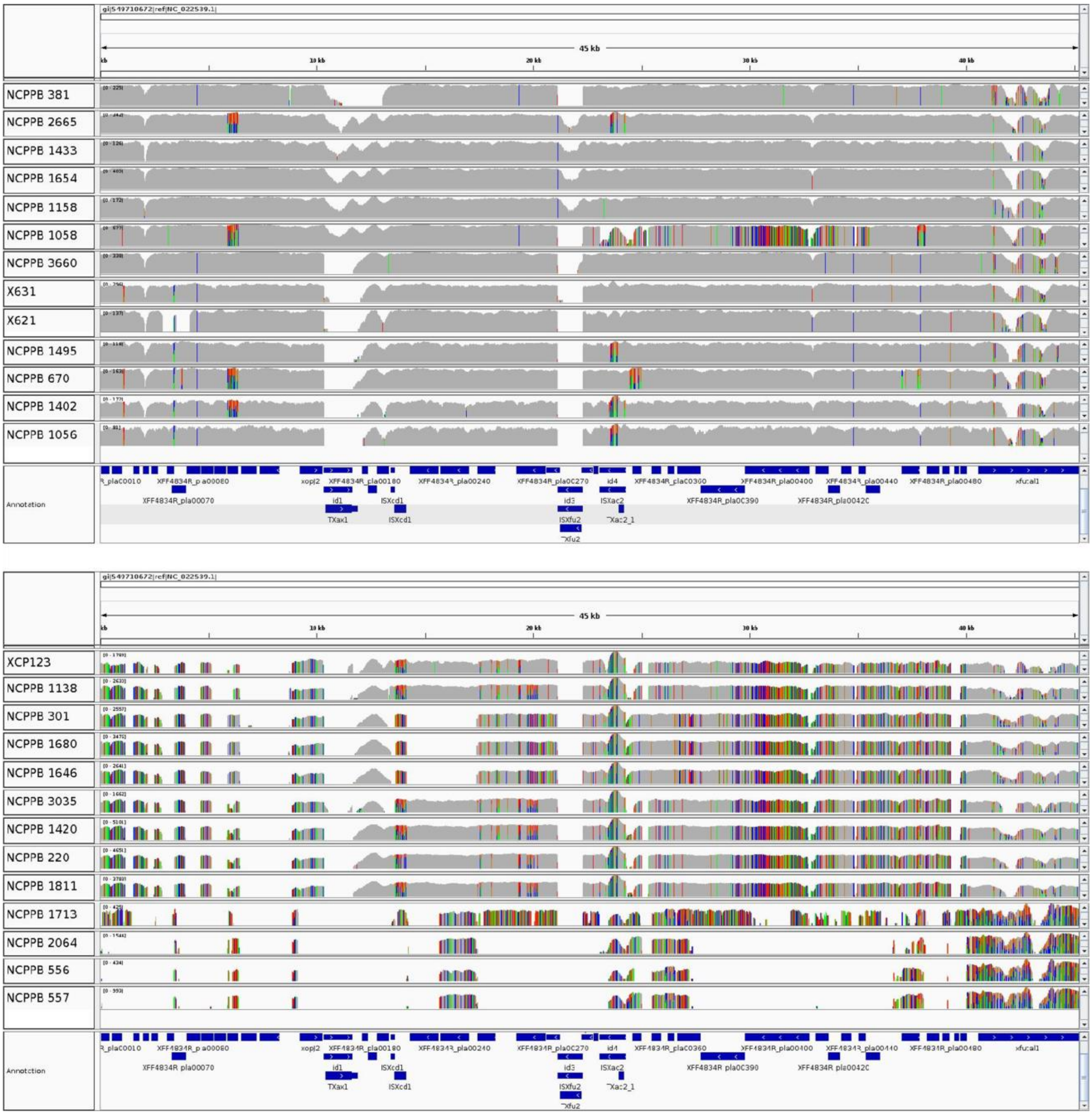


Figure S9 Conservation of *Xff* plasmid pla in the *Xff* and *Xap* isolates sequenced in the present study. The MiSeq sequence reads were aligned against the reference genome sequence of *Xff* 4834-R (Darrasse et al., 2013) using BWA-MEM (Li, 2013, 2014). The depth of coverage plots are visualised using IGV (Thorvaldsdóttir et al., 2013).

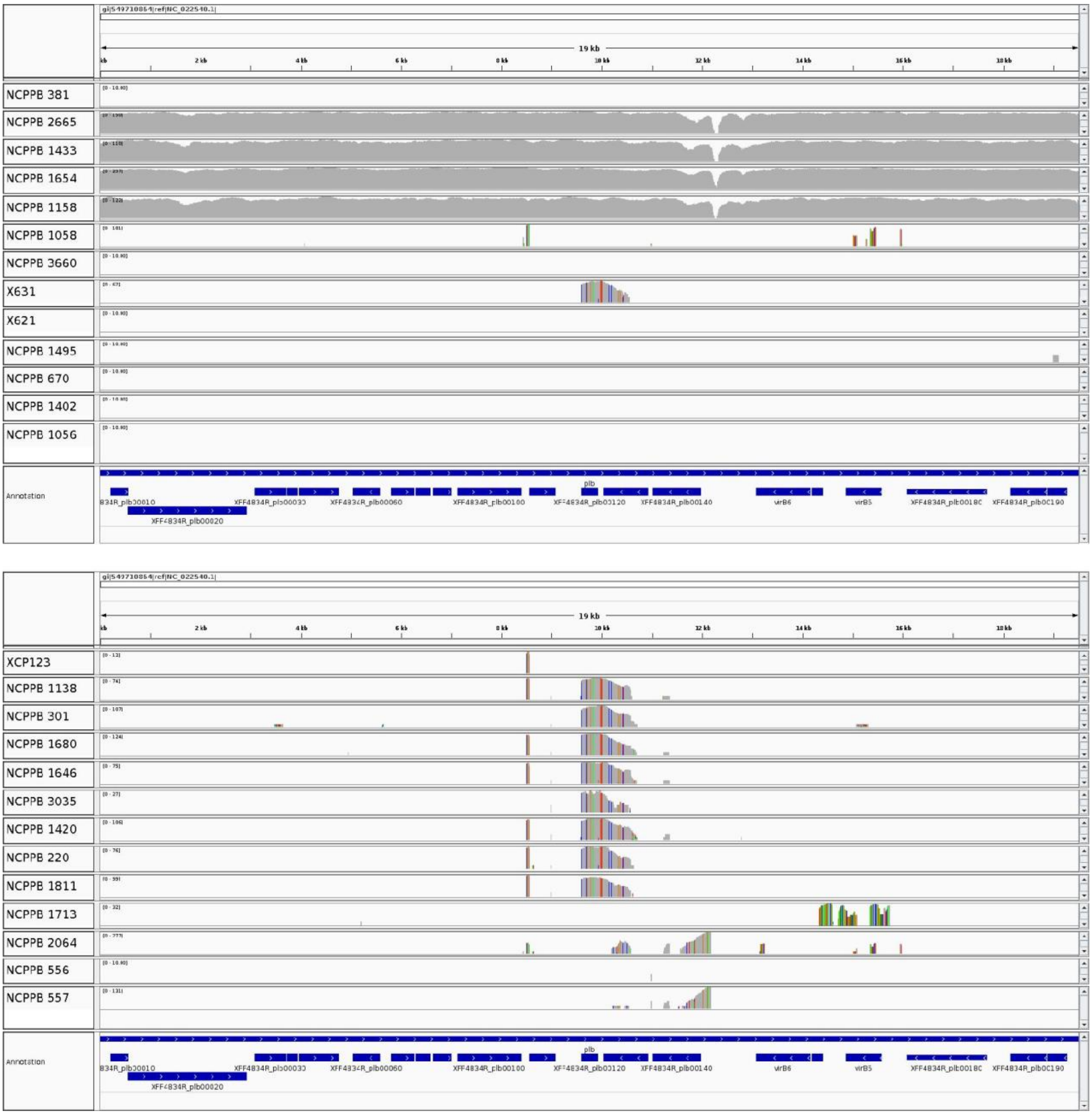


Figure S10 Conservation of *Xff* plasmid plb in the *Xff* and *Xap* isolates sequenced in the present study. The MiSeq sequence reads were aligned against the reference genome sequence of *Xff* 4834-R (Darrasse et al., 2013) using BWA-MEM (Li, 2013, 2014). The depth of coverage plots are visualised using IGV (Thorvaldsdóttir et al., 2013).

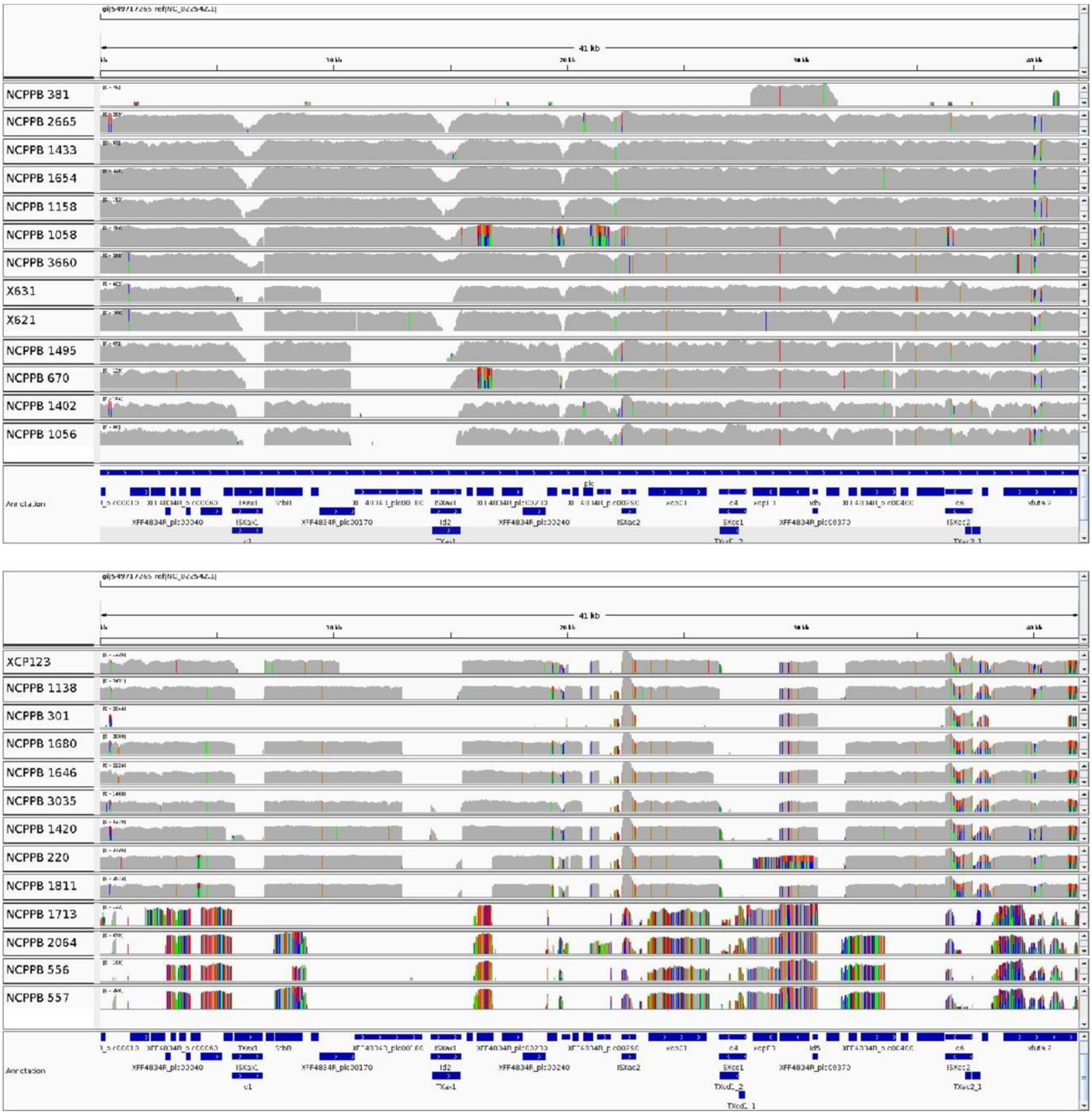


Figure S10 (cont) Conservation of *Xff* plasmid plc in the *Xff* and *Xap* isolates sequenced in the present study. The MiSeq sequence reads were aligned against the reference genome sequence of *Xff* 4834-R (Darrasse et al., 2013) using BWA-MEM (Li, 2013, 2014). The depth of coverage plots are visualised using IGV (Thorvaldsdóttir et al., 2013).

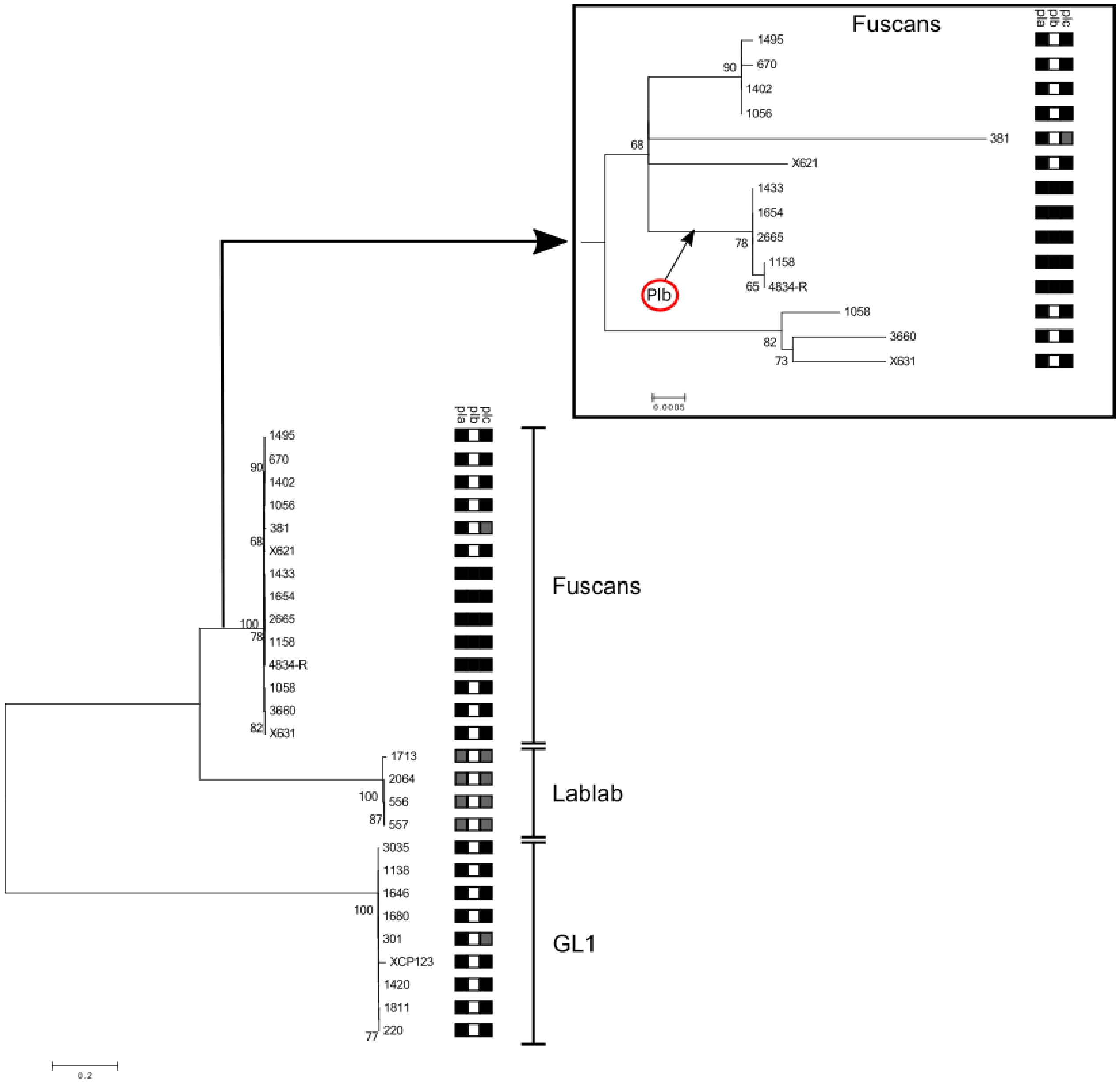


Figure S11. Distribution of plasmids found in *Xff* CFBP4884.
The distribution of each plasmid identified from *Xff* CFBP4884 across the three clades included in this study. The black boxes indicate presence of the plasmid, while white boxes indicate no evidence of the plasmid found in the assemblies. The grey boxes indicate partial sequence homology. The supplementary tree shows the topology of the fuscans clade in more detail, with the plasmid distribution mapped onto this for clarity.

Assembly (present study)	# contigs (>= 0 bp)	# contigs (>= 1000 bp)	Total length (>= 0 bp)	Total length (>= 1000 bp)	# contigs	Largest contig	total length	G+C (%)	N50	N75	L50	L75	# N's per 100 kbp
NCPPB2064 GCA_000786995.2	139	139	5,250,852	5,250,852	139	327,351	5,250,852	64.55	123,606	52,900	15	33	0
NCPPB557 GCA_000808655.2	143	143	5,226,308	5,226,308	143	326,992	5,226,308	64.55	116,632	47,769	16	33	0
NCPPB1128	71	71	4,950,425	4,950,425	71	475,671	4,950,425	68.89	107,247	74,331	14	27	0
NCPPB1713 GCA_000807875.2	144	144	5,291,882	5,291,882	144	356,409	5,291,882	64.61	96,285	52,695	19	38	0
NCPPB1159 GCA_000266285.1	128	128	4,797,980	4,797,980	128	257,906	4,797,980	65.18	96,063	46,574	17	36	0
NCPPB1654 GCA_000775185.2	134	134	5,032,953	5,032,953	134	367,357	5,032,953	64.76	81,530	40,461	16	38	0
NCPPB1680 GCA_000808695.2	158	158	5,117,812	5,117,812	158	224,929	5,117,812	64.9	77,618	47,328	24	45	0
NCPPB1420 GCA_000785935.2	165	165	5,174,926	5,174,926	165	224,974	5,174,926	64.86	76,066	42,067	23	46	0
NCPPB1056 GCA_000786945.2	134	134	5,150,070	5,150,070	134	233,487	5,150,070	64.77	75,549	44,652	20	42	0
NCPPB381 GCA_000788075.2	136	136	4,955,361	4,955,361	136	231,088	4,955,361	64.89	74,010	38,888	21	43	0
NCPPB1433 GCA_000775215.2	154	154	5,044,382	5,044,382	154	285,554	5,044,382	64.76	73,844	39,479	20	43	0
NCPPB1158 GCA_000775205.2	152	152	5,109,687	5,109,687	152	231,811	5,109,687	64.77	73,833	40,466	22	45	0
NCPPB3660 GCA_000786925.2	183	183	5,153,912	5,153,912	183	322,254	5,153,912	64.67	73,151	34,867	23	47	0
NCPPB670 GCA_000764875.2	152	152	5,233,004	5,233,004	152	201,426	5,233,004	64.68	71,533	38,197	22	48	0
NCPPB1058 GCA_000786935.2	168	168	5,180,142	5,180,142	168	324,288	5,180,142	64.7	71,334	35,921	25	49	0
NCPPB2665 GCA_000775195.2	154	154	5,092,836	5,092,836	154	319,856	5,092,836	64.72	71,115	34,305	19	45	0
NCPPB1495 GCA_000786915.2	158	158	5,197,608	5,197,608	158	194,645	5,197,608	64.64	70,760	36,664	24	50	0
X621 GCA_000817715.2	165	165	5,017,028	5,017,028	165	177,805	5,017,028	64.8	70,719	36,776	26	50	0
NCPPB1646 GCA_000785925.2	162	162	5,088,301	5,088,301	162	192,582	5,088,301	64.92	69,783	45,187	24	47	0
NCPPB1811 GCA_000808675.2	170	170	5,202,962	5,202,962	170	225,593	5,202,962	64.71	68,458	36,539	26	51	0
NCPPB301 GCA_000785945.2	153	153	5,047,533	5,047,533	153	192,590	5,047,533	64.95	67,789	45,183	25	48	0
XCP123 GCA_000827975.2	169	169	5,153,892	5,153,892	169	220,101	5,153,892	64.84	66,622	44,427	25	49	0
NCPPB1138 GCA_000808735.2	183	183	5,297,689	5,297,689	183	155,087	5,297,689	64.78	66,168	35,704	26	53	0
NCPPB220 GCA_000808715.2	191	191	5,364,294	5,364,294	191	151,739	5,364,294	64.61	61,941	40,326	30	56	0
X631 GCA_000827985.2	189	189	5,152,698	5,152,698	189	321,888	5,152,698	64.7	61,459	31,722	26	55	0
NCPPB3035 GCA_000774035.2	190	190	5,225,427	5,225,427	190	151,633	5,225,427	64.74	59,899	34,433	28	56	0
NCPPB556 GCA_000818835.2	288	288	5,191,650	5,191,650	288	166,874	5,191,650	64.58	41,494	18,726	38	88	0
NCPPB1402 GCA_000774025.2	240	240	5,342,561	5,342,561	240	113,545	5,342,561	64.68	39,404	20,951	43	89	0

Table S1 A. Information on assembly statistics for genomes used in this study.
Information on assembly statistics for genomes used in this study – statistics generated using QUAST.
All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Assembly (from Bart et. al, 2012)	# contigs (>= 0	bp) # contigs (>= 1000 bp)	Total length (>= 0 bp)	Total length (>= 1000 bp)	# contigs	Largest contig	Total length	GC (%)	N50	N75	L50	L75	# N's per 100 kbp
IBSBF1411	130	130	4,856,132	4,856,132	130	314,799	4,856,132	65.17	111,064	48,882	15	32	0
IBSBF2346	122	122	4,874,593	4,874,593	122	257,932	4,874,593	65.17	104,783	49,501	18	36	0
CFBP1851	123	123	4,798,436	4,798,436	123	281,156	4,798,436	65.18	96,104	47,325	17	34	0
NCPPB1159	128	128	4,797,980	4,797,980	128	257,906	4,797,980	65.18	96,063	46,574	17	36	0
IBSBF2672	131	130	4,856,510	4,855,603	131	257,571	4,856,510	65.16	95,444	45,918	18	37	0
IBSBF356	130	130	4,847,079	4,847,079	130	257,951	4,847,079	65.15	95,113	53,403	18	35	0
IBSBF2345	128	125	4,819,772	4,816,940	128	286,963	4,819,772	65.23	94,514	49,498	19	36	0
IBSBF320	121	121	4,803,726	4,803,726	121	257,673	4,803,726	65.19	94,274	49,412	17	36	0
UA226	129	129	4,951,862	4,951,862	129	290,136	4,951,862	65	88,566	49,656	18	38	0
IBSBF2820	150	150	4,930,429	4,930,429	150	257,651	4,930,429	65.11	88,415	42,506	19	40	0
IBSBF278	138	138	5,020,210	5,020,210	138	239,334	5,020,210	64.89	88,339	44,175	20	41	0
IBSBF436	149	149	4,903,327	4,903,327	149	257,580	4,903,327	65.15	83,955	44,597	19	39	0
IBSBF2670	139	139	4,902,902	4,902,902	139	280,128	4,902,902	65.13	83,729	44,599	17	36	0
UA324	146	146	4,916,013	4,916,013	146	161,357	4,916,013	65.09	77,947	42,353	22	44	0
IBSBF725	141	141	4,886,517	4,886,517	141	257,945	4,886,517	65.15	77,096	46,873	19	40	0
IBSBF2667	138	138	4,913,060	4,913,060	138	226,291	4,913,060	65.13	76,506	47,304	21	41	0
Xam669	137	137	4,820,297	4,820,297	137	201,435	4,820,297	65.21	75,427	42,095	21	43	0
UG27	131	131	4,903,342	4,903,342	131	268,832	4,903,342	65.06	74,810	44,155	19	39	0.02
UG24	145	144	4,909,909	4,908,917	145	233,940	4,909,909	65.04	74,286	43,883	20	41	0.02
IBSBF285	147	146	4,972,047	4,971,114	147	186,156	4,972,047	65.01	73,562	44,189	21	43	0.02
IBSBF2816	145	144	4,902,619	4,901,640	145	256,960	4,902,619	65.12	73,453	44,328	19	39	0
IBSBF726	140	137	4,832,905	4,829,984	140	257,968	4,832,905	65.18	72,346	42,156	21	42	0
UG21	148	148	5,010,468	5,010,468	148	237,252	5,010,468	65.06	71,375	39,255	23	47	0
ORSTX27	133	133	4,931,909	4,931,909	133	237,243	4,931,909	65.07	71,296	46,296	21	42	0
UA536	142	142	4,866,301	4,866,301	142	183,757	4,866,301	65.13	71,281	41,126	23	45	0
UA303	130	130	4,872,968	4,872,968	130	196,821	4,872,968	65.13	71,242	47,435	23	44	0
IBSBF2821	161	161	4,906,958	4,906,958	161	257,592	4,906,958	65.13	71,007	39,076	21	44	0
UG23	133	133	4,924,062	4,924,062	133	237,150	4,924,062	65.01	70,618	47,595	22	43	0
AT6B	163	163	4,946,093	4,946,093	163	237,170	4,946,093	65.01	69,689	34,822	22	47	0
UA306	144	144	5,065,231	5,065,231	144	235,368	5,065,231	65.01	69,670	43,868	23	45	0.02
ORST17	177	173	5,128,104	5,124,410	177	235,384	5,128,104	64.95	69,510	43,949	23	45	0
UA686	143	143	4,898,003	4,898,003	143	196,787	4,898,003	65.11	69,477	45,639	22	43	0
UG51	140	140	4,871,289	4,871,289	140	234,400	4,871,289	65.13	69,454	42,988	22	44	0
UA323	151	151	4,872,007	4,872,007	151	173,776	4,872,007	65.13	69,452	41,593	23	45	0
IBSBF289	155	155	4,847,529	4,847,529	155	226,288	4,847,529	65.14	69,253	35,540	23	46	0
CIO151	175	175	4,911,246	4,911,246	175	150,154	4,911,246	65.06	69,224	32,613	24	48	0
UG43	145	145	4,924,182	4,924,182	145	237,167	4,924,182	65.02	66,987	40,722	23	46	0
UG44	141	141	4,895,442	4,895,442	141	238,358	4,895,442	65.07	66,986	41,043	24	47	0

IBSBF2818	139	139	4,851,686	4,851,686	139	235,402	4,851,686	65.15	66,984	43,073	22	44	0
UA556	139	139	4,842,655	4,842,655	139	237,273	4,842,655	65.14	66,984	43,073	22	44	0
IBSBF2539	138	137	4,952,612	4,951,657	138	196,970	4,952,612	65.07	66,819	43,320	24	48	0
AFNC1360	149	149	4,921,075	4,921,075	149	235,381	4,921,075	65.07	65,052	42,988	25	48	0
Xam672	165	165	5,016,075	5,016,075	165	239,253	5,016,075	64.89	63,368	32,768	25	52	0
UA560	154	154	4,897,922	4,897,922	154	176,262	4,897,922	65.1	61,397	39,165	27	52	0
Xam1134	158	158	4,894,844	4,894,844	158	237,227	4,894,844	65.09	58,567	37,574	28	53	0
Xam668	195	195	4,954,860	4,954,860	195	276,859	4,954,860	64.92	46,240	27,846	29	63	0
Xam678	256	256	4,894,142	4,894,142	256	104,760	4,894,142	65.1	38,200	17,862	47	93	0
IBSBF2666	311	311	4,842,277	4,842,277	311	123,056	4,842,277	65.1	28,335	15,839	49	105	0
IBSBF2673	305	305	4,830,620	4,830,620	305	120,981	4,830,620	65.12	27,765	16,026	53	109	0
UG45	323	323	4,986,397	4,986,397	323	83,811	4,986,397	64.89	25,647	14,923	59	123	0
Thaiassembly.fna broken	324	322	4,874,623	4,872,860	324	84,544	4,874,623	65.16	25,423	14,825	62	124	0
IBSBF2665	434	434	4,836,797	4,836,797	434	111,173	4,836,797	65.14	17,985	10,227	83	172	0
IBSBF2822	464	463	4,831,349	4,830,360	464	57,666	4,831,349	65.13	17,213	9,515	93	188	0
IBSBF2819	557	557	4,852,810	4,852,810	557	59,730	4,852,810	65.12	13,719	7,675	104	224	0
IBSBF2538	573	573	4,833,429	4,833,429	573	50,456	4,833,429	65.1	12,527	7,291	122	249	0
IBSBF1182	648	646	4,773,692	4,771,727	648	46,493	4,773,692	65.04	11,601	6,560	128	261	0
UG28	676	672	4,722,734	4,719,002	676	54,751	4,722,734	65.14	10,796	6,238	131	276	0
IBSBF614	706	706	4,730,788	4,730,788	706	51,491	4,730,788	65.03	10,598	5,680	136	291	0
IBSBF1994	677	674	4,774,518	4,771,587	677	45,564	4,774,518	65.08	10,370	6,007	137	287	0
NG1	785	784	4,737,721	4,736,722	785	38,252	4,737,721	64.9	9,229	5,016	160	335	0
ORST4	767	766	4,650,210	4,649,231	767	36,169	4,650,210	65.03	9,012	4,950	161	336	0
IBSBF280	850	848	4,842,752	4,840,771	850	41,005	4,842,752	64.81	8,666	4,696	171	358	0
IBSBF321	888	883	4,593,056	4,588,103	888	35,503	4,593,056	64.9	7,483	4,318	183	389	0

Table S1 B. information on assembly statistics for the genomes published by hajri et al 2012

Comparison of assembly statistics for the genomes published by BART et al 2012 - statistics generated using QUAST.

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Table S2. Genes found in *Xff* but absent from *Xap* GL1 and lablab-associated Xap. (following page) Each of these genes was covered over at least 95% of its length by sequence reads from each of the sequenced Xff genomes but covered by no more than 30% of its length by aligned reads from any of the Xap genomes.

Breadth of coverage was based on alignments of the genomic sequences reads against the reference pan-genome using BWA-MEM.

Accession	Description
KGK64681	fimbrial protein
KGT57443	hypothetical protein
KGU47595.1	hypothetical protein
KGU48127.1	hypothetical protein
KGU49471.1	hypothetical protein
KGU50252.1	fimbrial protein
KGU50381.1	NADPH:quinone oxidoreductase
KGU50382.1	AraC family transcriptional regulator
KGU50383.1	N-ethylmaleimide reductase
KGU50384.1	TetR family transcriptional regulator
KGU50385.1	short-chain dehydrogenase
KGU50807.1	glyoxalase
KGU50814.1	hypothetical protein
KGU50868.1	hypothetical protein
KGU50875.1	methyltransferase
KGU50876.2	hypothetical protein
KGU51538.1	membrane protein
KGU51543.1	hypothetical protein
KGU51683.1	modification methylase
KGU51684.1	hypothetical protein
KGU51685.1	hypothetical protein
KGU51686.1	hypothetical protein
KGU51744.1	hypothetical protein
KGU52010.1	alpha/beta hydrolase
KGU52011.1	LysR family transcriptional regulator
KGU52012.1	FMN-dependent NADH-azoreductase
KGU52013.1	hypothetical protein
KGU52019.1	hypothetical protein
KGU52023.1	hypothetical protein
KGU52101.1	hypothetical protein
KGU52102.1	hypothetical protein
KGU52107.1	hypothetical protein
KGU52614.1	hypothetical protein
KGU52669.1	hypothetical protein
KGU52846.1	beta-lactamase
KGU52847.1	transcriptional regulator
KGU53094.1	hypothetical protein
KGU53261.1	methyltransferase
KGU53514.1	hypothetical protein
KGU53764.1	pilus assembly protein PilV
KGU53765.1	pilus assembly protein
KGU53766.1	pilus assembly protein PilE
KGU53769.1	pre-pilin like leader sequence
KGU53770.1	pilus assembly protein
KGU54237.1	integrase
KGU55371.1	galactarate dehydrogenase
KGU55372.2	glucarate dehydratase
KGU55373.1	glucarate transporter
KGU55374.1	2 5-dioxovalerate dehydrogenase
KGU55375.1	5-dehydro-4-deoxyglucarate dehydratase
KGU55376.1	LysR family transcriptional regulator
KGU55422.1	porin
KGU55563.1	hypothetical protein
KGU56157.1	TonB-dependent receptor
KGU56374.1	membrane protein
KGU56715.1	hypothetical protein
KGU56733.1	hypothetical protein
KGU56734.1	amine oxidase
KGU56735.1	polysaccharide biosynthesis protein
KGU56736.1	ribonuclease III
KGU56737.1	hypothetical protein
KGU56829.1	hypothetical protein
KGU56830.1	hypothetical protein
KKW48692.1	hypothetical protein
KKW48708.1	hypothetical protein

KKW48794.1	hypothetical protein
KKW48949.1	hypothetical protein
PRJNA263153:NB99 22880	hypothetical protein

Table S3. Genes found in *Xap* GL1 but absent from *Xff* and lablab-associated *Xap*.
Each of these genes was covered over at least 95% of its length by sequence reads from each of the sequenced *Xap* GL1 genomes but covered by no more than 30% of its length by aligned reads from any of the *Xff* or lablab-associated *Xap* genomes.
Breadth of coverage was based on alignments of the genomic sequences reads against the reference pan-genome using BWA-MEM.

Accession	Description
KHS31773	hypothetical protein
KHS32523	hypothetical protein
KHS32525	hypothetical protein
KHS32582	baseplate assembly protein
KHS32717	transposase
KHS32718	hypothetical protein
KHS32719	hypothetical protein
KHS32720	hypothetical protein
KHS32721	hypothetical protein
KHS32722	hypothetical protein
KHS32723	hypothetical protein
KHS33554	hypothetical protein
KHS33556	integrase
KHS33557	hypothetical protein
KHS33797	hypothetical protein
KHS33831	tail fiber assembly protein
KHS33832	hypothetical protein
KHS33862	hypothetical protein
KHS33863	hypothetical protein
KHS33864	hypothetical protein
KHS33930	hypothetical protein
KHS34155	hypothetical protein
KHS34162	hypothetical protein
KHS34252	hypothetical protein
KHS34301	hypothetical protein
KHS34370	hypothetical protein
KHS34371	integrase
KHS34433	hypothetical protein
KHS34435	BadM/Rrf2 family transcriptional regulator
KHS34436	thioredoxin reductase
KHS34456	ATPase
KHS34457	chemotaxis protein CheY
KHS34561	hypothetical protein
KHS34585	type III secretion system protein InvA
KHS34586	hypothetical protein
KHS34587	hypothetical protein
KHS34588	hypothetical protein
KHS34589	hypothetical protein
KHS34590	hypothetical protein
KHS34591	type III secretion system protein
KHS34592	hypothetical protein
KHS34593	hypothetical protein
KHS34594	hypothetical protein
KHS34595	hypothetical protein

KHS34596	hypothetical protein
KHS34597	hypothetical protein
KHS34598	hypothetical protein
KHS34599	hypothetical protein
KHS34600	hypothetical protein
KHS34601	hypothetical protein
KHS34602	hypothetical protein
KHS34627	protein-S-isoprenylcysteine methyltransferase
KHS34635	LuxR family transcriptional regulator
KHS34637	hypothetical protein
KHS34671	hypothetical protein
KHS34690	nitrite reductase
KHS34691	MFS transporter
KHS34692	nitrate ABC transporter substrate-binding
KHS34836	hypothetical protein
KHS34920	oxidoreductase
KHS34921	LysR family transcriptional regulator
KHS34930	glutamyl-tRNA(Gln) amidotransferase subunit A
KHS34931	membrane protein
KHS34937	transcriptional regulator
KHS34966	hypothetical protein
KHS35235	hypothetical protein
KHS35333	hypothetical protein
KHS35341	hypothetical protein
KHS35507	hypothetical protein
KHS35625	AraC family transcriptional regulator
KHS35626	peptidase S41
KHS35637	TonB-dependent receptor
KHS35638	hypothetical protein
KHS35639	hypothetical protein
KHS35733	hypothetical protein
KHS35747	hypothetical protein
KHS35763	hypothetical protein
KHS35765	hypothetical protein
KHS35777	hypothetical protein
KHS35962	hypothetical protein
KHS36059	membrane protein
KHS36099	hypothetical protein
KHS36103	hypothetical protein
KHS36305	peptide-binding protein
KHS36321	hypothetical protein
KHS36395	hypothetical protein
KHS36488	hypothetical protein
KHS36558	hypothetical protein
KHS36587	transcriptional regulator
KHS36599	hypothetical protein
KHS36611	hypothetical protein
KHS36756	hypothetical protein
KHS36820	hypothetical protein
KHS36905	hypothetical protein
KHS36918	hypothetical protein
KHS36926	cation transporter
KHS36933	transporter
KHS36941	RNA methyltransferase
KHS36962	coproporphyrinogen III oxidase
KHS36963	metalloenzyme domain-containing protein
KHS36964	hypothetical protein
KHS36965	ATPase AAA
KHS36967	hypothetical protein
KHS36974	hypothetical protein

KHS37170	hypothetical protein
KHS37202	RNA polymerase sigma70
KHS37316	glucan biosynthesis protein
KHS37329	membrane protein
KHS37355	hypothetical protein
KHS37356	sulfotransferase
KHS37357	ABC transporter ATP-binding protein
KHS37558	hypothetical protein
KHS37688	hypothetical protein
KHS37689	hypothetical protein
KHS37690	hypothetical protein
KHS37779	hypothetical protein
KHS37780	hypothetical protein
KHS37803	ArsR family transcriptional regulator
KHS37933	hypothetical protein
KHS37990	hypothetical protein
KHS38607	hypothetical protein
KHS38608	hypothetical protein
KHS38895	hypothetical protein
KHS38947	fimbrial protein
KHS38949	pilus assembly protein
KHS39041	3-deoxy-manno-octulosonate cytidylyltransferase
KHS39043	methyltransferase
KHS39101	hypothetical protein
KHS39108	addiction module protein
KHS39222	ketosteroid isomerase
KHS39223	LysR family transcriptional regulator
KHS39227	chloride channel protein
KHS39521	hypothetical protein
KHS39674	hypothetical protein
KHS39807	hypothetical protein
KHS39921	hypothetical protein
KHS40058	hypothetical protein
KHS40082	histidine kinase
KHS40083	hypothetical protein
KHS40090	chitinase
KHS40367	histidine kinase
KHS40369	Tfp pilus assembly protein PilW
KHS40886	hypothetical protein
KHS40930	hypothetical protein
KHS40957	hypothetical protein
KHS41130	hypothetical protein
KHS41166	hypothetical protein
KHS41214	hypothetical protein
KHS41218	transcriptional regulator
KHS41299	hypothetical protein
KHS41374	long-chain acyl-CoA synthetase
KHS41375	long-chain fatty acid--CoA ligase
KHS41376	short-chain dehydrogenase
KHS41377	tetratricopeptide repeat protein
KHS41378	transcriptional regulator
KHS41390	histidine kinase
KHS41417	aspartyl beta-hydroxylase
KHS41418	cytochrome c biogenesis factor
KHS41419	hypothetical protein
KHS41420	sulfotransferase
KHS41421	TonB-dependent receptor
PRJNA270010:RN19 23245	chemotaxis protein
PRJNA270010:RN19 23465	hypothetical protein
PRJNA270010:RN19 23605	hypothetical protein

PRJNA270010:RN19 23805	hypothetical protein
PRJNA270010:RN19 23905	pilus assembly protein
PRJNA270010:RN19 24020	hypothetical protein
PRJNA270010:RN19 24165	hypothetical protein
PRJNA270010:RN19 24235	hypothetical protein
PRJNA270010:RN19 24245	hypothetical protein
PRJNA270010:RN19 25155	hypothetical protein

Table S4. Genes found in lablab-associated *Xap* but absent from *Xff* and *Xap* GL1.

Each of these genes was covered over at least 95% of its length by sequence reads from each of the sequenced lablab-associated *Xap* genomes but covered by no more than 30% of its length by aligned reads from any of the *Xff* or *Xap* GL1 genomes.

Breadth of coverage was based on alignments of the genomic sequences reads against the reference pan-genome using BWA-MEM.

Accession	Description
KHF46216	hypothetical protein
KHF46253	hypothetical protein
KHF46840	hypothetical protein
KHF46860	histidine kinase
KHF49249	hypothetical protein
KHS05251	hypothetical protein
KHS05293	hypothetical protein
KHS05314	hypothetical protein
KHS05315	Type III restriction enzyme res subunit
KHS05316	DNA methylase
KHS05318	hypothetical protein
KHS05350	hypothetical protein
KHS05374	hypothetical protein
KHS05398	hypothetical protein
KHS05399	hypothetical protein
KHS05428	hypothetical protein
KHS05432	oxidoreductase
KHS05433	hypothetical protein
KHS05434	hypothetical protein
KHS05484	oxidoreductase
KHS05485	epimerase
KHS05489	pilus assembly protein PilW
KHS05559	peptidase M61
KHS05757	XRE family transcriptional regulator
KHS05871	GCN5 family acetyltransferase
KHS05996	hypothetical protein
KHS05997	radical SAM protein
KHS05998	hypothetical protein
KHS06001	integrase
KHS06149	hypothetical protein
KHS06445	phosphopantetheinyl transferase
KHS06764	enterocin
KHS06882	aldehyde oxidase
KHS06992	hypothetical protein
KHS07180	hypothetical protein
KHS07181	hypothetical protein
KHS07201	phage tail protein
KHS07203	arylsulfotransferase
KHS07204	hypothetical protein

KHS07206	SAM-dependent methlyltransferase
KHS07207	asparagine synthase
KHS07208	hypothetical protein
KHS07209	hypothetical protein
KHS07210	transposase
KHS07211	methyltransferase
KHS07212	membrane protein
KHS07213	asparagine synthase
KHS07214	hypothetical protein
KHS07215	membrane protein
KHS07216	glycosyl transferase family 1
KHS07217	UDP-N-acetyl-D-mannosamine transferase
KHS07218	ligase
KHS07219	glycosyl transferase family 1
KHS07220	UDP-phosphate glucose phosphotransferase
KHS07221	glycosyl transferase family 1
KHS07222	mannosyltransferase
KHS07252	hypothetical protein
KHS07253	hypothetical protein
KHS07254	transporter
KHS07255	sugar transporter
KHS07256	methyltransferase type 12
KHS07296	hypothetical protein
KHS07297	hypothetical protein
KHS07330	AraC family transcriptional regulator
KHS07331	hypothetical protein
KHS07332	START domain protein
KHS07333	hypothetical protein
KHS07334	short-chain dehydrogenase
KHS07338	type III secretion system effector protein
KHS07498	hypothetical protein
KHS07499	hypothetical protein
KHS07584	calcium-binding protein
KHS07783	membrane protein
KHS07836	hypothetical protein
KHS07979	hypothetical protein
KHS08090	hypothetical protein
KHS08206	CopG family transcriptional regulator
KHS08252	transcriptional regulator
KHS08317	hypothetical protein
KHS08379	hypothetical protein
KHS08423	hypothetical protein
KHS08765	hypothetical protein
KHS08766	hypothetical protein
KHS08767	glycosyl transferase family 9
KHS08773	hypothetical protein
KHS08981	ligand-gated channel
KHS09024	UDP kinase
KHS09025	flagellar biosynthesis protein FlhB
KHS09026	flagellar biosynthesis protein FlhA
KHS09027	flagellar hook-basal body protein
KHS09028	flagellar hook-basal body protein
KHS09029	flagellar basal body P-ring biosynthesis protein
KHS09031	flagellar P-ring protein FlgI
KHS09032	hypothetical protein
KHS09033	hypothetical protein
KHS09034	hypothetical protein
KHS09035	ATP synthase
KHS09036	hypothetical protein
KHS09037	flagellar hook capping protein

KHS09038	flagellar hook-basal body protein
KHS09039	hypothetical protein
KHS09040	hypothetical protein
KHS09041	flagellar biosynthesis protein flip
KHS09042	RNA polymerase sigma-70 factor
KHS09072	hypothetical protein
KHS09073	hypothetical protein
KHS09162	hypothetical protein
KHS09164	hypothetical protein
KHS09167	hypothetical protein
KHS09270	hypothetical protein
KHS31033	hypothetical protein
PRJNA268142:QQ30 23585	transposase
PRJNA268142:QQ30 23775	hypothetical protein
PRJNA268142:QQ30 25245	hypothetical protein
PRJNA268142:QQ30 26020	flagellar M-ring protein FliF
PRJNA268142:QQ30 26715	hypothetical protein
PRJNA269802:RM61 23105	hypothetical protein
PRJNA269802:RM61 23120	glucose sorbosone dehydrogenase
PRJNA269802:RM61 23125	peptidoglycan-binding protein
PRJNA269802:RM61 23135	hypothetical protein
PRJNA269802:RM61 23315	hypothetical protein
PRJNA269802:RM61 23485	hypothetical protein
PRJNA269802:RM61 23580	hypothetical protein
PRJNA269802:RM61 24490	hypothetical protein
PRJNA269802:RM61 24950	transposase
PRJNA269802:RM61 25105	hypothetical protein
PRJNA269802:RM61 25435	type III secretion system effector protein

Accession	Description
KGU48816.1	hypothetical protein
KGU50255.1	hypothetical protein
KGU50386.1	hypothetical protein
KGU50427.1	hypothetical protein
KGU50466.1	LysR family transcriptional regulator
KGU50467.1	short-chain dehydrogenase
KGU50793.1	transducer protein car
KGU50994.1	oxidoreductase
KGU50995.1	AraC family transcriptional regulator
KGU51344.1	integrase
KGU51377.2	ADP-ribosylation/crystallin J1
KGU51433.1	hypothetical protein
KGU51457.1	hypothetical protein
KGU51533.1	transposase
KGU51650.1	hypothetical protein
KGU51734.1	LysR family transcriptional regulator
KGU51748.1	alpha/beta hydrolase
KGU52009.1	endonuclease
KGU52015.1	hypothetical protein
KGU52017.1	hypothetical protein
KGU52018.1	sulfate transporter
KGU52020.1	methionine aminopeptidase
KGU52021.1	hypothetical protein
KGU52624.1	tail collar protein
KGU52632.2	hemagglutinin
KGU52827.1	adenylate cyclase
KGU53452.1	hypothetical protein
KGU53454.1	hypothetical protein
KGU53466.1	dihydroorotate dehydrogenase
KGU54122.1	membrane protein
KGU54128.1	methyltransferase
KGU54129.1	cytochrome P450
KGU54131.1	hypothetical protein
KGU54490.1	hypothetical protein
KGU55549.1	hypothetical protein
KGU55562.1	hypothetical protein
KGU56526.1	hypothetical protein
KGU56527.1	transcriptional regulator
KGU56539.1	biopolymer transporter ExbD
KGU56660.1	dipeptidyl-peptidase 7
KHS34695	F420H2:quinone oxidoreductase
KHS39585	plasmid stabilization protein
KKW48473.1	hypothetical protein
KKW48540.1	hypothetical protein
KKW48549.1	hypothetical protein
KKW48757.1	HpaF protein
KKW48760.1	hypothetical protein
KKW48810.1	resolvase
PRJNA266873:PK63 22130	transposase

Table S5. Genes found in *Xap* GL1 and *Xff* but absent from lablab-associated *Xap*.
Each of these genes was covered over at least 95% of its length by sequence reads from each of the sequenced *Xap* GL1 and *Xff* genomes but covered by no more than 30% of its length by aligned reads from any of the lablab-associated *Xap* genomes.
Breadth of coverage was based on alignments of the genomic sequences reads against the reference pan-genome using BWA-MEM.

Accession	Description
KHS05752	hypothetical protein
KHS33391	hypothetical protein
KHS33392	chemotaxis protein
KHS33542	LysR family transcriptional regulator
KHS34259	LysR family transcriptional regulator
KHS34847	allophanate hydrolase
KHS34848	urea carboxylase
KHS35371	ATPase AAA
KHS35992	hypothetical protein
KHS37131	membrane protein
KHS37132	membrane protein
KHS37135	short-chain dehydrogenase
KHS37136	FAD-linked oxidase
KHS37137	membrane protein
KHS37138	membrane protein
KHS37723	RNA polymerase sigma70
KHS37724	membrane protein
KHS38567	hypothetical protein
KHS38568	peptidase S8
KHS39898	hypothetical protein
KHS40273	hypothetical protein
KHS40963	peptidase S9
KHS41282	hypothetical protein
KHS41283	hypothetical protein
KHS41455	2OG-Fe(II) oxygenase
PRJNA268142:QQ30 24300	aminopeptidase N
PRJNA270010:RN19 24880	transposase

Table S6. Genes found in lablab-associated *Xap* and *Xap* GL1 but absent from *Xff*.
Each of these genes was covered over at least 95% of its length by sequence reads from each of the sequenced lablab-associated *Xap* and *Xap* GL1 genomes but covered by no more than 30% of its length by aligned reads from any of the *Xff* genomes.
Breadth of coverage was based on alignments of the genomic sequences reads against the reference pan-genome using BWA-MEM.

Table S7. Genes found in lablab-associated *Xff* and *Xap* GL1 but absent from *Xap* GL1.
Each of these genes was covered over at least 95% of its length by sequence reads from each of the sequenced lablab-associated and *Xff* genomes but covered by no more than 30% of its length by aligned reads from any of the *Xap* GL1 genomes.
Breadth of coverage was based on alignments of the genomic sequences reads against the reference pan-genome using BWA-MEM.

Accession	Description
KGK66335	hypothetical protein
KGU50240.1	transposase
KGU50511.1	membrane protein
KGU50514.1	membrane protein
KGU50785.1	short-chain dehydrogenase
KGU50800.1	death-on-curing protein

KGU50805.1	hypothetical protein
KGU50861.1	histidine kinase
KGU50862.1	D-Ala-D-Ala carboxypeptidase
KGU50874.1	transcriptional regulator
KGU50901.1	HmsH protein
KGU50902.1	hemin storage protein
KGU50903.1	N-glycosyltransferase
KGU50904.1	membrane protein
KGU50924.1	ABC transporter ATP-binding protein
KGU50925.1	peptide ABC transporter permease
KGU50926.1	ABC transporter permease
KGU50927.1	peptide ABC transporter substrate-binding
KGU50928.2	acyl-CoA dehydrogenase
KGU50929.1	ligand-gated channel
KGU50930.1	FMN reductase
KGU50931.1	alkanesulfonate monooxygenase
KGU50932.1	ABC transporter substrate-binding protein
KGU50933.1	ABC transporter permease
KGU50934.1	sulfonate ABC transporter ATP-binding protein
KGU50935.1	monooxygenase
KGU50936.1	Fis family transcriptional regulator
KGU50938.1	hypothetical protein
KGU51006.1	TonB-dependent receptor
KGU51239.1	UDP-glucose 6-dehydrogenase
KGU51380.1	hypothetical protein
KGU51386.1	hypothetical protein
KGU51481.1	membrane protein
KGU51655.1	glycerophosphodiester phosphodiesterase
KGU51667.2	glycosyl transferase
KGU51668.1	membrane protein
KGU51715.1	hypothetical protein
KGU51716.1	alpha/beta hydrolase
KGU51724.1	Oar protein
KGU51740.1	hypothetical protein
KGU51858.1	hypothetical protein
KGU51985.1	UDP-glucose 4-epimerase
KGU51986.1	glycosyl transferase
KGU51987.1	UDP-galactopyranose mutase
KGU51988.1	beta-glucosidase
KGU52001.1	ligand-gated channel
KGU52007.1	hypothetical protein
KGU52008.1	ABC-type phosphate transport system periplasmic
KGU52394.1	hypothetical protein
KGU53190.1	hypothetical protein
KGU53243.1	ketoacyl reductase
KGU53262.1	UDP-3-O-(3-hydroxymyristoyl) glucosamine
KGU53263.1	ribosomal subunit interface protein
KGU53264.1	acetyltransferase
KGU53265.1	3-oxoacyl-ACP reductase
KGU53266.1	3-oxoacyl-ACP reductase
KGU53267.1	3-oxoacyl-ACP synthase
KGU53341.1	TonB-dependent receptor
KGU53342.1	beta-galactosidase
KGU53343.1	membrane protein
KGU53368.1	aspartate aminotransferase
KGU53369.1	phenazine biosynthesis protein PhzF
KGU53370.1	permease
KGU53546.1	hypothetical protein
KGU53684.1	hypothetical protein
KGU54168.1	type III secretion system effector protein

KGU54174.1	hypothetical protein
KGU54281.2	hypothetical protein
KGU54427.1	LysR family transcriptional regulator
KGU54428.1	membrane protein
KGU54447.1	TonB-dependent receptor
KGU54455.1	flavodoxin
KGU54482.1	hypothetical protein
KGU54485.1	hypothetical protein
KGU54486.1	hypothetical protein
KGU54487.2	transcriptional regulator
KGU54770.1	beta glucosidase
KGU54771.1	ribokinase
KGU54773.2	hypothetical protein
KGU54774.1	TonB-dependent receptor
KGU54958.1	esterase
KGU54981.1	chemotaxis protein
KGU54983.1	sulfate transporter
KGU54984.1	peptidase M20
KGU55170.1	NAD(P) transhydrogenase subunit beta
KGU55175.1	transcriptional regulator
KGU55215.1	membrane protein
KGU55220.1	acetaldehyde dehydrogenase
KGU55225.1	beta-lactamase
KGU55237.1	aldehyde dehydrogenase
KGU55325.1	TonB-dependent receptor
KGU55326.1	nitrilotriacetate monooxygenase
KGU55327.1	L-glyceraldehyde 3-phosphate reductase
KGU55328.1	sulfonate ABC transporter substrate-binding
KGU55817.1	IroE protein
KGU55818.1	outer membrane receptor protein
KGU55837.1	hypothetical protein
KGU55974.1	hypothetical protein
KGU56088.1	hypothetical protein
KGU56145.1	alpha/beta hydrolase
KHS05367	beta-lactamase
KHS05368	transcriptional regulator
KHS05493	fimbrial protein
KHS05622	TetR family transcriptional regulator
KHS05658	DeoR faimly transcriptional regulator
KHS05953	alpha-N-acetylglucosaminidase
KHS06150	recombinase
KHS06212	hypothetical protein
KHS06460	chemotaxis protein
KHS06531	monooxygenase
KHS06947	hypothetical protein
KHS07522	hypothetical protein
KHS07912	hypothetical protein
KHS08114	hypothetical protein
KHS08290	hypothetical protein
KHS08753	hypothetical protein
KHS08754	glyoxalase
KHS08825	hypothetical protein
KKW48699.1	hydrolase

References

Bart, R., Cohn, M., Kassen, A., McCallum, E. J., Shybut, M., Petriello, A., Krasileva, K., Dahlbeck, D., Medina, C., Alicai, T., et al. (2012). High-throughput genomic sequencing of cassava bacterial blight strains identifies conserved effectors to target for durable resistance. *Proc. Natl. Acad. Sci. U. S. A.* 109, E1972–9..

Darling, A. C. E., Mau, B., Blattner, F. R., and Perna, N. T. (2004). Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14, 1394–1403.

Darling, A. C. E., Mau, B., and Perna, N. T. (2010). progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLoS One* 5, e11147.

Darrasse, A., Carrère, S., Barbe, V., Boureau, T., Arrieta-Ortiz, M. L., Bonneau, S., Briand, M., Brin, C., Cociancich, S., Durand, K., et al. (2013). Genome sequence of *Xanthomonas fuscans* subsp. *fuscans* strain 4834-R reveals that flagellar motility is not a general feature of xanthomonads. *BMC Genomics* 14, 761.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072–5.

Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., and Otto, T. D. (2013). REAPR: a universal tool for genome assembly evaluation.

Genome Biol. 14, R47.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 3. Available at: http://arxiv.org/abs/1303.3997 [Accessed July 20, 2014].

Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* 30, 1–9.

Rissman, A. I., Mau, B., Biehl, B. S., Darling, A. E., Glasner, J. D., and Perna, N. T. (2009). Reordering contigs of draft genomes using the Mauve aligner. *Bioinformatics* 25, 2071–3.

Da Silva, A. C. R., Ferro, J. A., Reinach, F. C., Farah, C. S., Furlan, L. R., Quaggio, R. B., Monteiro-Vitorello, C. B., Van Sluys, M. A., Almeida, N. F., Alves, L. M. C., et al. (2002). Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. *Nature* 417, 459–63.

Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings Bioinforma.* 14 , 178–192.

References

1. Gilbertson, R. L., Otoya, M. M., Pastor Corrales, M. A. & Maxwell, D. P. Genetic diversity in common blight bacteria is revealed by cloned repetitive DNA sequences. (1991).
2. Schaaffhausen, R. v. *Dolichos lablab* or hyacinth bean. *Econ. Bot.* **17**, 146 (1963).
3. Schaad, N. W. *et al.* Reclassification of *Xanthomonas campestris* pv. *citri* (ex Hasse 1915) Dye 1978 forms A, B/C/D, and E as *X. smithii* subsp. *citri* (ex Hasse) sp. nov. nom. rev. comb. nov., *X. fuscans* subsp. *aurantifolii* (ex Gabriel 1989) sp. nov. nom. rev. comb. nov., and *X. alfalfae* subsp. *citrumelo* (ex Riker and Jones) Gabriel et al., 1989 sp. nov. nom. rev. comb. nov.; *X. campestris* pv *malvacearum* (ex Smith 1901) Dye 1978 as *X. smithii* subsp. *smithii* nov. comb. nov. nom. nov.; *X. campestri*s pv. *alfalfae* (ex Riker and Jones, 1935) Dye 1978 as *X. alfalfae* subsp. *alfalfae* (ex Riker et al., 1935) sp. nov. nom. rev.; and "var. *fuscans*" of *X. campestris* pv. *phaseoli* (ex Smith, 1987) Dye 1978 as *X. fuscans* subsp. *fuscans* sp. nov. *Syst. Appl. Microbiol.* **28**, 494–518 (2005).
4. Rodriguez-R, L. M. *et al.* Genomes-based phylogeny of the genus *Xanthomonas*. *BMC Microbiol.* **12**, 43 (2012).
5. Hajri, A. *et al.* A «repertoire for repertoire» hypothesis: Repertoires of type three effectors are candidate determinants of host specificity in *Xanthomonas*. *PLoS One* **4**, e6632 (2009).

Chapter 4:

Identification of potential virulence markers from *Campylobacter jejuni* isolates

Work from this chapter was published in:

Harrison JW, Dung TT, Siddiqui F, Korbrisate S, Bukhari H, Tra MP, Hoang NV, Carrique-Mas J, Bryant J, Campbell JI, Studholme DJ, Wren BW, Baker S, Titball RW, Champion OL. Identification of possible virulence marker from*Campylobacter jejuni* isolates. Emerg Infect Dis 20(6):1026–1029

This paper was cited by:

1. Ungureanu, V. A. *et al.* Virulence of a T6SS *Campylobacter jejuni* chicken isolate from North Romania. *BMC Res. Notes* **12**, 1–7 (2019).
2. Sainato, R. *et al.* Epidemiology of *campylobacter* infections among children in Egypt. *Am. J. Trop. Med. Hyg.* **98**, 581–585 (2018).
3. Iglesias-Torrens, Y. *et al.* Population structure, antimicrobial resistance, and virulence-associated genes in *campylobacter jejuni* isolated from three ecological niches: Gastroenteritis patients, broilers, and wild birds. *Front. Microbiol.* **9**, 1–13 (2018).
4. Clark, C. G. *et al.* Comparison of genomes and proteomes of four whole genome-sequenced *Campylobacter* jejuni from different phylogenetic backgrounds. *PLoS One* **13**, 1–28 (2018).
5. Singh, A., Nisaa, K., Bhattacharyya, S. & Mallick, A. I. Immunogenicity and protective efficacy of mucosal delivery of recombinant hcp+ of *Campylobacter jejuni* Type VI secretion system (T6SS) in chickens. *Mol. Immunol.* **111**, 182–197 (2019).
6. Bokhari, H. Exploitation of microbial forensics and nanotechnology for the monitoring of emerging pathogens. *Crit. Rev. Microbiol.* **44**, 504–521

(2018).

7. Borges, V. *et al. Helicobacter pullorum* Isolated from Fresh Chicken Meat: Antibiotic Resistance and Genomic Traits of an Emerging Foodborne Pathogen. *Appl. Environ. Microbiol.* **81**, 8155–8163 (2015).

8. Sima, F. *et al.* A novel natural antimicrobial can reduce the in vitro and in vivo pathogenicity of T6SS positive *Campylobacter jejuni* and *Campylobacter coli* chicken isolates. *Front. Microbiol.* **9**, 1–11 (2018).

9. Singh, A. & Mallick, A. I. Role of putative virulence traits of *Campylobacter jejuni* in regulating differential host immune responses. *J. Microbiol.* **57**, 298–309 (2019).

10. Bronnec, V. *et al.* Adhesion, biofilm formation, and genomic features of *Campylobacter jejuni* Bf, an atypical strain able to grow under aerobic conditions. *Front. Microbiol.* **7**, 1–14 (2016).

11. Minh Nguyen, T. N. Thermophilic *Campylobacter* - Neglected Foodborne Pathogens in Cambodia, Laos and Vietnam. *Gastroenterol. Hepatol. Open Access* **8**, 1–8 (2018).

12. Nguyen, T. N. M. *et al.* Genotyping and antibiotic resistance of thermophilic *Campylobacter* isolated from chicken and pig meat in Vietnam. *Gut Pathog.* **8**, 1–11 (2016).

13. Corcionivoschi, N. *et al.* Virulence characteristics of hcp+ *Campylobacter jejuni* and *Campylobacter coli* isolates from retail chicken. *Gut Pathog.* **7**, 1–11 (2015).

14. Siddiqui, F. *et al.* Molecular detection identified a type six secretion system in *Campylobacter jejuni* from various sources but not from human cases. *J. Appl. Microbiol.* **118**, 1191–1198 (2015).

15. Noreen, Z. *et al.* Structural basis for the pathogenesis of *Campylobacter*

jejuni Hcp1, a structural and effector protein of the Type VI Secretion System. *FEBS J.* **285**, 4060–4070 (2018).

16. Ugarte-Ruiz, M. *et al.* Prevalence of Type VI Secretion System in Spanish *Campylobacter jejuni* Isolates. *Zoonoses Public Health* **62**, 497–500 (2015).
17. An, J.-U. *et al.* Dairy Cattle, a Potential Reservoir of Human Campylobacteriosis: Epidemiological and Molecular Characterization of *Campylobacter jejuni* From Cattle Farms. *Front. Microbiol.* **9**, 1–12 (2018).
18. Duong, V. T. *et al.* No Clinical benefit of empirical antimicrobial therapy for pediatric diarrhea in a high-usage, high-resistance setting. *Clin. Infect. Dis.* **66**, 504–511 (2018).
19. Kovanen, S. *et al.* Population genetics and characterization of *Campylobacter jejuni* isolates from western jackdaws and game birds in Finland. *Appl. Environ. Microbiol.* **85**, 1–16 (2019).
20. Agnetti, J. *et al.* Clinical impact of the type VI secretion system on virulence of *Campylobacter* species during infection. *BMC Infect. Dis.* **19**, 237 (2019).
21. Corry, J. E. L. & Atabay, H. I. Poultry as a source of *Campylobacter* and related organisms. *J. Appl. Microbiol.* **90**, 96S-114S (2001).

Introduction

The genus *Campylobacter* (meaning “curved bacteria”) is a group of Gram-negative, microaerophilic, spiral shaped, motile epsilon proteobacteria [1]. *Campylobacter* species are the principle bacterial cause of human foodborne enterocolitis worldwide [2] and as such cause untold suffering. Included within the *Campylobacter* genus there are seventeen species including well-known strains such as *Campylobacter coli*, *Campylobacter fetus* and *Campylobacter pylori*. *Campylobacter* species have been shown to colonise a number of diverse habitats, including livestock, poultry, humans [3]. However, the species primarily associated with human infection are *C. coli* and *C. jejuni* [4]. Within the genus and indeed within each species campylobacters display a variety of pathogenic effects and adaptation to a wide variety of ecological niches, all suggestive of a high level of genomic diversity within this genus.

Campylobacter jejuni is one of the most well-known of the *Campylobacter* species and much study has been devoted to its biology. The first full genome sequence of *C. jejuni* NCTC 11168 was published in 2000 [1]. The genome of *C. jejuni* is one of the smaller bacterial genomes possessing one circular chromosome of 1,641,481 b.p. with a GC content of around 31%. The genome of *C. jejuni* NCTC 11168 was predicted to code for 1654 proteins and 54 RNAs. The *C. jejuni* genome is unusual in that it codes for very few insertion sequences, phage associated sequences or repeat sequences. Comparative studies have highlighted other unusual features associated with this pathogen for example hypervariable homopolymeric tracks and unusual lipooligosaccharide biosynthesis clusters [4].

C. jejuni has been shown to be the major cause of *Campylobacter*-associated diarrhoea in humans, causing more than 640,000 cases of diarrhoea in 2011 in the United Kingdom alone [5] . This causes great distress for the sufferers and a heavy burden on health services. *C. jejuni* has been shown to be a zoonotic pathogen that forms part of the commensal flora in the gastrointestinal tract of birds such as chickens. This facilitates the infection process as a major route of *C. jejuni* transmission to humans is via the handling and consumption of undercooked or raw chicken [6].

Despite the fact that *C. jejuni* has been shown to be a dominant global diarrhoeal pathogen with great impact on human health and health services, unlike many other common enteric pathogens, the mechanisms of pathogenesis of *C. jejuni* are not well understood. Additionally, confounding the complete understanding of the biology of this important human pathogen, there is a clear bias in our understanding of *Campylobacter* epidemiology. The weight of research effort into this subject has predominantly been concerned with *C. jejuni* infection in high-income countries, neglecting the low-income countries where *C. jejuni* infection is also rife. This is perhaps surprising as diarrhoeal disease is a leading cause of morbidity and mortality among children in Asia and *C. jejuni* has been shown to be a major cause of this disease burden [7,8]. However, a potentially interesting association has arisen; there is a difference in disease phenotype in infected individuals in low- and middle-income countries versus high-income countries. Those individuals diagnosed with *Campylobacter* infections in low- and middle-income countries have been reported to suffer from non-inflammatory, watery diarrhoea. Conversely, *Campylobacter* infection diagnosed in Europe and North America are typically associated with inflammatory, bloody diarrhoea. This observation suggests that the mechanism of the disease is not identical but varies across the geographical locales [9].

These obsevations support a level of intra-species genetic variation between those isolates causing disease in the two geographical areas.

As stated above, a novel class of protein translocation system has recently been identified in Gram-negative bacteria, the type 6 secretion system (T6SS). The role of this novel protein translocation system has been suggested to include pathogen-pathogen and host-pathogen interactions. The T6SS has been found to play a major role in virulence in a range of pathogens, including *Vibrio cholera* [10–13](reviewed in [14,15]). Unlike other enteric pathogens, such as *Salmonella* spp. and *E. coli*, in *C. jejuni* classical virulence determinants such as type 3 secretion systems, insertion sequences or phage associated sequences have not been identified during genomic analysis (Parkhill et al., 2001). However, a functional T6SS was recently identified in *C. jejuni* (Lertpiriyapong et al., 2012). The newly discovered *C. jejuni* T6SS has several important roles in the virulence of this important enteric pathogen, including cell adhesion and invasion in colonic epithelial and macrophage cells and colonization of mice (Lertpiriyapong et al., 2012). However, the role the T6SS plays during the infection cycle in humans has not been investigated.

The hemolysin co-regulated protein (*hcp*), is a highly conserved component of all characterized T6SS, including the functional *C. jejuni* T6SS (Das et al., 2000; Ishikawa et al., 2012; Parkhill et al., 2001; Parsons & Heffron, 2005; Pukatzk et al., 2006). It is believed that the *hcp* gene encodes either part of the translocation apparatus, or a secreted effector protein (Records, 2011). Prior to this project, the T6SS was identified in isolates from global studies of campylobacteriosis confirming the results published by Lertpyiapong et al (O. Champion - personal communication).

The aim of this study was to address the bias in *C. jejuni* genome sequencing data towards strains isolated in high-income regions, thus increasing both the volume and diversity of *C. jejuni* genomic resources. To survey the presence of the newly identified T6SS gene cluster over the full diversity of sequenced *C. jejuni* strains including those added by this study. Further to this the aim was to determine if there was a molecular marker that could be used to identify strains harbouring this marker and thus the T6SS gene cluster. This marker will then be used to survey strains from the UK and the far east to determine first whether there was an association between strains harbouring a T6SS with the more virulent form of C. jejuni infection and if these strains were significantly associated with a particular geographic region.

Author contribution

The author conducted all bioinformatic analysis for this project. This included using bespoke scripts and pipeline code to conduct the initial quality control of raw sequencing reads and the *de novo* assembly and analysis of all 12 novel strains. The author also conducted a comprehensive bioinformatic screen of the type 3 secretion system gene cluster across a large panel of *C. jejuni* strains along with genomic comparisons, sequence extraction and MLSA.

The author also contributed significantly to the pre-project research, concept design and planning for the project along with the writing, editing and submission of manuscript and the production and editing of all figures and tables

Manuscript

Identification of Possible Virulence Marker from *Campylobacter jejuni* Isolates

James W. Harrison, Tran Thi Ngoc Dung, Fariha Siddiqui, Sunee Korbrisate, Habib Bukhari, My Phan Vu Tra, Nguyen Van Minh Hoang, Juan Carrique-Mas, Juliet Bryant, James I. Campbell, David J. Studholme, Brendan W. Wren, Stephen Baker, Richard W. Titball, and Olivia L. Champion

A novel protein translocation system, the type-6 secretion system (T6SS), may play a role in virulence of *Campylobacter jejuni*. We investigated 181 *C. jejuni* isolates from humans, chickens, and environmental sources in Vietnam, Thailand, Pakistan, and the United Kingdom for T6SS. The marker was most prevalent in human and chicken isolates from Vietnam.

Campylobacter species are the principal bacterial cause of human foodborne enterocolitis worldwide (*1*). Despite the global significance of *C. jejuni* as a leading cause of diarrheal disease (*2*), the mechanisms of pathogenesis of *C. jejuni* are not well understood. Research on *Campylobacter* epidemiology has largely been conducted in high-income countries and therefore may not be representative of global patterns.

Recently, a novel class of protein translocation system was identified in gram-negative bacteria. This system, named the type-6 secretion system (T6SS), has been found to play roles in pathogen–pathogen and host–pathogen interactions and has a major effect on virulence in a range of pathogens, including *Vibrio cholerae* (*3*–*6*) (reviewed in *7*,*8*). A functional T6SS was recently identified in *C. jejuni* (*9*,*10*) and found to have several roles in virulence, influencing cell adhesion, cytotoxicity toward erythrocytes, and colonization of mice (*9*,*10*). However, it is unknown whether T6SS changes the effects of these pathogens in human infection.

In this study, we aimed to determine whether presence of T6SS in *C. jejuni* is potentially a marker associated with more severe human disease. Moreover, because human infection with *C. jejuni* is often associated with contact with poultry, we investigated whether poultry harbor *C. jejuni* that possess T6SS.

Author affiliations: University of Exeter, Exeter, UK (J.W. Harrison, D.J. Studholme, R.W. Titball, O.L. Champion); The Hospital for Tropical Diseases, Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam (T.T.N. Dung, M.P.V. Tra, N.V.M. Hoang, J. Carrique-Mas, J. Bryant, J.I. Campbell, S. Baker); Comsats University, Islamabad, Pakistan (F. Siddiqui, H. Bukhari); Mahidol University, Bangkok, Thailand (S. Korbrisate); University of Oxford, Oxford, UK (J. Carrique-Mas, J. Bryant, J.I. Campbell, S. Baker); and London School of Hygiene and Tropical Medicine, London, UK (B.W. Wren)

DOI: http://dx.doi.org/10.3201/eid2006.130635

The Study

To partially address bias toward study of *C. jejuni* strains from high-income countries and the under-representation of strains from Asia in previous studies, we previously sequenced the genomes of 12 clinical isolates of *C. jejuni* from Asia: 4 from Thailand, 3 from Pakistan, and 5 from Vietnam (J. Harrison, unpub. data; Figure 1). We noted that 8 (67%) of these isolates possessed a cluster of genes homologous to the recently described T6SS (Figure 1). This finding was in contrast to findings regarding previously sequenced *C. jejuni* genomes; only 10 (14%) of 71 previously sequenced *C. jejuni* strains possessed an apparently intact T6SS gene cluster (Figure 1; full listing of genomes is in online Technical Appendix Table 1, wwwnc.cdc.gov/EID/article/20/6/13-0635-Techapp1.pdf). Several other strains from our study and previously sequenced strains contained ≥1 T6SS genes but not a complete T6SS cluster. Figure 1 shows the presence and absence of each T6SS gene in each available genome sequence (J. Harrison, unpub. data) and the previously sequenced strains. A nonrandom distribution of T6SS can be seen across the phylogenetic diversity of *C. jejuni*; T6SS is limited to certain clades, and degeneration of the T6SS gene cluster apparently occurs in parallel within several of those clades (Figure 1).

Our genome sequencing analysis indicated that strains possessing a complete T6SS cluster could be distinguished by the presence of the *hcp* gene (Figure 1) (*9*,*10*). Therefore, we used *hcp* as a proxy for determining the presence of a functional T6SS in 181 *C. jejuni* isolates from chickens, humans, and environmental sources (collections of the Oxford University Clinical Research Unit and the University of Exeter; online Technical Appendix Table 2). We designed and used a multiplex PCR (online Technical Appendix Table 3) to screen for the presence of *hcp* in these isolates; the conserved *C. jejuni* housekeeping gene, *gltA*, was used as a positive control.

Of the 181 isolates, 28 originated from chickens in the United Kingdom and 21 from chickens in Vietnam. The *hcp* gene was found significantly more often in isolates

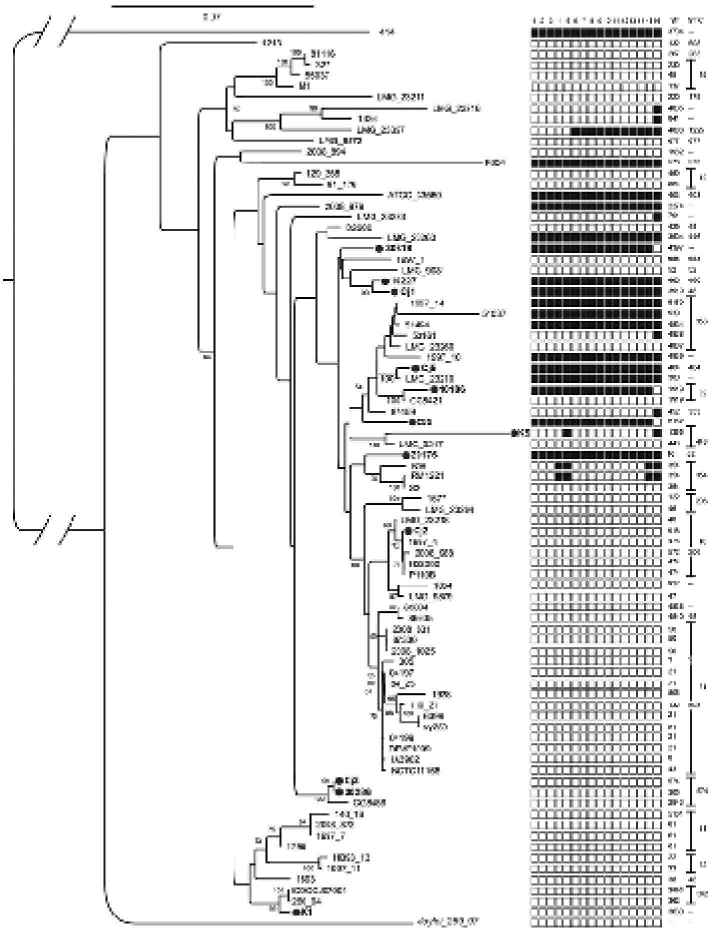


Figure 1. Distribution of the type-six secretion system (T6SS) marker across the phylogenetic diversity of *Campylobacter jejuni* strains, as determined by multilocus sequence analysis. We generated a maximum-likelihood tree from concatenated nucleotide alignments of 31 housekeeping genes; nucleotide sequences were aligned by using MUSCLE (www.drive5.com/muscle) and masked by using GBLOCKS (http://molevol.cmima.csic.es/castresana/Gblocks.html). Maximum-likelihood analysis was done by using the GTR model in PhyML (http://code.google.com/p/phyml/). Numbers on nodes denote bootstrap values (1,000 bootstrap replicates); values <50 are not shown. Black circles indicate strains whose genomes were sequenced in this study (GenBank accession nos. AUUQ00000000, AUUP00000000, AUUO00000000, AUUN00000000, AUUM00000000, AUUL00000000, AUUK00000000, AUUJ00000000, AUUI00000000, ARWS00000000, AUUH00000000, AUUG00000000). We inferred the presence/absence of each of the T6SS genes on the basis of TBLASTN (http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch) searches against the predicted proteins sequences from *C. jejuni* strain 414 (National Center for Biotechnology Information reference sequence no. NZ_CM000855). Presence or absence of each gene is indicated by a black or white square, respectively, for each strain: column 1, *hcp*; column 2, *icmF*_1; column 3, *icmF*_2; column 4, *vasK*; column 5, *FHA*; column 6, *vasF*; column 7, *vasE*; column 8, *vasD*; column 9, *impA*; column 10, *impD*; column 11, *impC*; columns 12 and 13, conserved hypotheticals; column 14, *vasA*; column 15, *vasB*; column 16, *vgrg*. The sequence type (ST) and ST complex (STC) columns represent global multilocus sequence types as described by the Oxford multilocus sequence typing scheme (http://pubmlst.org). ?, unknown ST; –, isolate could not be allocated to a specific ST or STC. Scale bar indicates nucleotide substitutions per site. Further details of the isolates are provided in online Technical Appendix Table 2 (wwwnc.cdc.gov/EID/article/20/6/13-0635-Techapp1.pdf).

from Vietnam (15 [71.4%] isolates) than in those from the United Kingdom (1 [3.5%] isolate) (p<0.01 by 2-sample *Z*-test; online Technical Appendix Figure 1). An additional 38 of the isolates were from humans in the United Kingdom and 33 from humans in Vietnam; again, the *hcp* gene was significantly more prevalent in isolates from Vietnam (20 [60.6%] isolates) than those from the United Kingdom (1 [2.6%] isolate) (p<0.01 by 2-sample *Z*-test; online Technical Appendix Figure 2).

We also found that patients infected with *hcp*-positive *C. jejuni* experienced bloody diarrhea more commonly than those infected with *hcp*-negative *C. jejuni*. For the 36 isolates for which detailed clinical data on patients were available, 6 (31.6%) of 19 patients in Vietnam who were infected with *hcp*-positive *C. jejuni* had bloody diarrhea, compared with 1 (5.9%) of 17 patients infected with *hcp*-negative *C. jejuni* (p<0.05 by 2-sample *Z*-test) (Figure 2). These results suggest a potential correlation between T6SS and bloody diarrhea, a serious clinical manifestation of the infection that results in higher rates of hospitalization and greater need for treatment with antimicrobial drugs (*11*). Moreover, *Campylobacter*-related septicemia developed in the 1 patient in the United Kingdom who was infected with a T6SS-positive strain (*11*). These data suggest that infection with the *C. jejuni* T6SS genotypic strains is associated with more severe disease. However, for sample bias to be ruled out, a comprehensive study is required in which the prevalence of T6SS is measured in *C. jejuni* samples from patients with mild and severe forms of infection.

We found a number of *C. jejuni* strains from humans and poultry that possessed the T6SS cluster, although some strains showed a slightly modified gene order (online Technical Appendix Table 1 and Figure 3). However, most (61 [85.9%] of 71) of the previously sequenced *C. jejuni* isolates lacked a complete T6SS gene cluster (Figure 1); this finding might explain why T6SS was not discovered in *C. jeuni* sooner. Conversely, our PCR-based study frequently identified the *hcp* marker in isolates from Thailand, Pakistan, and Vietnam (Table). We cannot be certain that all of the isolates with the *hcp* marker possessed a complete and functional T6SS gene cluster, but the *hcp* gene is consistently associated with the presence of a complete T6SS cluster in all available sequenced *C. jejuni* genomes (Figure 1). This correlation lends confidence to the use of *hcp* as a proxy.

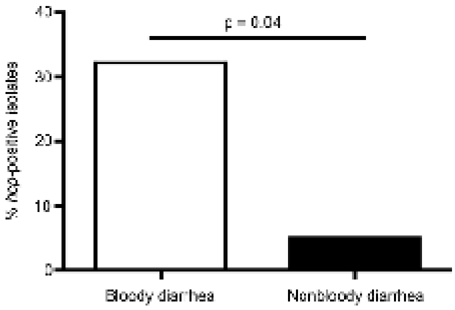


Figure 2. Percentage of *hcp*-positive *Campylobacter jejuni* strains isolated from patients in Vietnam who had bloody diarrhea and nonbloody diarrhea. Patients who were hospitalized because of *C. jejuni* infection were scored for the presence of bloody diarrhea or nonbloody diarrhea, and presence of the *hcp* type-six secretion system (T6SS) marker in strains isolated from the patients was determined. Of patients with bloody diarrhea, 32% were infected with *hcp*-positive strains; of patients with nonbloody diarrhea, 5% were infected with *hcp*-positive strains.

Poultry are a well-documented reservoir of human *Campylobacter* infection (*12*). We found that *Campylobacter* strains harboring the *hcp* marker were significantly associated with chickens in Asia. Large numbers of poultry are imported into North America and Europe from low-income countries, including Thailand (*13*). This process could introduce T6SS-positive *Campylobacter* genotypes into the food chains of the importing countries, posing a potential emerging threat to public health.

Conclusions

Our results suggest that the T6SS may be more prevalent in *C. jejuni* in Vietnam, Pakistan, and Thailand than in the United Kingdom. Furthermore, our results suggest that *hcp* may be a marker associated with severe human disease caused by *C. jejuni* infection in Vietnam, although there is no evidence that the association is causal. Chickens imported from these countries could be a source of *hcp*-positive strains and may have the potential to cause severe human infection.

Table. Overview of *Campylobacter jejuni* strains containing type-six secretion system genetic marker *hcp*, by country and isolate source

	No. *hcp*-positive strains/total no. strains (%)				
Isolate source	United Kingdom	Vietnam	Pakistan	Thailand	Total
Human	1/38 (2.6)	20/33 (60.6)	2/13 (15.4)	1/3 (33.3)	24/87 (27.6)
Chicken	1/28 (3.9)	15/21 (71.4)	1/2 (50)	0	17/51 (33.3)
Other	5/26 (19.2)	1/14 (7.1)	1/3 (33.3)	0	7/43 (16.3)
Total	7/92 (7.6)	36/68 (54.4)	4/18 22.2)	1/3 (33.3)	48/181 (26.5)

Acknowledgments

We thank Konrad Paszkiewicz and Karen Moore for assistance with whole-genome sequencing.

The work was partly supported by the UK Biotechnology and Biological Sciences Research Council, award BB/1024631/1 to R.T., D.S., and O.C.; by a Wellcome Trust Institutional Strategic Support Award (WT097835MF); and by a studentship awarded to J.H.

Mr Harrison is a PhD student at the University of Exeter under the supervision of D.S. His research focuses on using bioinformatic methods to investigate the comparative genomics of emerging diseases and plant-associated microbes.

References

1. Adak GK, Meakins SM, Yip H, Lopman BA, O'Brien SJ. Disease risks from foods, England and Wales, 1996–2000. Emerg Infect Dis. 2005;11:365–72. http://dx.doi.org/10.3201/eid1103.040191
2. Allos BM. *Campylobacter jejuni* infections: update on emerging issues and trends. Clin Infect Dis. 2001;32:1201–6. http://dx.doi.org/10.1086/319760
3. Das S, Chakrabortty A, Banerjee R, Roychoudhury S, Chaudhuri K. Comparison of global transcription responses allows identification of *Vibrio cholerae* genes differentially expressed following infection. FEMS Microbiol Lett. 2000;190:87–91. http://dx.doi.org/10.1111/j.1574-6968.2000.tb09267.x
4. Ishikawa T, Sabharwal D, Bröms J, Milton DL, Sjöstedt A, Uhlin BE, et al. Pathoadaptive conditional regulation of the type VI secretion system in *Vibrio cholerae* O1 strains. Infect Immun. 2012;80:575–84. http://dx.doi.org/10.1128/IAI.05510-11
5. Parsons DA, Heffron F. sciS, an icmF homolog in *Salmonella enterica* serovar *Typhimurium*, limits intracellular replication and decreases virulence. Infect Immun. 2005;73:4338–45. http://dx.doi.org/10.1128/IAI.73.7.4338-4345.2005
6. Pukatzki S, Ma AT, Sturtevant D, Krastins B, Sarracino D, Nelson WC, et al. Identification of a conserved bacterial protein secretion system in *Vibrio cholerae* using the *Dictyostelium* host model system. Proc Natl Acad Sci U S A. 2006;103:1528–33. http://dx.doi.org/10.1073/pnas.0510322103
7. Cascales E. The type VI secretion toolkit. EMBO Rep. 2008;9:735–41. http://dx.doi.org/10.1038/embor.2008.131
8. Mulder DT, Cooper CA, Coombes BK. Type VI secretion system-associated gene clusters contribute to pathogenesis of *Salmonella enterica* serovar *Typhimurium*. Infect Immun. 2012;80:1996–2007. http://dx.doi.org/10.1128/IAI.06205-11
9. Lertpiriyapong K, Gamazon ER, Feng Y, Park DS, Pang J, Botka G, et al. *Campylobacter jejuni* type VI secretion system: roles in adaptation to deoxycholic acid, host cell adherence, invasion, and in vivo colonization. PLoS ONE. 2012;7:e42842. http://dx.doi.org/10.1371/journal.pone.0042842
10. Bleumink-Pluym NMC, van Alphen LB, Bouwman LI, Wösten MMSM, van Putten JPM. Identification of a functional type VI secretion system in *Campylobacter jejuni* conferring capsule polysaccharide sensitive cytotoxicity. PLoS Pathog. 2013;9:e1003393. http://dx.doi.org/10.1371/journal.ppat.1003393
11. Kuşkonmaz B, Yurdakök K, Yalçin SS, Ozmert E. Comparison of acute bloody and watery diarrhea: a case control study. Turk J Pediatr. 2009;51:133–40.
12. Harris NV, Weiss NS, Nolan CM. The role of poultry and meats in the etiology of *Campylobacter jejuni/coli* enteritis. Am J Public Health. 1986;76:407–11. http://dx.doi.org/10.2105/AJPH.76.4.407
13. Food and Agriculture Organization of the United Nations. Agribusiness handbook: poultry meat and eggs. 2010 [cited 2013 Apr 1]. http://www.fao.org/fileadmin/user_upload/tci/docs/1_AH9-Poultry%20Meat%20&%20Eggs.pdf

Address for correspondence: Olivia L. Champion, University of Exeter, Geoffrey Pope Building, Stocker Road, Exeter EX4 4QD, UK; email: O.L.Champion@exeter.ac.uk

Supplementary information

strain name	source	country of origin	T6SS	Genome status	Ref	Hcp +ve
305	Turkey	Germany	negative	draft	(Takamiya et al., 2011)	
327	Turkey	Unknown	negative	draft	(Takamiya et al., 2011)	
414	Bank Vole	Unknown	positive	complete		yes
1213	Cow	USA	negative	draft		
1336	Bird	Unknown	positive	complete	(Hepworth et al., 2011)	
1577	Cow	USA	negative	draft		
1798	Cow	USA	negative	draft		
1854	Cow	USA	negative	draft		
1893	Cow	USA	negative	draft		
1928	Cow	USA	negative	draft		
04197	Unknown	Unknown	negative	draft		
04199	Unknown	Unknown	negative	draft		
6399	Unknown	Unknown	negative	draft		
51037	chicken	USA	positive	draft		yes
51494	chicken	USA	positive	draft		yes
53161	chicken	USA	positive	draft		
60004	chicken	USA	negative	draft		
81116	human	Unknown	negative	complete	(Pearson et al., 2007)	
86605	chicken	USA	negative	draft		
87330	chicken	USA	negative	draft		
87459	chicken	USA	positive	draft		
110_21	Unknown	USA	negative	draft		
129_258	Cow	USA	negative	draft		
140_16	Cow	USA	negative	draft		
1997_1	Human	USA	negative	draft		
1997_10	Human	USA	positive	draft		yes
1997_11	Human	USA	negative	draft		
1997_14	Human	USA	positive	draft		yes
1997_4	Human	USA	negative	draft		
1997_7	Human	USA	negative	draft		
2008_1025	Human	France	negative	draft		
2008_831	Human	France	negative	draft		
2008_872	Human	France	negative	draft		
2008_894	Human	France	negative	draft		
2008_979	Human	France	positive	draft		yes
2008_988	Human	France	negative	draft		
260_94	Human	S. Africa	negative	draft		
81_176	Human	Unknown	negative	Complete	(Russell, Blaser, Sarmiento, & Fox, 1989)	
84_25	Human	Unknown	negative	Complete		
ATCC_33560	Cow	Brussels	positive	draft		yes
CG8421	Human	Thailand	negative	draft	(Poly et al., 2008)	
CG8486	Human	Thailand	negative	draft	(Poly et al., 2007)	
D2600	Human	USA	negative	draft	(Jerome et al., 2012)	
DFVF1099	chicken	Unknown	negative	draft	(Takamiya et al., 2011)	
H22082	Human	New Zealand	negative	draft	(Takamiya et al., 2011)	
HB93_13	Human	China	negative	draft	(Burrough, Sahin, Plummer, Zhang, & Yaeger, 2009)	
IA3902	Sheep	USA	negative	Complete	(Luo et al., 2012)	
ICDCCJ07001	Human	China	negative	draft	(Zhang et al., 2010)	
LMG_23210	chicken	Belgium	positive	draft		Yes
LMG_23211	chicken	Belgium	negative	draft		
LMG_23216	chicken	Belgium	positive	draft		
LMG_23218	chicken	Belgium	negative	draft		
LMG_23223	chicken	Belgium	positive	draft		
LMG_23263	chicken	Bosnia and Herzegovina	positive	draft		yes
LMG_23264	Human	Slovenia	negative	draft		
LMG_23269	chicken	Belgium	negative	draft		
LMG_23357	water	Netherlands	positive	draft		
LMG_9081	human	USA	negative	draft		
LMG_9217	Human	Belgium	negative	draft		

LMG_9872	Human	Sweden	negative	draft		
LMG_9879	Human	Canada	negative	draft		
M1	Human/poultry	Unknown	negative	complete	(Friis et al., 2010)	
NCTC11168	Human	Unknown	negative	complete	(Gundogdu et al., 2007)	
NW	Human	USA	positive	draft	(Jerome et al., 2012)	
P110B	chicken	New Zealand	negative	draft	(Gundogdu et al., 2007)	
P854	chicken	UK	positive	draft		yes
RM1221	Unknown	Unknown	positive	complete	(Fouts et al., 2005)	
S3	poultry	Unknown	negative	complete	(Cooper, Cooper, Zuccolo, Law, & Joens, 2011)	
doylei 269 97	Human	Unknown	negative	complete		
xy259	Unknown	Unknown	negative	draft		
55037	chicken	USA	negative	draft		

Table S1. List of *C. jejuni* strains included in MLSA analysis.

Strain name	source	country of origin	T6SS	Genome status	Ref	Strain source
28766	Beach	UK	negative			This study
KSCattle8	Cattle	UK	negative			This study
11974	human	UK	negative			This study
13305	human	UK	negative			This study
11919	human	UK	negative			This study
30280	human	UK	negative			This study
11818	human	UK	negative			This study
12241	human	UK	negative			This study
99/188	human	UK	negative			This study
99/197	human	UK	negative			This study
99/97	human	UK	negative			This study
0 1/ 43	human	UK	negative			This study
99/189	human	UK	negative			This study
99/216	human	UK	negative			This study
94/229	human	UK	negative			This study
99/212	human	UK	negative			This study
BB1267	human	UK	negative			This study
31467	human	UK	negative			This study
31484	human	UK	negative			This study
32799	human	UK	negative			This study
31485	human	UK	negative			This study
33084	human	UK	positive			This study
93/372	human	UK	negative			This study
32787	human	UK	negative			This study
44119	human	UK	negative			This study
47693	human	UK	negative			This study
33106	human	UK	negative			This study
34007	human	UK	negative			This study
Hi40980306	human	UK	negative			This study
90843	human	UK	negative			This study
Hi40500471	human	UK	negative			This study
Hi40620306	human	UK	negative			This study
BB1267	human	UK	negative			This study
Hi81266	human	UK	negative			This study
Hi80586	human	UK	negative			This study
Hi80547	human	UK	negative			This study
Hi81006	human	UK	negative			This study
KSSAPSM6	human	UK	negative			This study
Hi81214	human	UK	negative			This study
KSSHPSM4	human	UK	negative			This study
99/118	Cow	UK	negative			This study
99/201	Cow	UK	negative			This study
99/202	Cow	UK	negative			This study
C0599 3095	Cow	UK	negative			This study

C085 40995	Cow	UK	negative			This study
1182 ENV	Env	UK	negative			This study
PS304	Pig	UK	negative			This study
PS623	Pig	UK	positive			This study
PS762	Pig	UK	negative			This study
PS830	Pig	UK	negative			This study
PS838	Pig	UK	negative			This study
PS843	Pig	UK	positive			This study
PS849	Pig	UK	positive			This study
PS852	Pig	UK	positive			This study
PS857	Pig	UK	positive			This study
C120/2	Poultry	UK	negative			This study
C132/1	Poultry	UK	negative			This study
D2/T/80	Poultry	UK	negative			This study
PS55491	Poultry	UK	positive			This study
A83515A	Poultry	UK	negative			This study
A1CF12	Poultry	UK	negative			This study
D502009A	Poultry	UK	negative			This study
C3/T2/8	Poultry	UK	negative			This study
D2/27B	Poultry	UK	negative			This study
C3/T/25	Poultry	UK	negative			This study
EX1286	Poultry	UK	negative			This study
MB1	Poultry	UK	negative			This study
MB2	Poultry	UK	negative			This study
MB3	Poultry	UK	negative			This study
MB4	Poultry	UK	negative			This study
MB5	Poultry	UK	negative			This study
MB6	Poultry	UK	negative			This study
MB7	Poultry	UK	negative			This study
MB8	Poultry	UK	negative			This study
MB9	Poultry	UK	negative			This study
MB12	Poultry	UK	negative			This study
MB13	Poultry	UK	negative			This study
MB14	Poultry	UK	negative			This study
MB15	Poultry	UK	negative			This study
MB16	Poultry	UK	negative			This study
MB17	Poultry	UK	negative			This study
MB18	Poultry	UK	negative			This study
S2160509901	Sheep	UK	negative			This study
S390209903	Sheep	UK	negative			This study
S1200409904	Sheep	UK	negative			This study
S8704099	Sheep	UK	negative			This study
S3720509904	Sheep	UK	negative			This study
S3790809901	Sheep	UK	negative			This study
S43503099	Sheep	UK	negative			This study
S4990109905	Sheep	UK	negative			This study
S58503099	Sheep	UK	negative			This study
Cj 54	Camel	Pakistan	negative			This study
N2	human	Pakistan	negative			This study
AKRH011	human	Pakistan	negative			This study
702	human	Pakistan	negative			This study
Y25	human	Pakistan	negative			This study
2960HF	human	Pakistan	negative			This study
712	human	Pakistan	negative			This study
K1	human	Pakistan	negative	draft		This study
K2	human	Pakistan	positive			This study
K4	human	Pakistan	negative			This study
K5	human	Pakistan	negative	draft		This study
K6	human	Pakistan	negative			This study
K7	human	Pakistan	negative			This study
K8	human	Pakistan	positive			This study
80	Poultry	Pakistan	negative			This study
255	Poultry	Pakistan	positive	draft		This study

Cj245	waste water	Pakistan	negative			This study
Cj 236	waste water	Pakistan	positive			This study
Cj1	human	Thailand	positive	draft		This study
Cj2	human	Thailand	negative	draft		This study
Cj3	human	Thailand	negative	draft		This study
Cj5	human	Thailand	positive	draft		This study
20157	human	Vietnam	positive			This study
30286	human	Vietnam	positive	draft		This study
30261	human	Vietnam	positive			This study
10227	human	Vietnam	positive	draft		This study
20160	human	Vietnam	negative			This study
30106	human	Vietnam	negative			This study
20288	human	Vietnam	negative			This study
30311	human	Vietnam	positive			This study
20283	human	Vietnam	positive			This study
10186	human	Vietnam	positive	draft		This study
20176	human	Vietnam	positive	draft		This study
20231	human	Vietnam	positive			This study
20301	human	Vietnam	positive			This study
30318	human	Vietnam	positive	draft		This study
20321	human	Vietnam	positive			This study
20332	human	Vietnam	negative			This study
30355	human	Vietnam	positive			This study
20319	human	Vietnam	positive			This study
20137	human	Vietnam	positive			This study
30391	human	Vietnam	negative			This study
30396	human	Vietnam	negative			This study
10275	human	Vietnam	negative			This study
20227	human	Vietnam	positive			This study
30446	human	Vietnam	positive			This study
20396	human	Vietnam	negative			This study
10126	human	Vietnam	positive			This study
20084	human	Vietnam	negative			This study
30431	human	Vietnam	negative			This study
30146	human	Vietnam	negative			This study
10070	human	Vietnam	negative			This study
10152	human	Vietnam	negative			This study
20245	human	Vietnam	positive			This study
71V103	Duck	Vietnam	negative			This study
71V42	Duck	Vietnam	negative			This study
71V489	Duck	Vietnam	negative			This study
71V151	Duck	Vietnam	negative			This study
71V135	Duck	Vietnam	negative			This study
71V445	Duck	Vietnam	negative			This study
71V484	Duck	Vietnam	negative			This study
71V420	Duck	Vietnam	negative			This study
71V409	Duck	Vietnam	negative			This study
71V397	Duck	Vietnam	negative			This study
71V49	Duck	Vietnam	negative			This study
71V69	Duck	Vietnam	negative			This study
72H57	Pig	Vietnam	negative			This study
71V110	Duck	Vietnam	positive			This study
71G139	Chicken	Vietnam	negative			This study
71G142	Chicken	Vietnam	positive			This study
71G356	Chicken	Vietnam	positive			This study
71G570	Chicken	Vietnam	positive			This study
71G784	Chicken	Vietnam	positive			This study
71G998	Chicken	Vietnam	positive			This study
71G1212	Chicken	Vietnam	positive			This study
71G1426	Chicken	Vietnam	positive			This study
71G1640	Chicken	Vietnam	positive			This study
71G1854	Chicken	Vietnam	positive			This study
71G2068	Chicken	Vietnam	positive			This study
71G2282	Chicken	Vietnam	positive			This study

71G326	Chicken	Vietnam	negative			This study
71G143	Chicken	Vietnam	positive			This study
71G329	Chicken	Vietnam	negative			This study
71G125	Chicken	Vietnam	positive			This study
71G124	Chicken	Vietnam	negative			This study
71G90	Chicken	Vietnam	positive			This study
71G30	Chicken	Vietnam	positive			This study
71G43	Chicken	Vietnam	negative			This study
72G117	Chicken	Vietnam	negative			This study

Table S2. List of 181 *C. jejuni* strains analyzed in this study

Table S3. List of primers

Primers (for target genes)	Primer Sequence (5’----- 3’)	Predicted Amplicon size	Tm	Reference
*gltA*F Cj	GCCCAAAGCCCATCAAGCGGA	142 bp	60	This Study
gltA F Cj	GCGCTTTGGGGTCATGCACA		58	This Study
Hcp F	CAAGCGGTGCATCTACTGAA	463 bp	60	This Study
Hcp R	TAAGCTTTGCCCTCTCTCCA		60	This Study

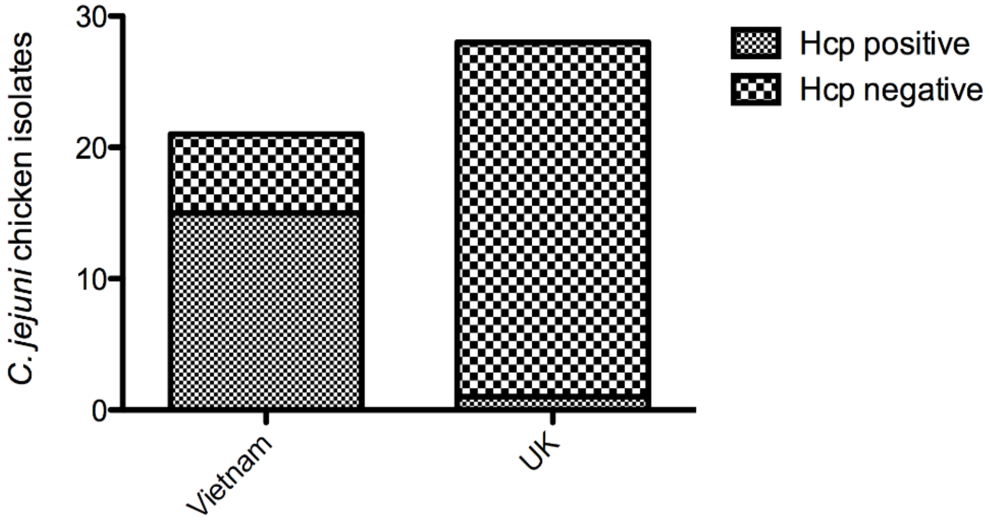


Figure S1: Prevalence of T6SS genetic marker *hcp* in *C. jejuni* isolated from chickens in Vietnam and the UK.

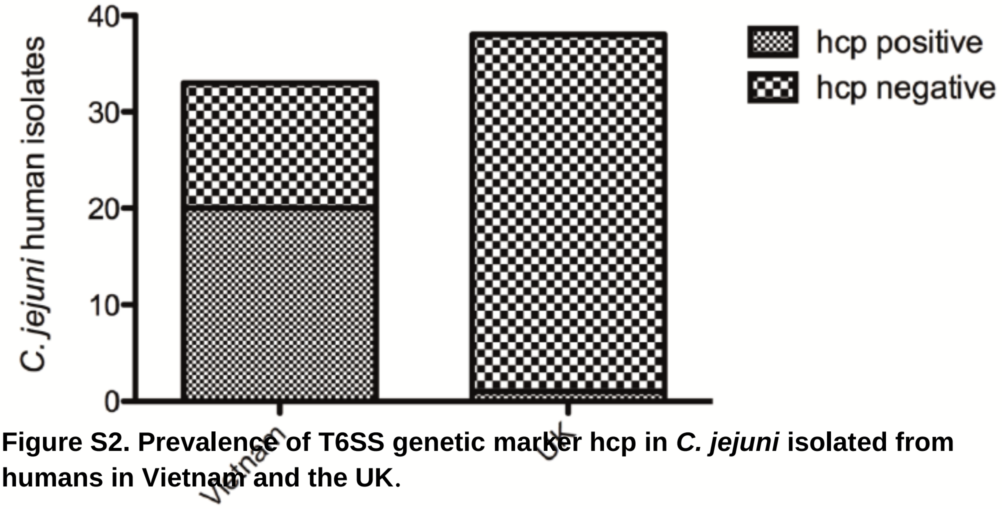


Figure S2. Prevalence of T6SS genetic marker hcp in *C. jejuni* isolated from humans in Vietnam and the UK.

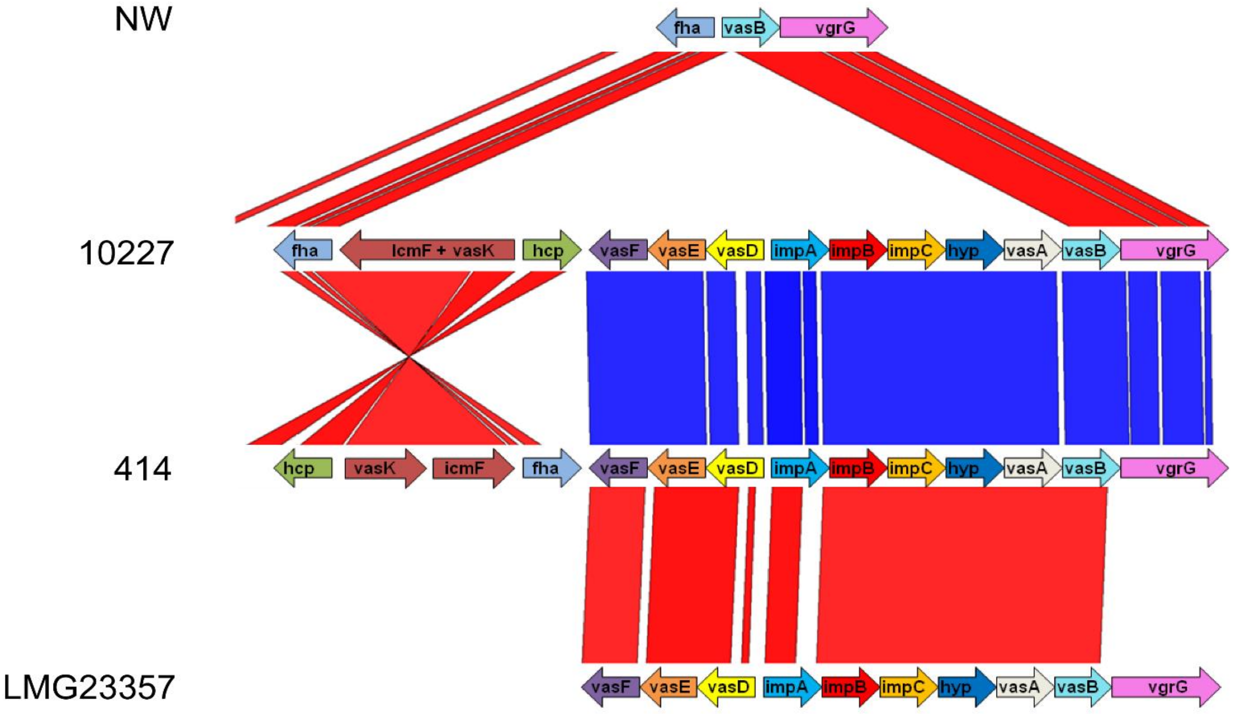


Figure S3. Comparison of the gene orders in the T6SS gene clusters found in *C. jejuni*.

Chapter 5:

The draft genome sequence of *Xanthomonas* species strain Nyagatare, isolated from diseased bean in Rwanda

Work from this chapter was published in:

Aritua V, Musoni A, Kabeja A, Butare L, Mukamuhirwa F, Gahakwa D, Kato F, Abang MM, Buruchara R, Sapp M, Harrison J, Studholme D.J., Smith J. (2015) The draft genome sequence of *Xanthomonas* species strain Nyagatare, isolated from diseased bean in Rwanda. FEMS Microbiol Lett 362(4):1–4.

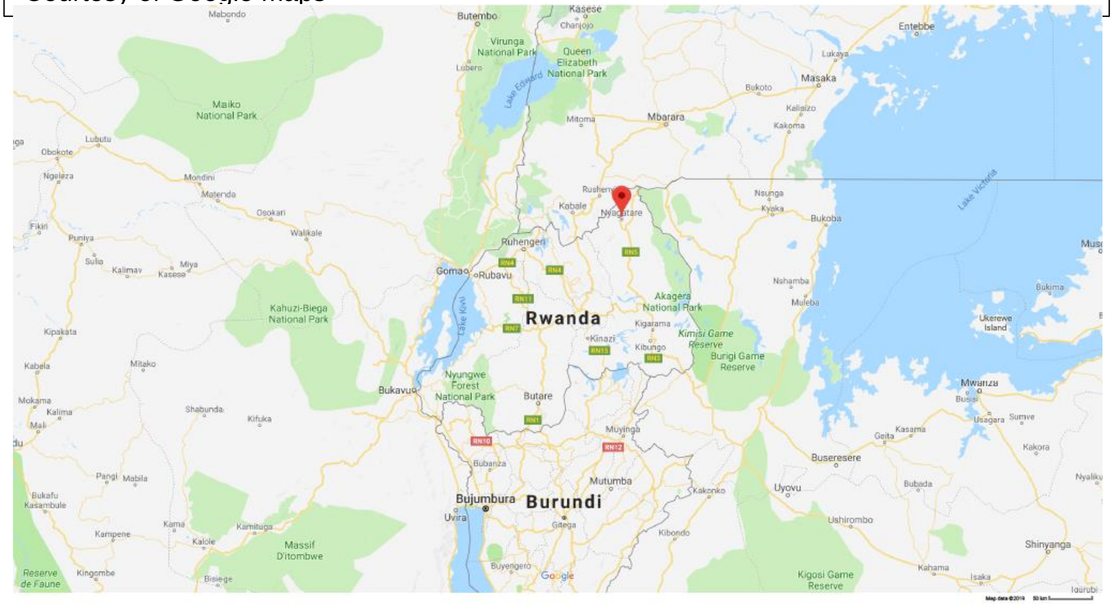
This paper was cited by:

1. Jacobs, J. M., Pesce, C., Lefeuvre, P. & Koebnik, R. Comparative genomics of a cannabis pathogen reveals insight into the evolution of pathogenicity in *Xanthomonas*. *Frontiers in Plant Science* **6**, 431 (2015).
2. Jacques, M.-A. *et al.* Using Ecology, Physiology, and Genomics to Understand Host Specificity in Xanthomonas. *Annu. Rev. Phytopathol.* **54**, 163–187 (2016).
3. Vicente, J. G., Rothwell, S., Holub, E. B. & Studholme, D. J. Pathogenic, phenotypic and molecular characterisation of *Xanthomonas nasturtii* sp. Nov. And *Xanthomonas floridensis* sp. Nov., new species of *Xanthomonas* associated with watercress production in Florida. *Int. J. Syst. Evol. Microbiol.* **67**, 3645–3654 (2017).
4. Meline, V. *et al.* Role of the acquisition of a type 3 secretion system in the emergence of novel pathogenic strains of *Xanthomonas*. *Mol. Plant Pathol.* **20**, 33–50 (2019).

Introduction

This chapter introduces a newly described species of *Xanthomonas*, isolated from an outbreak of disease on common bean (*Phaseolus vulgaris*) crops in the Nyagatare region of the east African country of Rwanda (Figure 7). This novel *Xanthomonas* isolate shows unusual disease symptoms and whilst at present does not yet pose a significant agro-economic threat in the region, there is the potential for this pathogen to spread and impact crop production and consequently the wellbeing of the local population. Also, given the global nature of agro-economy is perfectly feasible that pathogens such as this can be spread widely though the plant and seed trade therefore it is important to develop fast comprehensive methods to assess these threats. Tentatively named *Xanthomonas* sp. Nyagatare, this new isolate is the first representative to be sequenced from a newly described species level clade.

Figure 7: Map showing location of the Nyagatare region of Rwanda.
Courtesy of Google maps



This project presented an exciting opportunity to utilise next generation sequencing and bioinformatics analysis to investigate the genome of a novel emerging pathogen at an early stage of its outbreak timeline. It was hoped that this analysis would afford the opportunity to fully characterise the genome of a newly emerging pathogenic isolate of *Xanthomonas*, contributing to the understanding of the genomic variability of the genus and gaining an insight into the evolution and genomics of pathogenicity of *Xanthomonas*. Further to this, the characterisation and comparison of the Nyagatare species' genomic features with those other *Xanthomonas* species would provide valuable information to track the spread and assess the impact of this emerging threat to east African bean agriculture.

The use of modern sequencing technologies and bioinformatics analysis to inform the investigation into new and emerging pathogens has been carried out to great success in other fields, notably the *E. coli* outbreak in European countries in 2011[1] and the Ebola outbreak in western Africa in 2015 [2]. These studies were large scale and had, quite understandably, a huge amount of international support, cloud sourced manpower and computational weight to aid the analysis effort. It was hoped that this project could show that similar analysis could be performed on smaller outbreaks with less resources available, but still provide in depth analysis which could be of both scientific and practical use to track the spread and successfully inform containment strategies for an emerging pathogen.

It was hoped that this study detailed in this study would provide a valuable foundation for future work on this and other closely related *Xanthomonas* species. It would be very useful to carry out pathogenicity assays to confirm that this strain is indeed responsible for the disease outbreak on bean crops in Rwanda and fully assess its host range. This information along with the genome information would give a better idea of the threat posed by this emerging pathogen. It would be interesting to profile the secreted effectors of this strain, both experimentally and bioinformatically in order to obtain a complete picture of the effector complement and to compare this with other strains including those of the recently published Jacobs *et al*. study. This would help to elucidate the recent evolutionary history of this group

The aim of the work presented here was to use NGS technologies to characterise the genomics of a newly emerging bacterial pathogen of beans isolated during a recent outbreak in Rwanda. It was hoped to sequence, assemble and analyse this newly emerging pathogen. This would allow the classification and comparison of the Nyagatere strain with known xanthomonds, identifying virulence factors and genomic features which have facilitated the adaptation to this new ecological niche and identify potential molecular markers which could be used to track its spread and identify future outbreaks.

Author contribution

The author conducted all bioinformatic analysis for this project. This included using bespoke scripts and pipeline code to conduct the initial quality control of raw sequencing reads and the denovo assembly and analysis of this emerging strain. The author also conducted all analysis and bioinformatic comparisons of this strain with sequence databases.

The author also contributed significantly to the pre-project research, concept design and planning for the project along with the writing, editing and submission of manuscript and the production and editing of all figures and tables.

Manuscript

OXFORD
FEMS JOURNALS investing in science

FEMS Microbiology Letters, 362, 2015, fnu055

doi: 10.1093/femsle/fnu055
Advance Access Publication Date: 11 December 2014
Genome Announcement

GENOME ANNOUNCEMENT – Pathogens & Pathogenicity

The draft genome sequence of *Xanthomonas* species strain Nyagatare, isolated from diseased bean in Rwanda

Valente Aritua[1], Augustine Musoni[2], Alice Kabeja[2], Louis Butare[2], Floride Mukamuhirwa[2], Daphrose Gahakwa[2], Fred Kato[1], Mathew M. Abang[3], Robin Buruchara[4], Melanie Sapp[5], James Harrison[6], David J. Studholme[6,*] and Julian Smith[5]

[1]International Center for Tropical Agriculture, P.O. Box 6247, Kampala, Uganda, [2]Rwanda Agriculture Board, P.O. Box 5016, Kigali, Rwanda, [3]FAO Sub-regional Office for Eastern Africa, P.O. Box 5536, Addis Ababa, Ethiopia, [4]International Centre for Tropical Agriculture (CIAT) P.O. Box 823-00621, Nairobi, Kenya, [5]International Development, The Food and Environment Research Agency, Sand Hutton, York, YO41 1LZ, UK and [6]Biosciences, University of Exeter, Exeter EX4 4QD, Devon, UK

*Corresponding author: Biosciences, University of Exeter, Exeter EX4 4QD, Devon, UK. Tel: +44-(0)-1392-724678; Fax: +44-(0)-1392 263434; E-mail: d.j.studholme@exeter.ac.uk

One sentence summary: We present the genome sequence of the Nyagatare strain, a bacterial pathogen on beans that may be responsible for a mysterious disease emerging in Rwanda.

Editor: Skorn Mongkolsuk

ABSTRACT

We announce the genome sequence for *Xanthomonas* species strain Nyagatare, isolated from beans showing unusual disease symptoms in Rwanda. This strain represents the first sequenced genome belonging to an as-yet undescribed *Xanthomonas* species known as species-level clade 1. It has at least 100 kb of genomic sequence that shows little or no sequence similarity to other xanthomonads, including a unique lipopolysaccharide synthesis gene cluster. At least one genomic region appears to have been acquired from relatives of *Agrobacterium* or *Rhizobium* species. The genome encodes homologues of only three known type-three secretion system effectors: AvrBs2, XopF1 and AvrXv4. Availability of the genome sequence will facilitate development of molecular tools for detection and diagnostics for this newly discovered pathogen of beans and facilitate epidemiological investigations of a potential causal link between this pathogen and the disease outbreak.

Key words: common beans; bacterial canker; *Xanthomonas*; Rwanda

Common bean (*Phaseolus vulgaris*) is an important subsistence and cash crop for smallholder farmers in Rwanda, providing a major source of protein and micronutrients such as iron and zinc (Larochelle and Alwang 2014). In November 2013, farmers in Nyagatare District reported unusual disease on variety ISAR SCB 101 (RWR 2245). Leaf symptoms included curling of upper leaves, wilting, drying and dropping off. There were also brownish and white spots on affected leaves as well as brownish

Received: 21 October 2014; **Accepted:** 4 December 2014

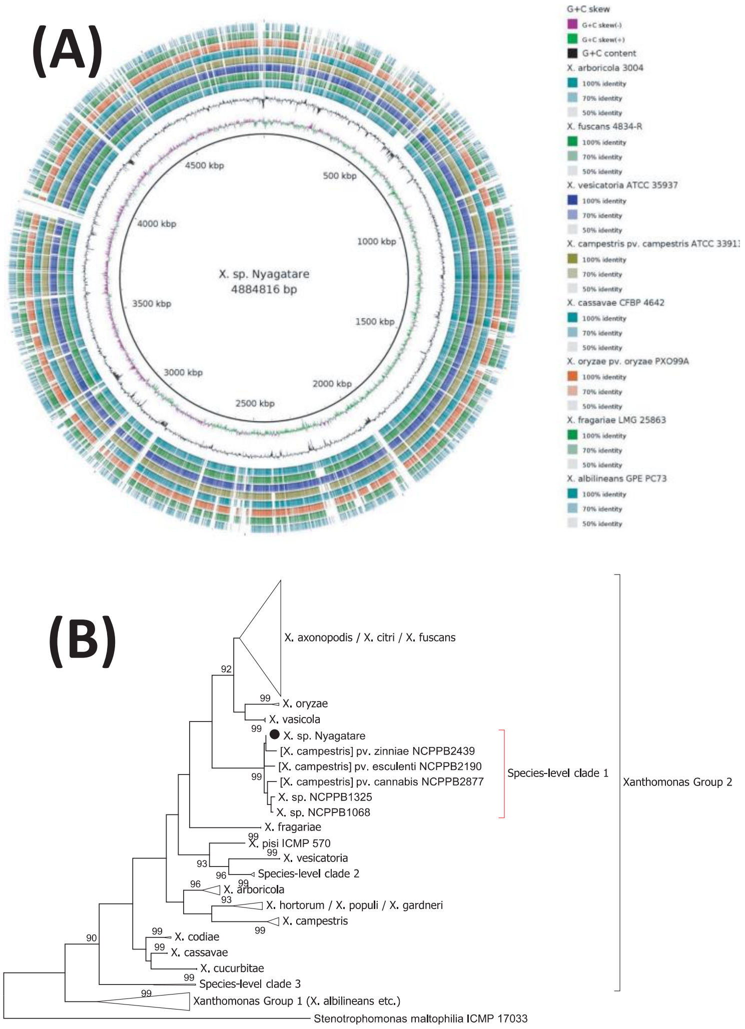


Figure 1. The genome sequence of *Xanthomonas* sp. Nyagatare. Panel **(A)** shows a global comparison of the Nyagatare genome sequence against representative previously sequenced *Xanthomonas* genomes. The genome sequences (Pieretti *et al.*, 2009; Song and Yang 2010; Potnis *et al.*, 2011; Bolot *et al.*, 2013; Darrasse *et al.*, 2013; Vandroemme *et al.*, 2013) were aligned against the Nyagatare genome assembly using BLASTN with an *E*-value threshold of 1×10^{-6}. The Nyagatare assembly had first been re-ordered against the *X. axonopodis* pv. *citri* 306 (da Silva *et al.*, 2002) reference sequence using the contig re-ordering function in Mauve (Rissman *et al.*, 2009). The alignments are visualized using BLAST Ring Image Generator (BRIG) (Alikhan *et al.*, 2011). Panel **(B)** shows the phylogenetic position of the Nyagatare strain based on comparison to previously sequenced *gyrB* genes (Parkinson *et al.*, 2009). Evolutionary history was inferred by using the maximum likelihood method based on the Tamura-Nei model (Tamura and Nei 1993). The tree with the highest log likelihood (−8634.7961) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained by applying the neighbor-joining method to a matrix of pairwise distances estimated using the maximum composite likelihood (MCL) approach. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 438 nucleotide sequences. All positions with less than 95% site coverage were eliminated. That is, fewer than 5% alignment gaps, missing data and ambiguous bases were allowed at any position. There were a total of 524 positions in the final dataset. Evolutionary analyses were conducted in MEGA6 (Tamura *et al.*, 2013). *Xanthomonas* group 1 and 2 as defined by Young and colleagues (Young *et al.*, 2008) are indicated by square brackets as is also species-level clade 1 as defined by Parkinson and colleagues (Parkinson *et al.*, 2009).

to dark necrosis on veins and margins. The stems and branches developed extensive white scabs, which later developed into grey gall-like structures. Green to dark-brown-black streaks and wounds that developed into cankers and necrotic tissues also developed on the stems. The pods developed grey scabs and spots coalescing into large swellings, similar to those on stems. Many of the pods were water soaked, aborted or poorly filled. On dissection, stem vascular tissues were untainted, suggesting that the pathogen is intercellular. A survey by the Rwanda Agriculture Board in November 2013 found that 6 of the 14 sectors of the Nyagatare District were affected. Although the implications were serious for farmers concerned, the overall situation was not yet alarming with no more than 15 ha being affected, but there is concern about possible future spread.

Bacteria were isolated from diseased plant material on YDC (yeast extract dextrose carbonate) medium at CIAT Pathology Laboratory, Uganda. Pathogenicity was demonstrated by inoculation of the isolated strain onto CAL96 beans under glasshouse conditions; symptoms are shown in the Supporting Information. Genomic DNA was sequenced to approximately 58-fold coverage using the Illumina MiSeq with Nextera XT Library Preparation, generating 663 444 pairs of 300-bp reads and assembled into 91 scaffolds with a total length of 4 885 384 bp and an N_{50} length of 101 745 bp using Velvet 1.2.10 (Zerbino and Birney 2008) followed by gap-filling using GapCloser version 1.12-r6 (Luo *et al.*, 2012). Data are available at GenBank under accession numbers GCA_000764855.1 and JRQI00000000.1.

To investigate the core and variable portions of the genome, we used *dnadiff* from the Mummer package (Delcher *et al.*, 2002) to perform pairwise sequence comparisons between the Nyagatare strain genome and all previously sequenced *Xanthomonas* genomes [results are tabulated in Fig. S1 (Supporting Information)]. The highest degree of shared accessory genome was with *X. arboricola* 3004 (73.73% of genome shared with Nyagatare). Fig. 1A also provides an overview of genomic conservation and variation. The genome with greatest sequence similarity was *X. cassavae* (Bolot *et al.*, 2013) with 89.16% nucleotide sequence identity. Average nucleotide identity (ANI) values, as calculated by JSpecies (Richter and Rosselló-Móra 2009), between members of a single species usually exceed 95%. The ANI values between Nyagatare and *X. cassavae* were 87.38% (ANIb) and 89.12% (ANIm). Between Nyagatare and *X. arboricola* 3004, ANIb was 85.54% and ANIm was 88.84%. Between Nagatare and *X. fuscans*, the respective values for ANIb and ANIm were 85.82 and 88.66%. Thus, strain Nyagatare does not belong to any of the previously sequenced species and is phylogenetically distinct from previously studied pathogens of common bean (that fall within the species *X. axonopodis* and *X. fuscans*). The lack of sequenced genomes with very high sequence similarity to strain Nyagatare precluded high-resolution phylogenomic analysis (Rodriguez-R *et al.*, 2012); however, the availability of an extensive database of sequences for the phylogenetic marker gene *gyrB* (Parkinson *et al.*, 2009) allowed us to more precisely examine its phylogenetic position. As illustrated in Fig. 1B, the Nyagatare strain falls within Parkinson's species-level clade 1 (Parkinson *et al.*, 2009), along with little-studied pathogens of *Zinnia elegans*, *Hibiscus esculentus*, *Cannabis sativa*, *Helianthus annuus* and *Nicotiana tabacum* (NCPPB strains 2439, 2190, 2877, 1325 and 1068).

Commensurate with its phylogenetic distinctness from previously sequenced *Xanthomonas* species, the Nyagatare strain has at least 100 kb of genomic sequence that shows little or no sequence similarity to other xanthomonads, as judged by BLASTN searches. This includes a 16.5-kb region located between *metB* and *etfA* (JRQI01000003.1 positions 48 238–64 812) harboring genes for lipopolysaccharide (LPS) synthesis that are quite distinct from any previously sequenced LPS synthesis gene cluster (Patil and Sonti 2004). Another example is a 2.3-kb region (JRQI01000032.1 positions 37 278–34 915) that shares 84% nucleotide sequence identity with the large chromosome of *Agrobacterium radiobacter* K84 (GenBank: CP000628.1), and similar levels of identity with several *Rhizobium* species, but shares no detectable sequence similarity with any available *Xanthomonas* sequences in the NCBI databases.

Virulence factors described in previously sequenced *Xanthomonas* genomes include effector proteins that are substrates of the type-III secretion system (T3SS) (White *et al.*, 2009). The Nyagatare genome encodes an apparently complete T3SS (Fig. S2, Supporting Information). Based on TBLASTN searches between the genome of the Nyagatare strain and Ralf Koebnik's catalogue of known T3SS effectors (http://www.xanthomonas.org/t3e.html), there are homologues of only three: AvrBs2 (73% identity between GenBank: CAJ21683.1 and JRQI01000008.1: 30 926 to 33 058), XopF1 (66% identity between CAJ22045.1 and NC00_3340) and an open reading frame (JRQI01000008.1 positions 38 866 to 39 942) encoding a protein with 87% amino-acid sequence identity to AvrXv4 which has only previously been reported in genomes of *X. euvesicatoria* (Astua-Monge *et al.*, 2000) and *X. perforans* (Potnis *et al.*, 2011).

In conclusion, we present a draft-quality genome sequence for the Nyagatare strain. This is the first genome sequence representing Parkinson's species-level clade 1, and as such its availability will aid the study of this as-yet undescribed candidate new species. Furthermore, this strain may be responsible for the mysterious disease emerging as a potentially serious threat to beans, an important subsistence crop. Availability of the genome sequence will facilitate development of molecular tools for detection and diagnostics thus enabling researchers to test for an epidemiological link between this strain and the disease.

SUPPLEMENTARY DATA

Supplementary data is available at FEMSLE online.

ACKNOWLEDGEMENTS

We thank Richard Thwaites (Fera) for facilitating the DNA sequencing.

FUNDING

Sequencing was made possible through the Canadian International Development Agency (now incorporated into the Department of Foreign Affairs, Trade and Development) support to Pan-Africa Bean Research Alliance. VA was supported on this work by the BBSRC SCPRID Bean Grant BB/J011568/1.

Conflict of interest statement. None declared.

REFERENCES

Alikhan N-F, Petty NK, Ben Zakour NL, *et al.* BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 2011;**12**:402.

Astua-Monge G, Minsavage G V, Stall RE, *et al.* Resistance of tomato and pepper to T3 strains of *Xanthomonas campestris* pv. *vesicatoria* is specified by a plant-inducible avirulence gene. *Mol Plant Microbe In* 2000;**13**:911–21.

Bolot S, Munoz Bodnar A, Cunnac S, *et al*. Draft genome sequence of the *Xanthomonas cassavae* type strain CFBP 4642. *Genome Announc* 2013;**1**: e00679–13.

Da Silva ACR, Ferro JA, Reinach FC, *et al*. Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. *Nature* 2002;**417**:459–63.

Darrasse A, Carrère S, Barbe V, *et al*. Genome sequence of *Xanthomonas fuscans* subsp. *fuscans* strain 4834-R reveals that flagellar motility is not a general feature of xanthomonads. *BMC Genomics* 2013;**14**:761.

Delcher AL, Phillippy A, Carlton J, *et al*. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* 2002;**30**:2478–83.

Larochelle C, Alwang J. Impacts of improved bean varieties on food security in Rwanda. *In: AAEA Annual Meeting*, Minneapolis, MN, 2014.

Luo R, Liu B, Xie Y, *et al*. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 2012;**1**:18.

Parkinson N, Cowie C, Heeney J, *et al*. Phylogenetic structure of *Xanthomonas* determined by comparison of gyrB sequences. *Int J Syst Evol Micr* 2009;**59**:264–74.

Patil P, Sonti R. Variation suggestive of horizontal gene transfer at a lipopolysaccharide (lps) biosynthetic locus in *Xanthomonas oryzae* pv. *oryzae*, the bacterial leaf blight pathogen of rice. *BMC Microbiol* 2004; **4**: 40.

Pieretti I, Royer M, Barbe V, *et al*. The complete genome sequence of *Xanthomonas albilineans* provides new insights into the reductive genome evolution of the xylem-limited Xanthomonadaceae. *BMC Genomics* 2009;**10**:616.

Potnis N, Krasileva K, Chow V, *et al*. Comparative genomics reveals diversity among xanthomonads infecting tomato and pepper. *BMC Genomics* 2011;**12**:146.

Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. *P Natl Acad Sci USA* 2009;**106**:19126–31.

Rissman AI, Mau B, Biehl BS, *et al*. Reordering contigs of draft genomes using the Mauve aligner. *Bioinformatics* 2009;**25**:2071–3.

Rodriguez-R LM, Grajales A, Arrieta-Ortiz M, *et al*. Genomes-based phylogeny of the genus *Xanthomonas*. *BMC Microbiol* 2012;**12**:43.

Song C, Yang B. Mutagenesis of 18 type III effectors reveals virulence function of XopZ(PXO99) in *Xanthomonas oryzae* pv. *oryzae*. *Mol Plant Microbe In* 2010;**23**:893–902.

Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* 1993;**10**:512–26.

Tamura K, Stecher G, Peterson D, *et al*. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* 2013;**30**:2725–9.

Vandroemme J, Cottyn B, Baeyen S, *et al*. Draft genome sequence of *Xanthomonas fragariae* reveals reductive evolution and distinct virulence-related gene content. *BMC Genomics* 2013;**14**:829.

White FF, Potnis N, Jones JB, *et al*. The type III effectors of *Xanthomonas*. *Mol Plant Pathol* 2009;**10**:749–66.

Young JM, Park D-C, Shearman HM, *et al*., A multilocus sequence analysis of the genus *Xanthomonas*. *Syst Appl Microbiol* 2008;**31**:366–77.

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

Supplementary information

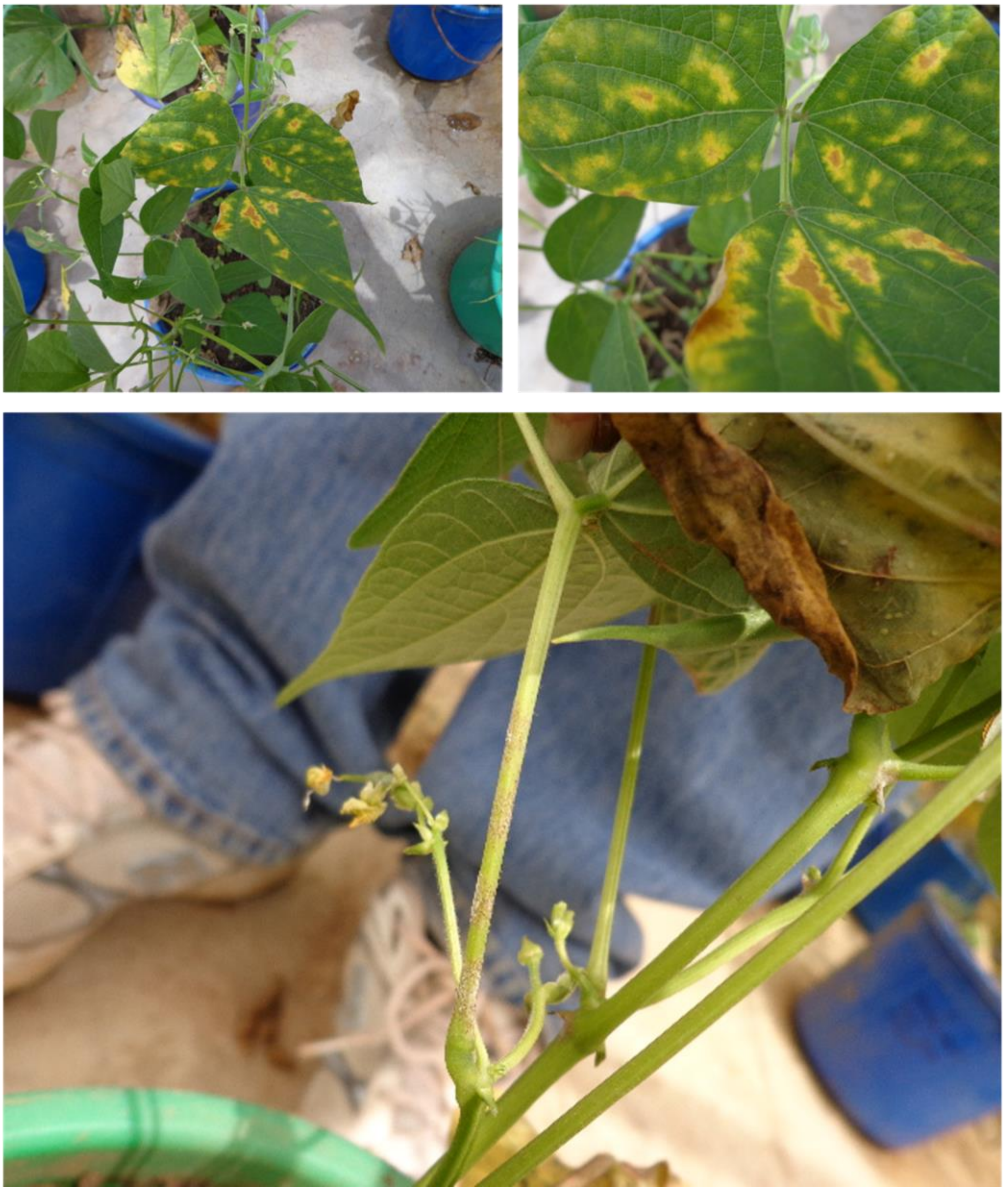


Figure S1: Disease symptoms following inoculation of CAL96 in glass house with *Xanthomonas* sp. strain Nyagatare

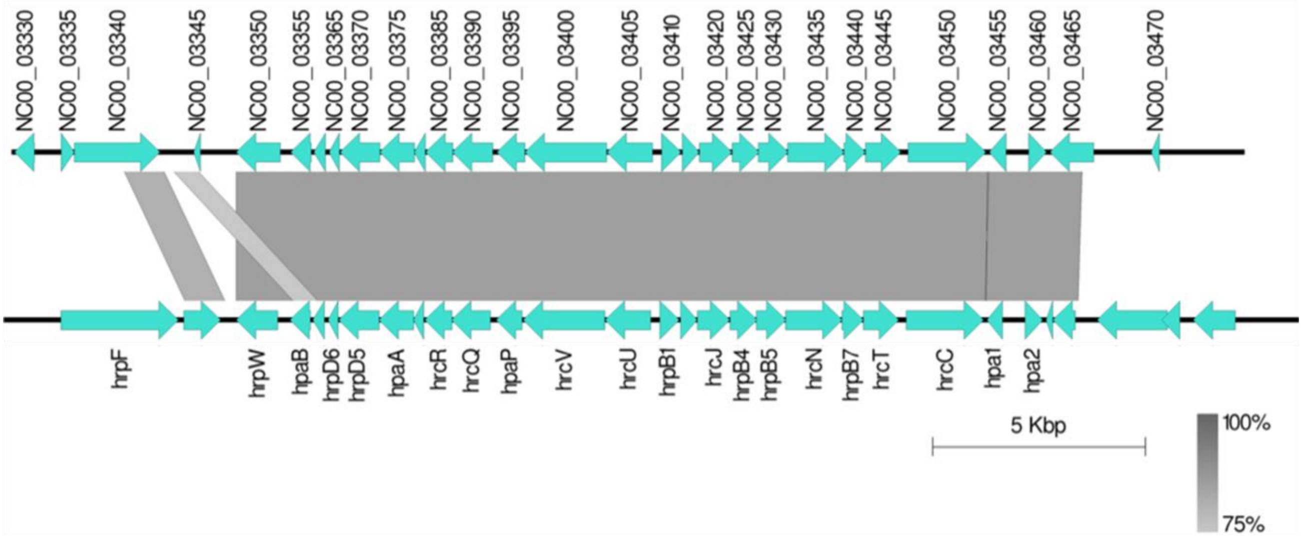


Figure S2: The genome sequence of *Xanthomonas* sp. strain Nyagatare encodes a Hrp type-three secretion system. The Nyagatare sequence (GenBank: KN265463.1) was aligned against the chromsome of *X. campestris* pv. *campestris* ATCC 33913 (GenBank: AE008922.1) using BLASTN and visualised using EasyFig [4].Nucleotide sequence identity of BLASTN hits is indicated by the grey scale gradient.

Bases aligned (Nyagatare)	% bases aligned (Nyagatare)	Reference genome sequence	Bases aligned (reference genome)	% bases aligned (reference genome)	% sequence identity
3513693	73.73	Xanthomonas_arboricola.GCF_000585435	3518320	72.08	88.89
3413084	71.13	Xanthomonas_axonopodis.GCF_000265805	3427722	70.22	88.71
3412068	71.11	Xanthomonas_axonopodis.GCF_000266285	3426141	70.19	88.71
3424828	71.06	Xanthomonas_axonopodis.GCF_000265925	3432830	70.33	88.72
3421358	70.59	Xanthomonas_axonopodis.GCF_000265725	3435639	70.38	88.71
3515768	70.57	Xanthomonas_fuscans_4834_R_uid222814.NC_022541	3463681	70.96	88.7
3401045	70.56	Xanthomonas_axonopodis.GCF_000266685	3415339	69.97	88.73
3422539	70.54	Xanthomonas_axonopodis.GCF_000266105	3424566	70.16	88.7
3419195	70.53	Xanthomonas_axonopodis.GCF_000266225	3435147	70.37	88.72
3415641	70.53	Xanthomonas_axonopodis.GCF_000266385	3425198	70.17	88.7
3426030	70.4	Xanthomonas_axonopodis.GCF_000266365	3437267	70.42	88.71
3381278	70.39	Xanthomonas_axonopodis.GCF_000265745	3388497	69.42	88.71
3400692	70.37	Xanthomonas_axonopodis.GCF_000266265	3413510	69.93	88.73
3427512	70.35	Xanthomonas_axonopodis.GCF_000266805	3435715	70.39	88.7
3427534	70.34	Xanthomonas_axonopodis.GCF_000265565	3439453	70.46	88.7
3428191	70.33	Xanthomonas_axonopodis.GCF_000265945	3435809	70.39	88.71
3414094	70.3	Xanthomonas_axonopodis.GCF_000265885	3427253	70.21	88.73
3422769	70.26	Xanthomonas_axonopodis.GCF_000266625	3435016	70.37	88.7
3411634	70.25	Xanthomonas_axonopodis.GCF_000266005	3423575	70.14	88.73
3440946	70.24	Xanthomonas_axonopodis.GCF_000265825	3454113	70.76	88.72
3258331	70.07	Xanthomonas_axonopodis.GCF_000266305	3275729	67.11	88.89
3415108	70.06	Xanthomonas_axonopodis.GCF_000266345	3427269	70.21	88.73
3430754	70.04	Xanthomonas_axonopodis.GCF_000266425	3441086	70.5	88.71
3430397	70.04	Xanthomonas_axonopodis.GCF_000266405	3438890	70.45	88.71
3422737	69.93	Xanthomonas_axonopodis.GCF_000266645	3435683	70.38	88.71
3380834	69.9	Xanthomonas_axonopodis.GCF_000265645	3394337	69.54	88.77
3420773	69.88	Xanthomonas_axonopodis.GCF_000266585	3429706	70.26	88.7
3373039	69.83	Xanthomonas_axonopodis.GCF_000266025	3387930	69.41	88.75
3416913	69.82	Xanthomonas_axonopodis.GCF_000266725	3427408	70.22	88.71
3422497	69.8	Xanthomonas_axonopodis.GCF_000266505	3432871	70.33	88.71
3464221	69.74	Xanthomonas_alfalfae.GCF_000225915	3453257	70.74	88.75

3464221	69.74	Xanthomonas_axonopodis_citrumelo_F1_uid73179.NC_016010	3453257	70.74	88.75
3425464	69.74	Xanthomonas_axonopodis.GCF_000266545	3433948	70.35	88.71
3423581	69.73	Xanthomonas_axonopodis.GCF_000266485	3433517	70.34	88.7
3427226	69.72	Xanthomonas_axonopodis.GCF_000266845	3436968	70.41	88.71
3374248	69.68	Xanthomonas_axonopodis.GCF_000265625	3389165	69.43	88.75
3423039	69.67	Xanthomonas_axonopodis.GCF_000266745	3432245	70.31	88.72
3364535	69.64	Xanthomonas_axonopodis.GCF_000266185	3379192	69.23	88.79
3401387	69.61	Xanthomonas_axonopodis.GCF_000266245	3414113	69.94	88.73
3494600	69.59	Xanthomonas_arboricola.GCF_000306055	3497748	71.66	88.85
3412101	69.59	Xanthomonas_axonopodis.GCF_000266765	3425490	70.18	88.72
3424606	69.59	Xanthomonas_axonopodis.GCF_000265785	3437171	70.42	88.71
3411483	69.57	Xanthomonas_axonopodis.GCF_000265685	3423933	70.14	88.73
3425566	69.57	Xanthomonas_axonopodis.GCF_000266465	3434035	70.35	88.7
3410500	69.56	Xanthomonas_axonopodis.GCF_000266085	3423524	70.14	88.73
3290043	69.55	Xanthomonas_axonopodis.GCF_000266785	3303152	67.67	88.86
3360530	69.53	Xanthomonas_axonopodis.GCF_000265965	3374661	69.13	88.78
3423594	69.53	Xanthomonas_axonopodis.GCF_000266565	3434402	70.36	88.7
3411148	69.52	Xanthomonas_axonopodis.GCF_000266165	3424010	70.15	88.73
3411684	69.47	Xanthomonas_axonopodis.GCF_000265845	3426428	70.2	88.71
3279812	69.45	Xanthomonas_axonopodis.GCF_000266525	3292370	67.45	88.78
3438687	69.43	Xanthomonas_axonopodis.GCF_000265985	3448036	70.64	88.7
3422447	69.39	Xanthomonas_axonopodis.GCF_000266325	3434269	70.36	88.7
3648742	69.34	Xanthomonas_perforans.GCF_000192045	3435002	70.37	88.8
3433330	69.33	Xanthomonas_axonopodis.GCF_000265585	3444908	70.57	88.72
3309492	69.32	Xanthomonas_axonopodis.GCF_000265905	3323519	68.09	88.85
3413380	69.23	Xanthomonas_axonopodis.GCF_000266145	3426468	70.2	88.73
3442937	69.22	Xanthomonas_axonopodis.GCF_000309905	3454476	70.77	88.65
3421346	69.17	Xanthomonas_axonopodis.GCF_000265765	3432924	70.33	88.7
3174245	69.11	Xanthomonas_axonopodis.GCF_000265665	3189384	65.34	88.95
3349134	69.01	Xanthomonas_axonopodis.GCF_000266125	3362790	68.89	88.81
3430842	69	Xanthomonas_axonopodis.GCF_000266205	3439334	70.46	88.7
3415913	68.94	Xanthomonas_axonopodis.GCF_000266665	3428629	70.24	88.71
3362668	68.91	Xanthomonas_fuscans.GCF_000175135	3364440	68.93	88.72
3285573	68.83	Xanthomonas_axonopodis.GCF_000265865	3300619	67.62	88.84
3440136	68.58	Xanthomonas_axonopodis.GCF_000266705	3448387	70.65	88.7
3392380	68.58	Xanthomonas_axonopodis.GCF_000285775	3398927	69.63	88.67
3217502	68.56	Xanthomonas_campestris.GCF_000277875	3227335	66.12	88.47
3440289	68.53	Xanthomonas_axonopodis.GCF_000266045	3449271	70.66	88.7
3458796	68.46	Xanthomonas_hortorum.GCF_000505565	3458379	70.85	88.29
3242204	68.43	Xanthomonas_axonopodis.GCF_000265705	3254668	66.68	88.89
3425021	68.36	Xanthomonas_axonopodis.GCF_000266445	3432284	70.32	88.71
3405714	68.3	Xanthomonas_axonopodis.GCF_000266605	3413109	69.92	88.74
3214929	67.99	Xanthomonas_campestris.GCF_000277955	3223155	66.03	88.46
3229866	67.97	Xanthomonas_campestris.GCF_000277915	3232569	66.22	88.42

3500220	67.92	Xanthomonas_axonopodis_Xac29_1_uid193774.NC_020800	3489867	71.49	88.69
3445501	67.83	Xanthomonas_alfalfae.GCF_000488955	3450913	70.7	88.79
3224364	67.76	Xanthomonas_campestris.GCF_000277895	3224951	66.07	88.41
3395149	67.73	Xanthomonas_fuscans.GCF_000175155	3396203	69.58	88.73
3425456	67.63	Xanthomonas_axonopodis.GCF_000266825	3437684	70.43	88.7
3452428	67.46	Xanthomonas_axonopodis.GCF_000309925	3464660	70.98	88.65
3730618	67.45	Xanthomonas_vesicatoria.GCF_000192025	3610523	73.97	88.94
3486665	67.37	Xanthomonas_axonopodis_citri_306_uid57889.NC_003919	3490140	71.5	88.69
3536912	67.33	Xanthomonas_axonopodis.GCF_000495275	3527850	72.27	88.66
3258804	67.29	Xanthomonas_axonopodis.GCF_000266065	3270729	67.01	88.87
3425828	67.02	Xanthomonas_citri.GCF_000263335	3436894	70.41	88.66
3231561	66.96	Xanthomonas_vasicola.GCF_000278035	3235516	66.28	88.45
3498283	66.93	Xanthomonas_arboricola.GCF_000355635	3496237	71.63	88.91
3282758	66.89	Xanthomonas_campestris.GCF_000233635	3215723	65.88	88.49
3426184	66.81	Xanthomonas_axonopodis.GCF_000265605	3434495	70.36	88.71
3199446	66.76	Xanthomonas_campestris.GCF_000277975	3201415	65.59	88.53
3440203	66.43	Xanthomonas_campestris_vesicatoria_85_10_uid58321.NC_007508	3436431	70.4	88.77
3177476	66.28	Xanthomonas_campestris.GCF_000159815	3191469	65.38	88.33
3486665	66.11	Xanthomonas_citri.GCF_000007165	3490140	71.5	88.69
3500220	66.09	Xanthomonas_axonopodis.GCF_000348585	3489867	71.49	88.69
3259839	65.97	Xanthomonas_campestris.GCF_000221965	3235136	66.28	87.78
3259839	65.97	Xanthomonas_campestris_raphani_756C_uid159539.NC_017271	3235136	66.28	87.78
3510122	65.96	Xanthomonas_citri_Aw12879_uid194444.NC_020815	3504578	71.8	88.67
3122181	65.84	Xanthomonas_campestris.GCF_000277935	3123091	63.98	88.22
2856081	65.75	Xanthomonas_oryzae.GCF_000511585	2865786	58.71	88.55
3632228	65.7	Xanthomonas_gardneri.GCF_000192065	3409330	69.84	88.39
3447860	65.41	Xanthomonas_cassavae.GCF_000454545	3464534	70.98	89.16
3444615	65.41	Xanthomonas_axonopodis.GCF_000259445	3446198	70.6	88.75
3510122	65.02	Xanthomonas_citri.GCF_000349225	3504578	71.8	88.67
2779966	64.87	Xanthomonas_oryzae.GCF_000482445	2786277	57.08	88.56
3204799	64.64	Xanthomonas_vasicola.GCF_000278015	3212538	65.81	88.42
3199708	64.62	Xanthomonas_vasicola.GCF_000277995	3206648	65.69	88.41
3094619	64.42	Xanthomonas_vasicola.GCF_000278075	3096324	63.43	88.63
2837982	64.24	Xanthomonas_oryzae.GCF_000507025	2851686	58.42	88.58
3216393	64.03	Xanthomonas_campestris.GCF_000263835	3218504	65.94	87.78
3245426	63.9	Xanthomonas_campestris.GCF_000070605	3223488	66.04	87.77
3245426	63.9	Xanthomonas_campestris_uid61643.NC_010688	3223488	66.04	87.77
3189992	63.78	Xanthomonas_campestris.GCF_000321125	3198870	65.53	87.73
2925863	63.75	Xanthomonas_oryzae.GCF_000212755	2940610	60.24	88.6
3225566	63.54	Xanthomonas_campestris.GCF_000007145	3215592	65.88	87.73
3225566	63.54	Xanthomonas_campestris_ATCC_33913_uid57887.NC_003902	3215592	65.88	87.73
3443225	63.53	Xanthomonas_euvesicatoria.GCF_000009165	3437634	70.42	88.77
2898804	63.09	Xanthomonas_oryzae.GCF_000212775	2911183	59.64	88.66
3224075	62.62	Xanthomonas_campestris.GCF_000012105	3210905	65.78	87.72

3224075	62.62	Xanthomonas_campestris_8004_uid57595.NC_007086	3210905	65.78	87.72
2928763	60.62	Xanthomonas_oryzae.GCF_000168315	2876862	58.94	88.53
2928763	60.62	Xanthomonas_oryzae_oryzicola_BLS256_uid54411.NC_017267	2876862	58.94	88.53
2994263	60.48	Xanthomonas_vasicola.GCF_000278055	2992850	61.31	88.81
2519966	60.25	Xanthomonas_fragariae.GCF_000376745	2480723	50.82	87.58
3217719	58.74	Xanthomonas_vasicola.GCF_000159795	3224905	66.07	88.41
2872100	58.14	Xanthomonas_oryzae_MAFF_311018_uid58547.NC_007705	2791472	57.19	88.55
2872100	58.14	Xanthomonas_oryzae.GCF_000010025	2791472	57.19	88.55
2866762	58.01	Xanthomonas_oryzae_KACC_10331_uid58155.NC_006834	2790542	57.17	88.55
2866762	58.01	Xanthomonas_oryzae.GCF_000007385	2790542	57.17	88.55
2979829	56.87	Xanthomonas_oryzae_PXO99A_uid59131.NC_010717	2793070	57.22	88.55
2979829	56.87	Xanthomonas_oryzae.GCF_000019585	2793070	57.22	88.55
2309083	46.5	Xanthomonas_sp._M97.GCF_000401255	2313467	47.39	86.24
1437893	35.06	Xanthomonas_translucens.GCF_000313775	1444452	29.59	85.42
1484597	33.23	Xanthomonas_translucens.GCF_000334075	1481033	30.34	85.27
1472756	32.99	Xanthomonas_translucens.GCF_000331775	1470877	30.13	85.4
1515567	32.21	Xanthomonas_sp._SHU166.GCF_000364685	1513699	31.01	85.35
1571991	32.09	Xanthomonas_sacchari.GCF_000225975	1566458	32.09	85.25
1185367	32.02	Xanthomonas_sp._NCPPB1131.GCF_000226895	1185379	24.28	86.08
1545738	31.96	Xanthomonas_sp._SHU199.GCF_000364665	1540783	31.57	85.31
1551429	31.95	Xanthomonas_sp._SHU308.GCF_000364645	1549168	31.74	85.3
1466854	31.55	Xanthomonas_sp._NCPPB1132.GCF_000226915	1462817	29.97	85.54
722661	21.14	Pseudoxanthomonas_suwonensis_11_1_uid62105.NC_014924	716302	14.67	84.41
686299	19.88	Pseudoxanthomonas_spadix_BD_a59_uid75113.NC_016147	685905	14.05	84.42
714597	18.96	Xanthomonas_albilineans_GPE_PC73_uid43163.NC_013722	710312	14.55	84.63
714597	18.55	Xanthomonas_albilineans.GCF_000087965	710312	14.55	84.63

Supplementary table 1: Summary of results of dnadiff comparisons between the Nyagatare genome assembly versus previously sequenced xanthomonad genomes.

References

1. Cheung, M. K., Li, L., Nong, W. & Kwan, H. S. 2011 German *Escherichia coli* O104:H4 outbreak: whole-genome phylogeny without alignment. *BMC Res. Notes* **4**, 533 (2011).
2. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–32 (2016).
3. Jacobs, J. M., Pesce, C., Lefeuvre, P. & Koebnik, R. Comparative genomics of a cannabis pathogen reveals insight into the evolution of pathogenicity in *Xanthomonas*. **6**, 1–13 (2015).
4. Sullivan, M. J., Petty, N. K. & Beatson, S. A. Easyfig: a genome comparison visualizer. *Bioinformatics* **27**, 1009–1010 (2011).

Chapter 6:

Assessing the performance of the Oxford Nanopore Technologies MinION

Work from this chapter was published in:

Laver, T., Harrison, J., O'Neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., & Studholme, D. J. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular detection and quantification*, *3*, 1–8.

This paper was cited by:

1. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics, Proteomics Bioinforma.* **13**, 278–289 (2015).
2. Wattam, A. R. *et al.* Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Res.* **45**, D535–D542 (2017).
3. Sović, I. *et al.* Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nat. Commun.* **7**, (2016).
4. Ye, C., Hill, C. M., Wu, S., Ruan, J. & Ma, Z. DBG2OLC: Efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Sci. Rep.* **6**, 1–9 (2016).
5. Marschall, T. *et al.* Computational pan-genomics: Status, promises and challenges. *Brief. Bioinform.* **19**, 118–135 (2018).
6. Bleidorn, C. Third generation sequencing: Technology and its potential impact on evolutionary biodiversity research. *Syst. Biodivers.* **14**, 1–8 (2016).
7. Gomez-Escribano, J. P., Alt, S. & Bibb, M. J. Next generation sequencing of actinobacteria for the discovery of novel natural products. *Mar. Drugs* **14**, 6–8 (2016).
8. Salmela, L., Walve, R., Rivals, E., Ukkonen, E. & Sahinalp, C. Accurate self-correction of

errors in long reads using de Bruijn graphs. *Bioinformatics* **33**, 799–806 (2017).

9. Ambardar, S., Gupta, R., Trakroo, D., Lal, R. & Vakhlu, J. High Throughput Sequencing: An Overview of Sequencing Chemistry. *Indian J. Microbiol.* **56**, 394–404 (2016).

10. Cretu Stancu, M. *et al.* Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat. Commun.* **8**, 1–13 (2017).

11. Rang, F. J., Kloosterman, W. P. & de Ridder, J. From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* **19**, 1–11 (2018).

12. Sohn, J. Il & Nam, J. W. The present and future of de novo whole-genome assembly. *Brief. Bioinform.* **19**, 23–40 (2018).

13. Sunagar, K., Morgenstern, D., Reitzel, A. M. & Moran, Y. Ecological venomics: How genomics, transcriptomics and proteomics can shed new light on the ecology and evolution of venom. *J. Proteomics* **135**, 62–72 (2016).

14. Kennedy, E., Dong, Z., Tennant, C. & Timp, G. Reading the primary structure of a protein with 0.07 nm 3 resolution using a subnanometre-diameter pore. *Nat. Nanotechnol.* **11**, 968–976 (2016).

15. Dodsworth, S. Genome skimming for next-generation biodiversity analysis. *Trends Plant Sci.* **20**, 525–527 (2015).

16. Weirather, J. L. *et al.* Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* **6**, 100 (2017).

17. Escalona, M., Rocha, S. & Posada, D. A comparison of tools for the simulation of genomic next-generation sequencing data. *Nat. Rev. Genet.* **17**, 459–469 (2016).

18. Camilla, L.C. *et al.* MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Research* **4**, 1075 (2015).

19. Hugerth, L. W. & Andersson, A. F. Analysing microbial community composition through amplicon sequencing: From sampling to hypothesis testing. *Front. Microbiol.* **8**, 1–22 (2017).

20. Frenken, T. *et al.* Integrating chytrid fungal parasites into plankton ecology: research gaps and needs. *Environ. Microbiol.* **19**, 3802–3822 (2017).

21. Lee, R. S. & Behr, M. A. The implications of whole-genome sequencing in the control of tuberculosis. *Ther. Adv. Infect. Dis.* **3**, 47–62 (2015).

22. Paridah, M. . *et al.* We are IntechOpen , the world ' s leading publisher of Open Access books Built by scientists , for scientists TOP 1 %. *Intech* **i**, 13 (2016).

23. Kchouk, M., Gibrat, J. F. & Elloumi, M. Generations of Sequencing Technologies: From First to Next Generation. *Biol. Med.* **09**, (2017).

24. Stoiber, M. *et al.* De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. *bioRxiv* 094672 (2016).

25. Friedrich, S. M., Zec, H. C. & Wang, T. H. Analysis of single nucleic acid molecules in micro- and nano-fluidics. *Lab Chip* **16**, 790–811 (2016).

26. Owens, M. M. *et al.* Pitfalls of haplotype phasing from amplicon-based long-read sequencing. *Sci. Rep.* **6**, 1–6 (2016).

27. Liu, Q., Zhang, P., Wang, D., Gu, W. & Wang, K. Interrogating the 'unsequenceable' genomic trinucleotide repeat disorders by long-read sequencing. *Genome Med.* **9**, 65 (2017).

28. Magi, A., Semeraro, R., Mingrino, A., Giusti, B. & D'Aurizio, R. Nanopore sequencing data analysis: State of the art, applications and challenges. *Brief. Bioinform.* **19**, 1256–1272 (2017).

29. Mitsuhashi, S. *et al.* A portable system for rapid bacterial composition analysis using a nanopore-based sequencer and laptop computer. *Sci. Rep.* **7**, 5657 (2017).

30. Mak, S. S. T. *et al.* Comparative performance of the BGISEQ-500 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. *Gigascience* **6**, 1–13 (2017).

31. Parker, J., Helmstetter, A. J., Devey, Di., Wilkinson, T. & Papadopulos, A. S. T. Field-based species identification of closely-related plants using real-time nanopore sequencing. *Sci. Rep.* **7**, 8345 (2017).

32. Chu, J., Mohamadi, H., Warren, R. L., Yang, C. & Birol, I. Innovations and challenges in detecting long read overlaps: An evaluation of the state-of-the-art. *Bioinformatics* **33**, 1261–1270 (2017).

33. Tomaszkiewicz, M., Medvedev, P. & Makova, K. D. Y and W Chromosome Assemblies: Approaches and Discoveries. *Trends Genet.* **33**, 266–282 (2017).

34. Sović, I., Križanović, K., Skala, K. & Šikić, M. Evaluation of hybrid and non-hybrid methods for de novo assembly of nanopore reads. *Bioinformatics* **32**, 2582–2589 (2016).

35. Jansson, J. K., White, R. A., Baker, E. S., Callister, S. J. & Moore, R. J. The past, present and future of microbiome analyses. *Nat. Protoc.* **11**, 2049–2053 (2016).

36. Amha, Y. M. *et al.* Inhibition of anaerobic digestion processes: Applications of molecular tools. *Bioresour. Technol.* **247**, 999–1014 (2018).

37. Deschamps, S. *et al.* Characterization, correction and de novo assembly of an Oxford Nanopore genomic dataset from Agrobacterium tumefaciens. *Sci. Rep.* **6**, 28625 (2016).

38. Hossein TabatabaeiYazdi, S. M., Gabrys, R. & Milenkovic, O. Portable and Error-Free DNA-Based Data Storage. *Sci. Rep.* **7**, 5011 (2017).

39. Zhang, D., Bi, H., Liu, B. & Qiao, L. Detection of Pathogenic Microorganisms by Microfluidics Based Analytical Methods. *Anal. Chem.* **90**, 5512–5520 (2018).

40. Križanović, K., Echchiki, A., Roux, J. & Šikić, M. Evaluation of tools for long read RNA-seq splice-aware alignment. *Bioinformatics* **34**, 748–754 (2018).

41. Ye, C. & Ma, Z. (Sam). Sparc: a sparsity-based consensus algorithm for long erroneous

sequencing reads. *PeerJ* **4**, e2016 (2016).

42. Moldován, N. *et al.* Third-generation Sequencing Reveals Extensive Polycistronism and Transcriptional Overlapping in a Baculovirus. *Sci. Rep.* **8**, 8604 (2018).

43. Massaia, A. & Xue, Y. Human Y chromosome copy number variation in the next generation sequencing era and beyond. *Hum. Genet.* **136**, 591–603 (2017).

44. Debladis, E., Llauro, C., Carpentier, M. C., Mirouze, M. & Panaud, O. Detection of active transposable elements in *Arabidopsis thaliana* using Oxford Nanopore Sequencing technology. *BMC Genomics* **18**, 537 (2017).

45. Pomerantz, A. *et al.* Real-time DNA barcoding in a rainforest using nanopore sequencing: Opportunities for rapid biodiversity assessments and local capacity building. *Gigascience* **7**, 1–14 (2018).

46. Lear, G. *et al.* Methods for the extraction, storage, amplification and sequencing of DNA from environmental samples. *N. Z. J. Ecol.* **42**, 10-50A (2018).

47. Carter, J.-M. & Hussain, S. Robust long-read native DNA sequencing using the ONT CsgG Nanopore system. *Wellcome open Res.* **2**, 23 (2017).

48. Alvarenga, D. O., Fiore, M. F. & Varani, A. M. A metagenomic approach to cyanobacterial genomics. *Frontiers in Microbiology* **8**, 809 (2017).

49. Barrett, C. F., Bacon, C. D., Antonelli, A., Cano, Á. & Hofmann, T. An introduction to plant phylogenomics with a focus on palms. *Bot. J. Linn. Soc.* **182**, 234–255 (2016).

50. Wilkinson, M. J. *et al.* Supplementary Information Replacing Sanger with Next Generation Sequencing to improve coverage and quality of reference DNA barcodes for plants Supplementary Tables S1 , S2 , S3 , S4 , S5 , S6. *Sci. Rep.* **7**, 46040 (2017).

51. Moldován, N. *et al.* Multi-platform sequencing approach reveals a novel transcriptome profile in *pseudorabies* virus. *Frontiers in Microbiology* **8**, 2708 (2018).

52. Nguyen, M. *et al.* Developing an in silico minimum inhibitory concentration panel test for

Klebsiella pneumonia. *Sci. Rep.* **8**, 421 (2018).

53. Alikian, M., Gale, R. P., Apperley, J. F., Foroni, L. & Alikian, M. Molecular techniques for the personalised management of patients with chronic myeloid leukaemia. *Biomol. Detect. Quantif.* **11**, 4–20 (2017).

54. Agah, S., Zheng, M., Pasquali, M. & Kolomeisky, A. B. DNA sequencing by nanopores: Advances and challenges. *J. Phys. D. Appl. Phys.* **49**, 413001 (2016).

55. Van den Bergh, B., Swings, T., Fauvart, M. & Michiels, J. Experimental Design, Population Dynamics, and Diversity in Microbial Experimental Evolution. *Microbiol. Mol. Biol. Rev.* **82**, e00008-18 (2018).

56. Valles-Colomer, M. *et al.* Meta-omics in inflammatory bowel disease research: Applications, challenges, and guidelines. *J. Crohn’s Colitis* **10**, 735–746 (2016).

57. Lavezzo, E., Barzon, L., Toppo, S. & Palù, G. Third generation sequencing technologies applied to diagnostic microbiology: benefits and challenges in applications and data analysis. *Expert Rev. Mol. Diagn.* **16**, 1011–1023 (2016).

58. Vasylyeva, T. I., Friedman, S. R., Paraskevis, D. & Magiorkinis, G. Integrating molecular epidemiology and social network analysis to study infectious diseases: Towards a socio-molecular era for public health. *Infect. Genet. Evol.* **46**, 248–255 (2016).

59. Jünemann, S. *et al.* Bioinformatics for NGS-based metagenomics and the application to biogas research. *J. Biotechnol.* **261**, 10–23 (2017).

60. Merker, M., Kohl, T. A., Niemann, S. & Supply, P. The evolution of strain typing in the *mycobacterium tuberculosis* complex. in *Advances in Experimental Medicine and Biology* **1019**, 43–78 (Springer, 2017).

61. D’Agostino, D., Morganti, L., Corni, E., Cesini, D. & Merelli, I. Combining Edge and Cloud computing for low-power, cost-effective metagenomics analysis. *Futur. Gener. Comput. Syst.* **90**, 79–85 (2019).

62. White, R. A. *et al.* The state of rhizospheric science in the era of multi-omics: A practical guide to omics technologies. *Rhizosphere* **3**, 212–221 (2017).

63. Ma, X., Stachler, E. & Bibby, K. Evaluation of Oxford Nanopore MinION Sequencing for 16S rRNA Microbiome Characterization (preprint). *bioRxiv* 099960 (2017).

64. Amin, M. R., Skiena, S. & Schatz, M. C. NanoBLASTer: Fast alignment and characterization of Oxford Nanopore single molecule sequencing reads. in *2016 IEEE 6th International Conference on Computational Advances in Bio and Medical Sciences, ICCABS 2016* 1–6 (IEEE, 2016).

65. Deonovic, B., Wang, Y., Weirather, J., Wang, X. J. & Au, K. F. IDP-ASE: Haplotyping and quantifying allele-specific expression at the gene and gene isoform level by hybrid sequencing. *Nucleic Acids Res.* **45**, e32–e32 (2017).

66. Zhao, C., Liu, F. & Pyle, A. M. An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. *Rna* **24**, 183–185 (2018).

67. Liu, M. & Darling, A. Metagenomic Chromosome Conformation Capture (3C): techniques, applications, and challenges. *F1000Research* **4**, 1377 (2015).

68. Adams, I. & Fox, A. Diagnosis of plant viruses using next-generation sequencing and metagenomic analysis. in *Current Research Topics in Plant Virology* 323–335 (Springer, 2016).

69. Morrison, J., Watts, G., Hobbs, G. & Dawnay, N. Field-based detection of biological samples for forensic analysis: Established techniques, novel tools, and future innovations. *Forensic Sci. Int.* **285**, 147–160 (2018).

70. Brady, K. T. & Reiner, J. E. Improving the prospects of cleavage-based nanopore sequencing engines. *J. Chem. Phys.* **143**, 08B608_1 (2015).

71. Redin, D. *et al.* Droplet Barcode Sequencing for targeted linked-read haplotyping of single DNA molecules. *Nucleic Acids Res.* **45**, e125–e125 (2017).

72. Galata, V. *et al.* BusyBee Web: metagenomic data analysis by bootstrapped supervised binning and annotation. *Nucleic Acids Res.* **45**, W171–W179 (2017).

73. Doyle, L. E. & Marsili, E. Weak electricigens: A new avenue for bioelectrochemical research. *Bioresour. Technol.* **258**, 354–364 (2018).

74. Lindberg, M. R. *et al.* A comparison and integration of MiSeq and MinION platforms for sequencing single source and mixed mitochondrial genomes. *PLoS One* **11**, e0167600 (2016).

75. Austin, C. M. *et al.* De novo genome assembly and annotation of Australia's largest freshwater fish, the Murray cod (*Maccullochella peelii*), from Illumina and Nanopore sequencing read. *Gigascience* **6**, gix063 (2017).

76. Krishnakumar, R. *et al.* Systematic and stochastic influences on the performance of the MinION nanopore sequencer across a range of nucleotide bias. *Sci. Rep.* **8**, 3159 (2018).

77. Fu, S. *et al.* IDP-denovo: De novo transcriptome assembly and isoform annotation by hybrid sequencing. *Bioinformatics* **34**, 2168–2176 (2018).

78. Jagadeesan, B. *et al.* The use of next generation sequencing for improving food safety: Translation into practice. *Food Microbiol.* **79**, 96–115 (2019).

79. Weik, F., Kesselheim, S. & Holm, C. A coarse-grained DNA model for the prediction of current signals in DNA translocation experiments. *J. Chem. Phys.* **145**, 194106 (2016).

80. Di Donato, A., Filippone, E., Ercolano, M. R. & Frusciante, L. Genome Sequencing of Ancient Plant Remains: Findings, Uses and Potential Applications for the Study and Improvement of Modern Crops. *Front. Plant Sci.* **9**, 441 (2018).

81. Szkop, K. J. & Nobeli, I. Untranslated Parts of Genes Interpreted: Making Heads or Tails of High-Throughput Transcriptomic Data via Computational Methods. *BioEssays* **39**, 1700090 (2017).

82. Arnold, C. Considerations in centralizing whole genome sequencing for microbiology in a

public health setting. *Expert Review of Molecular Diagnostics* **16**, 619–621 (2016).

83. Krachunov, M., Nisheva, M. & Vassilev, D. Application of Machine Learning Models in Error and Variant Detection in High-Variation Genomics Datasets. *Computers* **6**, 29 (2017).

84. Cook, D. E. *et al.* Long-Read Annotation: Automated Eukaryotic Genome Annotation Based on Long-Read cDNA Sequencing. *Plant Physiol.* **179**, 38–54 (2019).

85. Ji, W. *et al.* Rapid and Accurate Sequencing of Enterovirus Genomes Using MinION Nanopore Sequencer *. *Biomed Env. Sci* **30**, 718–726 (2017).

86. Helbing, S., Lattorff, H. M. G., Moritz, R. F. A. & Buttstedt, A. Comparative analyses of the major royal jelly protein gene cluster in three *Apis* species with long amplicon sequencing. *DNA Res.* **24**, 279–287 (2017).

87. Siegler, R. A New Conceptual Framework for Sarcoma. *Front. Genet.* **6**, 257–260 (2015).

88. Magner, A., Duda, J., Szpankowski, W. & Grama, A. Fundamental Bounds for Sequence Reconstruction From Nanopore Sequencers. *IEEE Trans. Mol. Biol. Multi-Scale Commun.* **2**, 92–106 (2017).

89. Zhao QiongYi, and Gratten, J., and Restuadi Restuadi & and Li Xuan. Mapping and differential expression analysis from short-read RNA-Seq data in model organisms. *Quant. Biol.* **4**, 22–35 (2016).

90. Hu, H., Scheben, A. & Edwards, D. Advances in Integrating Genomics and Bioinformatics in the Plant Breeding Pipeline. *Agriculture* **8**, 75 (2018).

91. Susilawati, T. N. *et al.* Deep sequencing approach for investigating infectious agents causing fever. *Eur. J. Clin. Microbiol. Infect. Dis.* **35**, 1137–1149 (2016).

92. Gonzalez, C. *et al.* Barcoding analysis of HIV drug resistance mutations using Oxford Nanopore MinION (ONT) sequencing. *bioRxiv* 240077 (2018).

93. Moldován, N. *et al.* Multiplatform next-generation sequencing identifies novel RNA molecules and transcript isoforms of the endogenous retrovirus isolated from cultured cells.

FEMS Microbiol. Lett. **365**, fny013 (2018).

94. van Aerle, R. & Santos, E. M. Advances in the application of high-throughput sequencing in invertebrate virology. *J. Invertebr. Pathol.* **147**, 145–156 (2017).

95. Maggi, E., Patterson, N. E. & Montagna, C. Technological advances in precision medicine and drug development. *Expert Rev. Precis. Med. Drug Dev.* **1**, 331–343 (2016).

96. Fu, S., Wang, A. & Au, K. F. A comparative evaluation of hybrid error correction methods for error-prone long reads. *Genome Biol.* **20**, 26 (2019).

97. Rice, E. S. & Green, R. E. New Approaches for Genome Assembly and Scaffolding. *Annu. Rev. Anim. Biosci.* **7**, 17–40 (2018).

98. Song, E. J., Lee, E. S. & Nam, Y. Do. Progress of analytical tools and techniques for human gut microbiome research. *J. Microbiol.* **56**, 693–705 (2018).

99. Parker, J., Helmstetter, A. J., Devey, D. S. & Papadopulos, A. S. T. Field-based species identification in eukaryotes using single molecule, real-time sequencing. *bioRxiv* 107656 (2017).

100. Van den Berge, K. *et al.* RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis. *Annu. Rev. Biomed. Data Sci.* **2**, (2019).

101. Cali, D. S., Kim, J. S., Ghose, S., Alkan, C. & Mutlu, O. Nanopore Sequencing Technology and Tools for Genome Assembly: Computational Analysis of the Current State, Bottlenecks and Future Directions. *arXiv Prepr. arXiv* **1711**, (2017).

102. Bainomugisa, A. *et al.* A complete high-quality MinION nanopore assembly of an extensively drug-resistant *Mycobacterium tuberculosis* Beijing lineage strain identifies novel variation in repetitive PE/PPE gene regions. *Microb. Genomics* **4**, (2018).

103. Macqueen, D. J. *et al.* Nanopore sequencing for rapid diagnostics of salmonid RNA viruses. *Sci. Rep.* **8**, 1–9 (2018).

104. Milicchio, F. & Prosperi, M. Efficient data structures for mobile de novo genome assembly

by third-generation sequencing. *Procedia Comput. Sci.* **110**, 440–447 (2017).

105. Krachunov, M., Nisheva, M. & Vassilev, D. Machine learning-driven noise separation in high variation genomics sequencing datasets. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **11089 LNAI**, 173–185 (Springer, 2018).

106. Brenner, J. & Putonti, C. HAsh-MaP-ERadicator: Filtering non-target sequences from next generation sequencing reads. in *Proceedings - 2015 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2015* 1100–1101 (IEEE, 2015).

107. Lecluze, E., Jégou, B., Rolland, A. D. & Chalmel, F. New transcriptomic tools to understand testis development and functions. *Mol. Cell. Endocrinol.* **468**, 47–59 (2018).

108. Afshar, P. T. & Wong, W. H. COSINE: non-seeding method for mapping long noisy sequences. *Nucleic Acids Res.* **45**, e132 (2017).

109. Samson, R. *et al.* Metagenomic insights to understand transient influence of Yamuna River on taxonomic and functional aspects of bacterial and archaeal communities of River Ganges. *Sci. Total Environ.* **674**, 288–299 (2019).

110. Patel, A. *et al.* MinION rapid sequencing: Review of potential applications in neurosurgery. *Surg. Neurol. Int.* **9**, 157 (2018).

111. Starostik, P. Clinical mutation assay of tumors. *Anticancer. Drugs* **28**, 1–10 (2017).

112. Owen, S. V, Perez-Sepulveda, B. M. & Adriaenssens, E. M. Detection of Bacteriophages: Sequence-Based Systems. *Bacteriophages Biol. Technol. Ther.* 1–25 (2018).

113. Cali, D. S., Kim, J. S., Ghose, S., Alkan, C. & Mutlu, O. Nanopore Sequencing Technology and Tools for Genome Assembly: Computational Analysis of the Current State, Bottlenecks and Future Directions. in *Pacific Symposium on Biocomputing Poster Session* (2017).

114. Chu, J. Overlapping long sequence reads: Current in- novations and challenges in developing sensi- tive, specific and scalable algorithms. *Bioarxiv* 1–3 (2016).

115. Faucon, P., Trevino, R., Balachandran, P., Standage-Beier, K. & Wang, X. High Accuracy Base Calls in Nanopore Sequencing. in *Proceedings of the 6th International Conference on Bioinformatics and Biomedical Science* 12–16 (ACM, 2017).

116. Corni, E. *et al.* Low-power portable devices for metagenomics analysis: Fog computing makes bioinformatics ready for the Internet of Things. *Futur. Gener. Comput. Syst.* **88**, 467–478 (2018).

117. Yeh, C.-M., Liu, Z.-J. & Tsai, W.-C. Advanced Applications of Next-Generation Sequencing Technologies to Orchid Biology. *Curr. Issues Mol. Biol.* 51–70 (2018).

118. Hu, Y. O. O. *et al.* Stationary and portable sequencing-based approaches for tracing wastewater contamination in urban stormwater systems. *Sci. Rep.* **8**, 11907 (2018).

119. Zhao, F. & Bajic, V. B. The Value and Significance of Metagenomics of Marine Environments. *Genomics, Proteomics Bioinforma.* **13**, 271–274 (2015).

120. Yamazaki, H., Esashika, K. & Saiki, T. A 150 nm ultraviolet excitation volume on a porous silicon membrane for direct optical observation of dna coil relaxation during capture into nanopores. *Nano Futur.* **1**, 11001 (2017).

121. Stein, D. Nanopore Sequencing: Forcing Improved Resolution. *Biophys. J.* **109**, 2001–2002 (2015).

122. Yanagi, I., Hamamura, H., Akahori, R. & Takeda, K. I. Two-step breakdown of a SiN membrane for nanopore fabrication: Formation of thin portion and penetration. *Sci. Rep.* **8**, (2018).

123. Rau, T., Weik, F. & Holm, C. A dsDNA model optimized for electrokinetic applications. *Soft Matter* **13**, 3918–3926 (2017).

124. Huang, Y. C., Dang, V. D., Chang, N. C. & Wang, J. Multiple large inversions and breakpoint rewiring of gene expression in the evolution of the fire ant social supergene. *Proc. R. Soc. B Biol. Sci.* **285**, 20180221 (2018).

125. Lewandowski, K. *et al.* The Effect of Nucleic Acid Extraction Platforms and Sample Storage on the Integrity of Viral RNA for Use in Whole Genome Sequencing. *J. Mol. Diagnostics* **19**, 303–312 (2017).

126. Bleidorn, C. & Bleidorn, C. Assembly and Data Quality. in *Phylogenomics* 81–103 (Springer, 2017).

127. Baldwin-Brown, J. G., Weeks, S. C. & Long, A. D. A New Standard for Crustacean Genomes: The Highly Contiguous, Annotated Genome Assembly of the Clam Shrimp *Eulimnadia texana* Reveals HOX Gene Order and Identifies the Sex Chromosome. *Genome Biol. Evol.* **10**, 143–156 (2017).

128. Delehelle, F., Cussat-Blanc, S., Alliot, J. M., Luga, H. & Balaresque, P. ASGART: Fast and parallel genome scale segmental duplications mapping. *Bioinformatics* **34**, 2708–2714 (2018).

129. Dutta, G. *et al.* Microfluidic Devices for Label-Free DNA Detection. *Chemosensors* **6**, 43 (2018).

130. Kaplan, R., Yavits, L. & Ginosar, R. RASSA: Resistive Pre-Alignment Accelerator for Approximate DNA Long Read Mapping. *IEEE Micro* (2018).

131. H., G. & G., M. T. Next-generation sequencing platforms for latest livestock reference genome assemblies. *African J. Biotechnol.* **17**, 1232–1240 (2018).

132. Arango-Argoty, G. A. *et al.* NanoARG: a web service for detecting and contextualizing antimicrobial resistance genes from nanopore-derived metagenomes. *Microbiome* **7**, 88 (2019).

133. Najafi, A. *et al.* Fundamental Limits of Pooled-DNA Sequencing. *arXiv Prepr. arXiv1604.04735* (2016).

134. Radko, S. P. *et al.* Prospects for the use of third generation sequencers for quantitative profiling of transcriptome. *Biomed. Chem. Res. Methods* **1**, e00086 (2018).

135. Hansen, S. *et al.* Combination random isothermal amplification and nanopore sequencing for rapid identification of the causative agent of an outbreak. *J. Clin. Virol.* **106**, 23–27 (2018).

136. Horbal, L. & Luzhetskyy, A. The Genetic System of Actinobacteria. in *Biology and Biotechnology of Actinobacteria* 79–121 (Springer, 2017). -1_5

137. Kaplan, R., Yavits, L. & Ginosar, R. RASSA: Resistive Pre-Alignment Accelerator for Approximate DNA Long Read Mapping. *IEEE Micro* (2018).

138. Hehir-kwa, J. Y., Tops, B. B. J., Kemmeren, P., Tops, B. B. J. & The, P. K. Expert Review of Molecular Diagnostics The clinical implementation of copy number detection in the age of next-generation sequencing. *Expert Rev. Mol. Diagn.* **18**, 907–915 (2018).

139. Handler, K. *et al.* Single-Cell Transcriptomics in Cancer Immunobiology: The Future of Precision Oncology. *Front. Immunol.* **9**, 2582 (2018).

140. Karaca, M. & Ince, A. G. Molecular markers in Salvia L.: Past, present and future. in *Salvia Biotechnology* 291–398 (Springer, 2018).

141. Arango-Argoty, G. A. *et al.* NanoARG: A web service for identification of antimicrobial resistance elements from nanopore-derived environmental metagenomes. *bioRxiv* 483248 (2018).

142. Harel, N., Meir, M., Gophna, U. & Stern, A. Sequencing Complete Genomes of RNA Viruses With MinION Nanopore: Finding Associations Between Mutations. *bioRxiv* 575480 (2019).

143. Druley, T. E. Minimal Residual Disease Testing. *Minimal Residual Dis. Test.* (2018).

144. Suwinski, P. *et al.* Advancing Personalized Medicine Through the Application of Whole Exome Sequencing and Big Data Analytics. *Front. Genet.* **10**, 49 (2019).

145. Nellist, C. F. Disease Resistance in Polyploid Strawberry. in *The Genomes of Rosaceous Berries and Their Wild Relatives* 79–94 (Springer, 2018).

146. Gilchrist, C. A. The E. histolytica Genome Structure and Virulence. *Curr. Trop. Med. Reports* **3**, 158–163 (2016).

147. Krizanovic, K., Sovic, I., Krpelnik, I. & Sikic, M. RNA Transcriptome Mapping with GraphMap. in *bioRxiv* (2017).

148. Franus, W., Nowak, R. M. & Kuśmirek, W. Scaffolding algorithm using second- and third-generation reads. in *Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2018* **10808**, 82 (International Society for Optics and Photonics, 2018).

149. Walsh, D. M. *et al.* the Airway Microbiome After Burn and Inhalation Injury. *Chapel Hill* (2016).

150. Nowak, R. M., Forc, M. & Kuśmirek, W. *De Novo* genome assembly for third generation sequencing data. in *Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2018* **10808**, 112 (International Society for Optics and Photonics, 2018).

151. Jahn, S. C. & Starostik, P. Clinical Lung Cancer Mutation Detection. *A Glob. Sci. Vis. - Prev. Diagnosis, Treat. Lung Cancer* 83 (2017).

152. Zascavage, R. R., Thorson, K. & Planz, J. V. Nanopore sequencing: An enrichment-free alternative to mitochondrial DNA sequencing. *Electrophoresis* **40**, 272–280 (2019).

153. Dufort y Álvarez, G. *et al.* Compression of Nanopore FASTQ Files. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **11465 LNBI**, 36–47 (Springer, 2019).

154. Song, S., Guo, Y., Kim, J. S., Wang, X. & Wood, T. K. Phages Mediate Bacterial Self-Recognition. *Cell Rep.* **27**, 737-749.e4 (2019).

155. Sagar, S. When two is better than one. *Drapers* 20 (2008).

156. Dilthey, A., Meyer, S. & Kaasch, A. J. Increasing the efficiency of long-read sequencing for

hybrid assembly with k-mer-based multiplexing. *bioRxiv* 680827 (2019).

157. Krizanovic, K., Sovic, I., Krpelnik, I. & Sikic, M. RNA Transcriptome Mapping with GraphMap. *bioRxiv* 160085 (2017).

158. Baldwin-Brown, J. G., Weeks, S. C. & Long, A. D. A New Standard for Crustacean Genomes: The Highly Contiguous, Annotated Genome Assembly of the Clam Shrimp *Eulimnadia texana* Reveals HOX Gene Order and Identifies the Sex Chromosome. *Genome Biol. Evol.* **10**, 143–156 (2017).

159. Tajiri, M. Comparison of High-Throughput Sequencing for Phage Display Peptide Screening on Two Commercially Available Platforms. *Int. J. Pept. Res. Ther.* 1–7 (2019).

160. Peng, M., Zhang, Y.-P., Ma, Z. (Sam), Li, L. & Ye, C. Hybrid assembly of ultra-long nanopore reads augmented with 10×-genomics contigs: Demonstrated with a human genome. *Genomics* 1–6 (2018).

161. Mishra, D. C. *et al.* Strategies and Tools for Sequencing and Assembly of Plant Genomes. in *The Potato Genome* 81–93 (Springer, 2017).

162. Liu, B., Liu, Y., Zang, T. & Wang, Y. deSALT: fast and accurate long transcriptomic read alignment with de Bruijn graph-based index. *bioRxiv* 612176 (2019).

163. Greig, D. R., Jenkins, C., Gharbia, S. & Dallman, T. J. Comparison of single nucleotide variants identified by Illumina and Oxford Nanopore technologies in the context of a potential outbreak of Shiga Toxin Producing Escherichia coli. *bioRxiv* 570192 (2019).

164. Maroilley, T. & Tarailo-Graovac, M. Uncovering Missing Heritability in Rare Diseases. *Genes (Basel).* **10**, 275 (2019).

165. Liao, X. *et al.* Current challenges and solutions of de novo assembly. *Quant. Biol.* 1–20 (2019).

166. Krachunov, M., Nisheva, M. & Vassilev, D. Machine learning models for error detection in metagenomics and polyploid sequencing data. *Inf.* **10**, 110 (2019).

167. Banin, A. N. *et al.* Development of a Versatile, Near Full Genome Amplification and Sequencing Approach for a Broad Variety of HIV-1 Group M Variants. *Viruses* **11**, 317 (2019).

168. Tapinos, A. *et al.* The Utility of Data Transformation for Alignment, *De Novo* Assembly and Classification of Short Read Virus Sequences. *Viruses* **11**, 394 (2019).

169. Murray, K., Dunigan, D. D. & Sayood, K. Dictionary coded profiles and their use with nanopore sequencers. in *IEEE International Conference on Electro Information Technology* 422–426 (IEEE, 2017).

170. Krachunov, M., Nisheva, M. & Vassilev, D. Application of Machine Learning Models in Error and Variant Detection in High-Variation Genomics Datasets. *Computers* **6**, 29 (2017).

171. Deveson, I. W. *et al.* Chiral DNA sequences as commutable controls for clinical genomics. *Nat. Commun.* **10**, 1342 (2019).

172. Coughlan, S. Pathogen genomics of Methicillin resistant Staphylococcus aureus and Leishmania. *DNA* **540**, G21E (2017).

173. Deveson. Chiral DNA sequences as commutable reference standards for clinical genomics. *bioRxiv* **August 31**, 404285 (2018).

174. Goldstein, S., Beka, L., Graf, J. & Klassen, J. L. Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. *BMC Genomics* **20**, 23 (2019).

175. Johri, S., Doane, M., Allen, L. & Dinsdale, E. Taking Advantage of the Genomics Revolution for Monitoring and Conservation of Chondrichthyan Populations. *Diversity* **11**, 49 (2019).

176. Corresp, J. I. K., Smolander, O. & Pereira, P. A. B. gapFinisher : a reliable gap filling pipeline for SSPACE- LongRead scaffolder output. *PeerJ Prepr.* **5**, e3467v1 (2017).

177. Krych, L. *et al.* Finally, Bulk Typing of Bacterial Species down to Strain Level using ON-rep-seq. *bioRxiv* 34 (2018).

178. Voorhuijzen-Harink, M. M. *et al.* Toward on-site food authentication using nanopore sequencing. *Food Chem. X* **2**, 100035 (2019).

179. Pancrace, C., Gugger, M. & Calteau, A. Genomics of NRPS/PKS Biosynthetic Gene Clusters in Cyanobacteria. *Cyanobacteria Omi. Manip.* **32**, 55–74 (2016).

180. Sullivan, R., Yau, W. Y., O'Connor, E. & Houlden, H. Spinocerebellar ataxia: an update. *J. Neurol.* **266**, 533–544 (2019).

181. Kiel, M. *et al.* Identification of novel biomarkers for priority serotypes of Shiga toxin-producing *Escherichia coli* and the development of multiplex PCR for their detection. *Front. Microbiol.* **9**, 1321 (2018).

182. Geng, Y. *et al.* A crowdsourcing method for correcting sequencing errors for the third-generation sequencing data. in *Proceedings - 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017* **2017-January**, 1626–1633 (IEEE, 2017).

183. Rofeh, J. *et al.* Microfluidic block copolymer membrane arrays for nanopore DNA sequencing. *Appl. Phys. Lett.* **114**, 213701 (2019).

184. Anderson, M. W. Emerging Next-Generation Sequencing Technologies. in *Genomic Applications in Pathology* 29–39 (Springer, 2019).

185. Gatzmann, F. *et al.* The methylome of the marbled crayfish links gene body methylation to stable expression of poorly accessible genes. *Epigenetics Chromatin* **11**, 57 (2018).

186. Schonrock, N. *et al.* The RNA modification landscape in human disease. *Rna* **23**, 1754–1769 (2017).

187. Lu, H., Giordano, F. & Ning, Z. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics Bioinforma.* **14**, 265–279 (2016).

Introduction

In 2006 the release of the first examples of next generation sequencing technologies heralded a wave of novel discovery in the field of biological and medical sciences. In the early stages of the development of these systems it was clear that the technology had huge potential. Although these early iterations had a great deal of promise, there were limitations. Whilst the Illumina short read sequencing technology could produce a large volume of data in comparison to existing technology, systematic constraints meant that the data generated had to be treated in a very different way and new bioinformatic techniques had to be developed. The read lengths generated by the early Illumina machines were short (~32 bp) and the error rate was high. Pyrosequencing could produce longer reads (~100 bp or even longer) but could only produce a fraction of the volume of data. However, these new technologies, experimental protocols and data analysis pipelines quickly matured and particularly Illumina went on to dominate the sequencing landscape for the next decade.

The move towards third generation sequencing technologies was the next major development, these new technologies released in the early 2010’s aimed to address the major limitations of the NGS by producing ultra-long reads. These ultra-long sequence reads would resolve questions which previous technologies had been unable to address such as the accurate scaffolding of genome assemblies generated using short read sequencing and the resolution of long repeat regions of genomic sequence or haplotypes. There are/were two main competitors, Pacific Biosciences Single Molecule Real Time (SMRT) sequencing and Oxford Nanopore Technologies (ONT) nanopore sequencing.

ONT comercially released their first sequencer, the MinION in May 2015 but released it to selected research groups prior to this in an early access evaluation program (MinION Early Access Program) in 2014. The MinION was the first commercial sequencer to utilise nanopore sequencing: the determination of the sequence of discreet molecules of DNA from the changes in electrical field as the bases pass through protein nanopores in an electrically resistant polymer membrane. The MinION does not measure each base individually, instead it records the changes per 5 bases (5mers) (or more recently 6mers) that pass through the pore. This process complicates the post sequencing analysis somewhat. As the molecule of DNA or RNA moves through the nanopore one base at a time, the micro changes in current across 5mer is recorded. These then require complicated algorithms to determine the single base sequence. Sequencing single molecules of template DNA eradicated the need for a PCR amplification step in the sequencing process therefore avoiding the biases of this process.

Prior to this study limited data had been published concerning the volume, quality and limitations of the data produced by the ONT MinION Mikheyev and Tin succeeded in sequencing the Lambda Phage genome although this study suggested that less than 1% of data generated by their sequencing run was alignable to the reference [1]. Loman *et al* and Ashton *et al* were more optimistic, claiming to sequence the entire *E. coli* genome and resolve some interesting genomic features of the *Salmonella typhi* genome respectively [2,3].

Although there was limited data from projects utilising this exciting novel sequencing technology, the potential seemed huge, particularly for the field of bacterial genomics. The possibility of ultra-long, single molecule reads of DNA sequence had the potential to answer many important questions which researchers had been unable to answer due to the short read limitations of previous technologies. Certain facets of bacterial DNA sequence including long repetitive regions and multiple sequence repeats within the genomes or plasmids of bacterial strains meant the analysis of short read data was unable to generate contiguous assemblies of these strains. The ONT MinION offered these long contiguous reads which could be used to sequence these regions and close assemblies. Not only this but the portable nature and low logistical overhead of the MinION meant that there was the potential to utilise the MinION in the field to track bacterial disease outbreaks and as an *in situ* environmental sensor.

The aim of the study covered in this chapter was to assess the potential of the ONT MinION as a tool for use in the field of bacterial sequencing. The work covered thus far in this thesis utilised NGS technologies to investigate bacterial genomics. However, the advent of third generation sequencing presented new opportunities to further advance the field. The ONT MinION in particular posses`sed exciting characteristics which could potentially be utilised to conduct studies such as those in this thesis quickly in the field characterising novel bacterial outbreaks and tracking their spread. The aim was to fully evaluate the volume and error rate of the sequence data generated and characterise any systematic biases which may need to be taken into consideration. During the MAP the sequencing chemistry was evolving at a fast pace, we focussed on the R6 version (the latest version available at the time). We aimed to assess the utility of Minion data for bacterial genomics and meta-genomics. To this end we sequenced a mix of three species with a variety of G + C contents: *Borrelia burgdorferi* (28.6%), *Streptomyces avermitilis* (70.7%) and *E. coli* (50.8%).

Author contribution

The author conducted all initial concept design, planning and pre project research. The author was also responsible for all bioinformatic analysis on this project, using bespoke scripts and pipeline code to assess, analyse and optimise the performance of this exciting novel technology. The project included the assessment and optimisation of initial tools to process the novel data types produced and where required the production of bespoke scripts to handle the unique analysis opportunities presented by this project.

The author also contributed significantly to the pre-project research, concept design and planning for the project along with the writing, editing and submission of manuscript and the production and editing of all figures and tables

Manuscript:

Biomolecular Detection and Quantification 3 (2015) 1–8

Contents lists available at ScienceDirect

Biomolecular Detection and Quantification

journal homepage: www.elsevier.com/locate/bdq

Original Article

Assessing the performance of the Oxford Nanopore Technologies MinION

T. Laver[a,*,1], J. Harrison[a,1], P.A. O'Neill[a,b], K. Moore[a,b], A. Farbos[a,b], K. Paszkiewicz[a,b], D.J. Studholme[a]

[a] *Biosciences, University of Exeter, Geoffrey Pope Building, Stocker Road, Exeter EX4 4QD, UK*
[b] *Wellcome Trust Biomedical Informatics Hub, Geoffrey Pope Building, Stocker Road, University of Exeter, Exeter EX4 4QD, UK*

ARTICLE INFO

Article history:
Received 4 December 2014
Received in revised form 5 February 2015
Accepted 18 February 2015

Keywords:
DNA sequencing
MinION
Nanopore

ABSTRACT

The Oxford Nanopore Technologies (ONT) MinION is a new sequencing technology that potentially offers read lengths of tens of kilobases (kb) limited only by the length of DNA molecules presented to it. The device has a low capital cost, is by far the most portable DNA sequencer available, and can produce data in real-time. It has numerous prospective applications including improving genome sequence assemblies and resolution of repeat-rich regions. Before such a technology is widely adopted, it is important to assess its performance and limitations in respect of throughput and accuracy. In this study we assessed the performance of the MinION by re-sequencing three bacterial genomes, with very different nucleotide compositions ranging from 28.6% to 70.7%; the high G + C strain was underrepresented in the sequencing reads. We estimate the error rate of the MinION (after base calling) to be 38.2%. Mean and median read lengths were 2 kb and 1 kb respectively, while the longest single read was 98 kb. The whole length of a 5 kb rRNA operon was covered by a single read. As the first nanopore-based single molecule sequencer available to researchers, the MinION is an exciting prospect; however, the current error rate limits its ability to compete with existing sequencing technologies, though we do show that MinION sequence reads can enhance contiguity of de novo assembly when used in conjunction with Illumina MiSeq data.

1. Introduction

The Oxford Nanopore Technologies (ONT) MinION [20] is a new sequencing technology that is currently available as part of an early access and development scheme: the MinION Access Programme [21]. This programme allowed early access to the MinION for participating sequencing centres. The results produced by this study are based on the first round of the ONT MinION Access Programme, using the company's R6 sequencing chemistry.

The MinION will most likely be the first commercially available sequencer that uses nanopores. Nanopore sequencing has been shown to be able to discriminate individual nucleotides by measuring the change in electrical conductivity as DNA molecules pass through the pore [23,28]. Nanopore sequencing does not rely on sequencing by synthesis as most current major technologies do. Laszlo et al. [13] sequenced the phi X 174 genome using another nanopore based technology, demonstrating that nanopore sequencing can produce long reads that are accurate enough to enable them to be aligned back to their reference genomes.

The MinION has several attributes that give it the potential to replace or complement existing sequencing technologies for some applications. The technology offers read lengths of tens of kilobases, with theoretically no instrument-imposed limitation on the size of reads that can be generated. The MinION uses nanopores to sequence a single DNA molecule per pore [11]; this has significant potential advantages over the current widely used sequencing technologies (Ion Torrent, Illumina), which rely on sequencing clusters of amplified DNA molecules. Sequencing a single molecule removes the necessity for PCR amplification and its associated biases [1]. The device has a low capital cost, is by far the most portable DNA sequencer available and can produce data in real-time, although at this stage the samples still require library preparation prior to sequencing – a process that has yet to be optimised. It has applications in scaffolding genome sequences

Abbreviations: NRPS, non-ribosomal peptide synthase; ONT, Oxford Nanopore Technologies.

* Corresponding author. Tel.: +44 7706943765.
E-mail addresses: twl207@exeter.ac.uk (T. Laver), jh288@exeter.ac.uk (J. Harrison), P.A.O'Neill@exeter.ac.uk (P.A. O'Neill), K.A.Moore@exeter.ac.uk (K. Moore), A.Farbos@exeter.ac.uk (A. Farbos), K.H.Paszkiewicz@exeter.ac.uk (K. Paszkiewicz), D.J.Studholme@exeter.ac.uk (D.J. Studholme).

[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.bdq.2015.02.001

assembled from short reads [3,31] and resolving repeat sequences or haplotypes, being able to span ambiguous regions in a single read, as has been demonstrated for PacBio [27,9]. Future developments may include use in real-time medical diagnostics and forensics, as well as prospective applications as an environmental DNA sensor.

As the MinION is still in its testing stage there is very limited data published data on its performance. Mikheyev and Tin [19] sequenced the lambda phage genome, reporting that, when unalignable reads are taken into account, less than 1% of the sequence produced by the MinION is identical to the reference. Quick et al. [24] were able to sequence an *Escherichia coli* genome demonstrating that the MinION is able to sequence entire bacterial genomes. Ashton et al. [2] used the MinION to resolve the structure and chromosomal insertion site of an antibiotic resistance island in *Salmonella typhi*. They estimated the median accuracy of their MinION data to be between 61.6% and 71.5% based on mapping back to the reference. De novo genome assembly using MinION reads has been demonstrated to achieve improved assembly compared to Illumina sequencing alone by [7].

During the MinION DNA library preparation hairpin structures are added to the end of the double stranded fragments, these fragments are then denatured resulting in one length of single stranded DNA consisting of the forward strand followed by the hairpin sequence then the reverse strand [24]. The MinION generates up to three different types of read for each fragment of DNA that passes through a pore: 'Template', 'Complement' and 'Two Direction'. Initially, the forward strand is sequenced generating the Template read then the hairpin structure is read through followed by the reverse strand, generating the Complement read. Finally, the ONT base calling software attempts to call a consensus sequence of the Template and Complement reads; this resulting consensus sequence is referred to as a Two Direction read. Not all fragments that pass through the pore result in generation of all three read types; some only result in the Template read as output, others in Template and Complement, while only a small minority produce Template, Complement and Two Direction reads. One objective of the current study was to assess whether there were differences between the three types of read, such as read G + C content, read length and error rate.

Extreme G + C content is known to affect the performance of DNA sequencers [1]. To investigate whether the MinION was affected by the nucleotide composition of the target DNA this study re-sequenced a mix of three bacteria with a range G + C content *Borrelia burgdorferi* (28.6%), *Streptomyces avermitilis* (70.7%) and *E. coli* (50.8%).

2. Methods

2.1. Bacterial DNA

Bacterial DNA was obtained from American Type Culture Collection (ATCC) for *S. avermitilis* (ATCC 35210), *B. burgdorferi* (ATCC 31267) and *E. coli* K-12 (ATCC 10798).

2.2. MinION sequencing

1 μg DNA was fragmented using Covaris g-tube centrifuged at 5000 × *g* for 60 s. 5 μl lambda phage spike-in DNA (CS, ONT) was added to each sample. Fragments were end-repaired and adenylated using NEXTflex Rapid DNAseq kit (Newmarket Scientific #5144-02), purified and concentrated using Ampure XP beads (Beckmann Coulter). Size distribution was checked on a Bioanalyser 7500 DNA chip (Agilent Technologies) (Supplementary Fig. 1) and the concentration determined using the Qubit BR assay (Life Technologies) before pooling DNA from each species in 50 μl: *S. avermitilis* 576 ng, *B. burgdorferi* 560 ng and *E. coli* 530 ng.

The ONT protocol was followed unless indicated and all reactions carried out at room temperature. Adapters were ligated to the adenylated DNA and purified using 0.4 × volume Ampure XP beads (Beckman Coulter); beads were washed with ONT-supplied wash buffer, and eluted in 25 μl ONT supplied elution buffer. Tether was annealed for 10 min and the library conditioned with the HP motor for 30 min. This pre-sequencing mix was stored briefly on ice. Immediately before sequencing, 6 μl pre-sequencing mix, 140 μl EP and 4 μl fuel mix were mixed very gently before loading on to the MinION flowcell. Additional input material was added to the MinION flowcell at 16 h 33 min.

2.3. MiSeq sequencing

For each species (*S. avermitilis*, *B. burgdorferi* and *E. coli*) Illumina fragment libraries were prepared and those containing insert sized averaging 550 bp were selected. DNA was sequenced (300 bp Paired End) on a MiSeq using v3 reagents. Supplementary Table 1 details the number of reads produced for each species.

Data available at the SRA: *B. burgdorferi* SRR1772332, *E. coli* SRR1770413, *S. avermitilis* SRR1770414.

2.4. Alignment of MinION reads against reference genome sequences

After sequencing and base calling reads were converted to fasta using Poretools [17] then aligned against a database of the closest available reference genomes for those species: *B. burgdorferi* ATCC 31267 (NC_001318) [5], *S. avermitilis* ATCC 35210 (NC_003155) [10] and *E. coli* strain MG1655 (NC_000913) [26], plus the 3.56 kb sequence of the lambda phage spike-in. The alignment of the MinION reads to the reference genomes was carried out using the LAST alignment software [6,12], as in [24]. The best alignment for each read was selected based on alignment score. Using LAST we aligned 12,632 reads (26.8%) and 38280405 (40.7%) bases. LAST was designed to cope well with long error-prone reads, resulting in higher mapping rates than alignment software designed for short high-fidelity reads such as BWA [15] or Bowtie2 (Langmead and Salzberg, 2012). An update for BWA mem [15] has been released designed for ONT reads. While its author suggests its performance will typically still be inferior to LAST [16] our results suggest the alignment rate is comparable, making it another viable option for aligning MinION reads (Supplementary Table 2).

2.5. Calculation of error rates from LAST sequence alignments

To calculate the error rate we counted the number of mismatch positions in the gapped alignment of a read to a reference sequence, thus it is a measure of substitution, insertion and deletion errors. The error rates were then expressed as a percentage of the length of reference sequence aligned against. Some recorded errors may in fact be genuine differences between our DNA samples and the published reference genome sequences, either due to real polymorphism or errors in the published reference sequences. To estimate the frequency of such false-positive errors we re-sequenced each of our genomic DNA samples using the Illumina MiSeq and hence ascertained the number of discrepancies between our DNA samples and the published reference sequences. The MiSeq reads were aligned against the reference genome sequences using Bowtie2 (Langmead and Salzberg, 2012) and differences to the references were evaluated using SAMtools and BCFtools [14]. Table 1 shows the number of short variations between our data and the published reference genomes, these suggest that approximately 0.009% of the 'errors' in the MinION data are not errors but genuine differences.

Table 1
Differences to published references genomes.

Species	SNPs	Indels
S. avermitilis	722	148
E. coli	402	17
B. burgdorferi	144	16

Clearly this small number of false-positive errors does not substantially affect the overall estimate of sequencing error-rate.

2.6. Calculating G + C content versus coverage

To investigate a potential bias against extreme G + C sequences we split the *E. coli* and *B. burgdorferi* genomes into 1000 bp windows using BEDTools [25] then using the LAST alignment of the MinION reads against the reference genome sequences we evaluated the coverage depth of the alignment for those windows using BEDTools.

2.7. Assembling E. coli using MinION reads

The Illumina MiSeq *E. coli* paired end reads were combined where possible using FLASH [18] resulting in 71456 overlapped reads and 562512 uncombined paired end reads. We extracted the MinION reads that aligned to *E. coli.* We generated an assembly using Spades 3.5.0 [22] (ONT MinION specific setting) with these MinION reads and the Illumina MiSeq data. The assembly was evaluated using QUAST [8].

Table 2
Summary statistics for the MinION reads.

Read type	Read count	Mean length (bp)	Standard deviation of length (bp)	Maximum length (bp)
Template	35,946	1951	3007	98,366
Complement	8270	1827	2549	44,769
Two direction	2877	3088	2958	28,365

3. Results and discussion

3.1. Overview of sequence data

We constructed a sequencing library containing genomic DNA from three bacterial strains in equal quantities, as described in Section 2. This single MinION run generated Template sequence reads for 35,946 different DNA fragments, but only 23.0% produced Complement reads and only 8.0% yielded Two Direction reads (Table 2). The longest single read generated was 98,366 bp. As shown in Fig. 1 reads of this extreme length were the exception and not representative of the distribution; the majority of reads for all three read types have read lengths of less than 2000 bp.

3.2. S. avermitilis sequences were under-represented

By aligning MinION sequence reads against published reference genomes, we tried to assign each read to its most likely genome of origin (i.e. *B. burgdorferi*, *S. avermitilis* or *E. coli*). Reads from *S. avermitilis* were clearly under-represented (Table 3), as there were equal abundances (by mass) of each bacterial genome in the

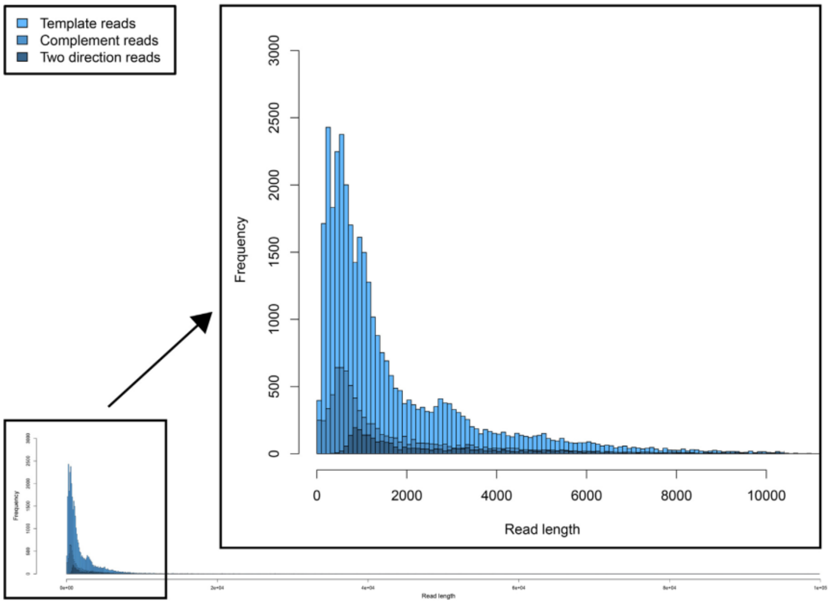


Fig. 1. Distribution of MinION read lengths. Frequency distributions of lengths of reads obtained from the MinION run. Data shown for each of the three read types Template, Complement and Two direction, superimposed.

Table 3
The number of reads of each type which aligned to each species.

Read type	*S. avermitilis*	*E. coli*	*B. burgdorferi*	Lambda	Unaligned
Template	226	2703	6752	1246	25,018
Complement	44	203	773	28	7222
Two direction	0	268	317	71	2221

sequencing library. Given the high G + C content of *S. avermitilis* compared to *B. burgdorferi* or *E. coli*, this suggests that G + C content may be the explanatory factor. However, this analysis does not exclude the possibility that some other property of the *S. avermitilis* DNA was responsible (e.g. methylation or other modification of the DNA). It is also not clear whether the under-representation arises from fewer *S. avermitilis* DNA molecules being sequenced (e.g. because they are out-competed for pores) or if the DNA was sequenced with a higher error rate resulting in lower alignment rates. The overall error rate of the aligned reads is 38.2% but is higher for the *S. avermitilis* reads (Table 4) (4.3 and 5.2 percentage points higher for the template and complement reads respectively). However when the aligned portions of all the reads are examined there is no clear correlation between G + C content and error rate (correlation coefficient of 0.198) (see also supplementary Figs. 2 and 3).

To further explore whether there was a bias against high G + C sequences we split the *E. coli* and *B. burgdorferi* genomes into windows and evaluated the relationship between coverage depth and G + C content. The correlation between high G + C and lower coverage was very weak for *E. coli* (correlation coefficient of −0.0171) while for *B. burgdorferi* it was in the opposite direction (correlation coefficient of 0.444), suggesting that if there is any trend at all, it is that extreme G + C results in lower coverage (Supplementary Figs. 4 and 5). The lack of windows in *E. coli* and *B. burgdorferi* with G + C content as high as *S. avermitilis* prevents a true examination of the effect of extreme G + C using this method.

3.3. G + C content of reads is not the same as the sequence aligned to

The mean G + C content of the MinION reads is 47.2%. As shown in Fig. 2 the GC content of the reads does not correspond to the G + C content of all of the input genomes; extreme G + C sequences which would be expected to be generated from *B. burgdorferi* and *S. avermitilis* are not present in the reads. However as shown in Table, the most likely genome of origin for many reads is *B. burgdorferi*, suggesting that reads were in fact generated from this genome. The lack of extreme G + C reads appears to be due, at least in part, to the fact that the G + C content of the aligned portion of a read is different to that of the section of the reference to which it aligns (Fig. 3). As shown in Fig. 4 the distribution of G + C content for the aligned sections of reads (Fig. 4A) is different to that of the sections of reference sequence to which they align (Fig. 4B); the extremes of G + C content found in the reference seem to be shifted towards intermediate G + C in the reads. This could be caused by substitution errors in the sequencing effectively inserting random bases in the reads which will result in reads with more intermediate G + C content than the sequenced DNA fragment.

Table 4
Error rate of reads split by type and species aligned to.

Read type	*S. avermitilis* (%)	*E. coli* (%)	*B. burgdorferi* (%)	Lambda (%)
Template	42.5	38.2	38.4	36.9
Complement	43.4	38.2	38.2	38.0
Two direction	NA	37.3	40.8	38.4

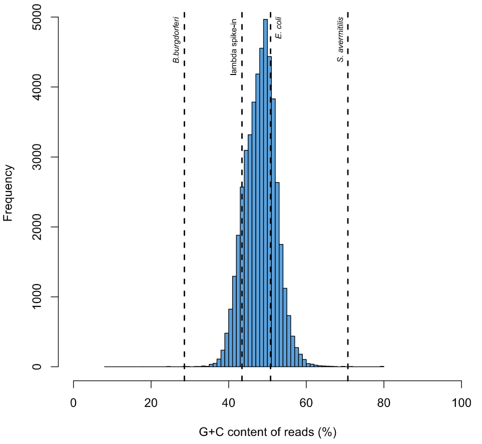


Fig. 2. G + C content of MinION reads. A frequency distribution of the G + C content of reads generated by the MinION run. Mean G + C content of each reference genome is included for comparison.

3.4. 25 genes covered by single MinION read

The long read lengths generated by the MinION have important possible applications not available to traditional short reads sequencing technologies. These reads (up to 98 kb in this study) are more than enough to span important genomic features such as secondary metabolite clusters, repeat rich regions and operons. Several interesting classes of bacterial genes are long and modular, made up of multiple partially repeated segments; these genes include non-ribosomal peptide synthase (NRPS) and TAL effectors. Because of the repetitive nature of these gene sequences, they are notoriously difficult to assemble using short-read sequencing

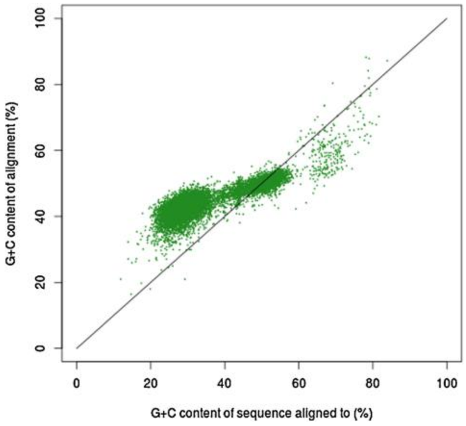


Fig. 3. G + C content of aligned portions of MinION reads against corresponding reference sequence. Plot of G + C content of the aligned portion of a read versus the G + C content of the section of the reference to which it aligns. Included is a line to demonstrate the relationship if the two were equal.

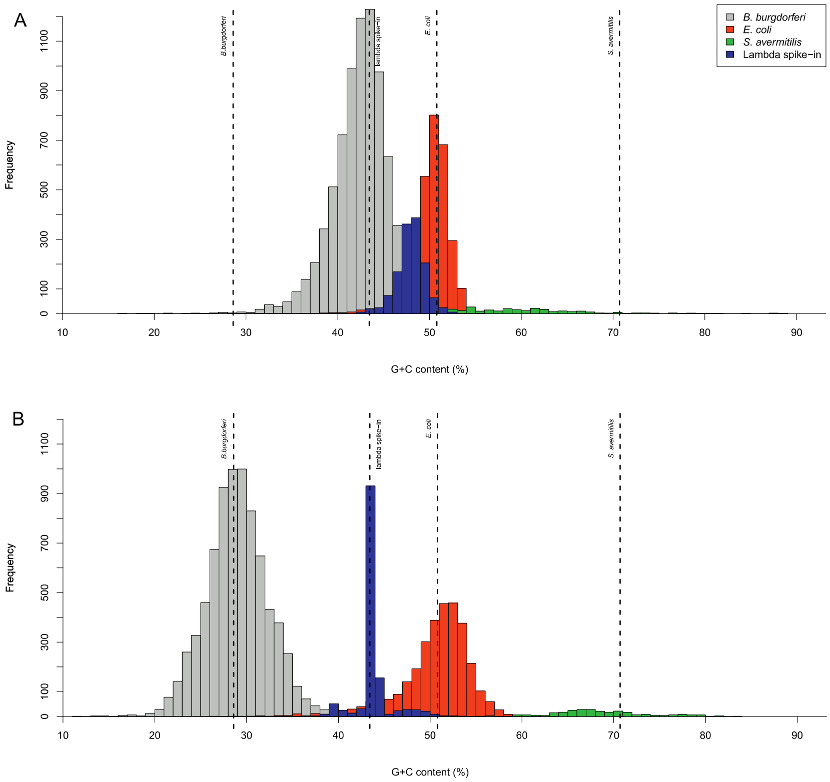


Fig. 4. G + C content of aligned portions of MinION reads and the reference sequence aligned to. Frequency distribution of the G + C content of aligned portions of the reads (A) and the G + C content of the sections of the reference genome they align to (B). Mean G + C content of each reference genome is included for comparison.

technologies. For example, Fig. 5A shows a section of the alignment generated from the MinION sequencing data. Highlighted is a single MinION read aligned to a 20,016 bp region of the reference genome spanning the entire length of one copy of the *E. coli* rDNA operon (5088 bp in length). Fig. 5B shows a NRPS gene cluster in the *E. coli* genome which is 53,661 bp in length and contains 49 genes. This MinION run has generated reads which span large portions of the cluster, one of which covers 28,134 bp of this NRPS cluster including 25 of its constituent genes. Repetitive regions are problematic when trying to assemble genomic data using short read sequencing technologies as it is not possible for one “short read” to span an entire region of interest [30].

3.5. MinION reads improved *E. coli de novo* assembly

To demonstrate how the long reads produced by the MinION can be used to improve genome assemblies we extracted the MinION reads which aligned to the *E. coli* genome and used these in a combined assembly with Illumina MiSeq data. The resulting assembly had 84 contigs of at least 200 bp, a longest contig of 442595 bp and an N50 of 199079 bp compared to the assembly using only MiSeq data which contained 116 contigs, whose longest contig was 299472 bp with an N50 of 159445 bp. However when the assemblies were evaluated using QUAST [8] the results show seven more misassemblies in the MinION aided assembly. These findings show that even at this early stage in the development of this technology, the MinION can offer substantial improvement in assembly length.

3.6. The error rate of the aligned reads remains constant over a MinION run

In order to evaluate the performance of the MinION over the duration of a run and whether there are characteristics of the data which vary over run time, a time series was generated. Higher mean read lengths were observed during the first 8 hours of operation (Fig. 6), perhaps suggesting that, if read length is your primary concern the initial stages of a run are optimal for this purpose. The alignment rate varies across the run time (Fig. 7) while the number of reads generated falls off towards the end of the run. However the error rate of the aligned reads remains relatively consistent throughout the run, suggesting that the quality of the data at the end of the run will not necessarily be any worse than at the beginning, so running the machine for as long as convenient will be beneficial rather than detrimental.

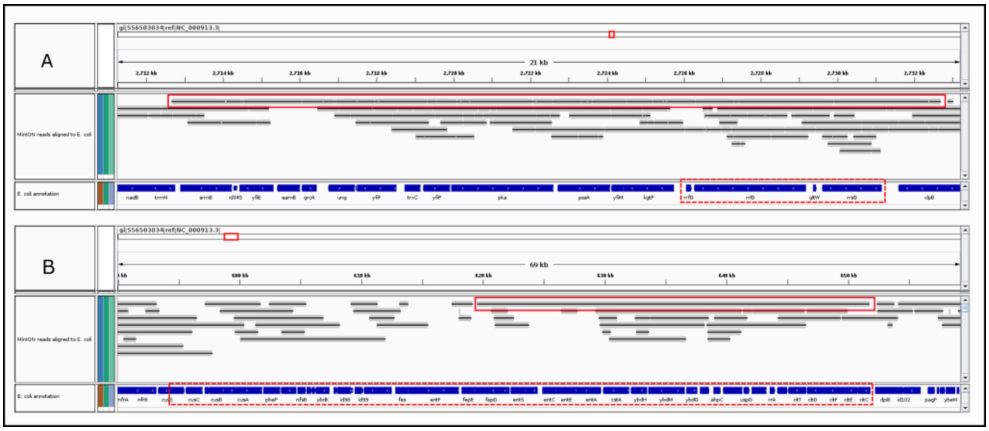

A
21 kb
MinION reads aligned to E. coli
E. coli annotation
B
69 kb
MinION reads aligned to E. coli
E. coli annotation

Fig. 5. Single MinION reads able to span important genes. Images generated using IGV [29] showing the alignment of MinION reads to the *E. coli* reference genome. (A) A rDNA operon. (B) A NRPS gene cluster. Highlighted with continuous red lines are the reads spanning the relevant sections and the dashed lines highlight the genomic regions of interest.

3.7. MinION quality scores do not follow the Phred scale

The per base quality scores of other sequencing technologies correspond with the Phred scale [4] where scores indicate a specific likelihood of error for that base; for example a Phred score of 20 indicates there will be 1 error for every 100 bases with that score. The MinION quality scores do not follow Phred expected error rates; the same quality score for the MinION does not equate to the same error rate as Phred (see supplementary Fig. 8).

3.8. Comparisons to publically available MinION data

The error rates measured on our MinION data are similar to those for other public data on the MinION. Using our methods on the data published by Mikheyev and Tin [19] we calculated their error rate for single direction reads as 40.2% based on 35.1% read aligned (25.4% bases aligned), while 32.7% of their Two Direction reads aligned (11.4% bases aligned) with 40.1% error. This data was generated using the same R6 MinION chemistry as the data published in this study.

Due to the experimental nature of the MinION, the sequencing chemistry is rapidly evolving. The data presented in this study was generated using R6 sequencing chemistry; to explore if our results for error rate and the effect of G + C content held true for the R7 chemistry we evaluated data from [24]. Re-analysing this data with our methods resulted in 57.8% of template reads aligned (55.4% of bases aligned), with 37.5% error, but more promisingly their High Quality Two Direction reads resulted in 82.5% of reads aligned (82.3% of bases aligned), with 26.6% error. This suggests that the error rate for the high quality reads is improving as the technology evolves. As we have already been able to demonstrate that MinION reads can both cover biologically important genes and be used to generate improved genome assemblies the technology will only have more applications as it improves.

To explore if the issues with extreme G + C content sequences that were suggested by our data were still present for the updated

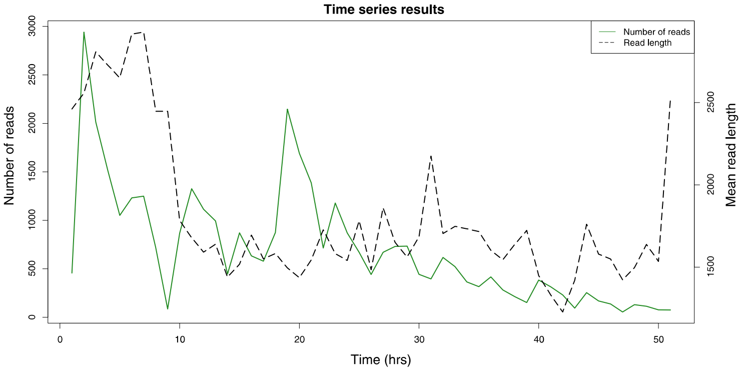


Fig. 6. Read data over time during the MinION run. Plot of number of reads and their mean length generated per hour during the MinION. Additional input material was added to the MinION flowcell at 16 h 33 min.

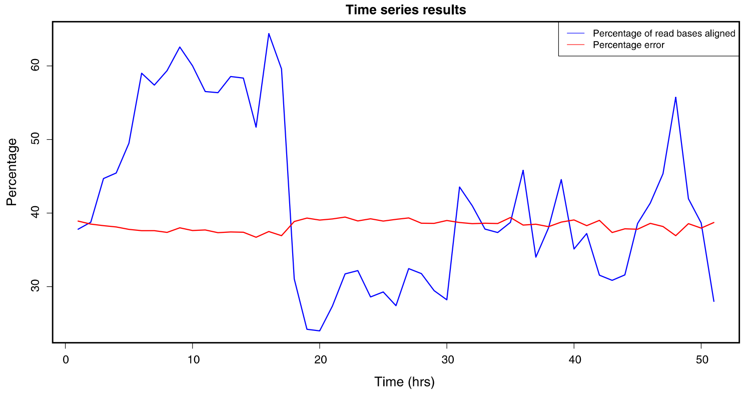


Fig. 7. Fluctuation in alignment and error rates over time during a MinION run. Plot showing percentage of read bases aligned per hour during the MinION run based on the alignment by LAST and their error rate. Additional input material was added to the MinION flowcell at 16 h 33 min.

chemistry we repeated our evaluation of coverage versus G + C content across windows of the *E. coli* genome for the R7 data. The results suggest a weak correlation between G + C content and depth of coverage (correlation of coefficient of −0.141 for Template reads and −0.0816 for High Quality Two Direction) a similar finding to our results gained from the R6 chemistry (Supplementary Figs. 6 and 7).

4. Conclusions

Our results demonstrate that in spite of its high error rate the MinION is able to generate extremely long reads, is able to span regions of interest in a single read and is able to improve the contiguity of genome assemblies. As well as the high error rate, the MinION's possible difficulties with high G + C content sequences, demonstrated in this study, will also need to be addressed before the device is put into widespread use.

Our analysis of data generated by [24] on the R7 MinION chemistry suggests that the error rate for the High Quality Two Direction reads is improving as the technology evolves, although we suggest the potential issues with sequencing extreme G + C sequences is still present. The lower error rate generated from the Two Direction reads produced with the updated MinION chemistry gives cause for optimism that future version of the MinION might be able to generate reads with a greatly reduced error rate while still retaining the long read length and low per unit costs that make this such an exciting technological prospect.

Acknowledgments

This work was funded by the University of Exeter Sequencing Service as part of the Oxford Nanopore MAP programme. The University of Exeter Sequencing Service is supported by the Wellcome Trust Institutional Strategic Support Fund (WT097835MF), Wellcome Trust Multi User Equipment Award (WT101650MA) and BBSRC LOLA award (BB/K003240/1). TL was supported by the BBSRC Industrial Case Studentship award BB/H016120/1. JH was supported by a BBSRC PhD studentship (BB/F017367/1).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bdq.2015.02.001.

References

[1] Aird D, Ross MG, Chen W, Danielsson M, Fennell T, Russ C, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol 2011;12(2):R18.
[2] Ashton PM, Nair S, Dallman T, Rubinio S, Rabsch W, Mwaigwisya S, et al. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. Nat Biotechnol 2015;33:296–300.
[3] Boetzer M, Pirovano W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinform 2014;15(211):1–9.
[4] Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using Phred I accuracy assessment. Genome Res 1998;8(3):175–85.
[5] Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature 1997;390:580–6.
[6] Frith MC, Hamada M, Horton P. Parameters for accurate genome alignment. BMC Bioinform 2010;11(80):1–14.
[7] Goodwin SS, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz M, McCombie R. Oxford nanopore sequencing and de novo assembly of a eukaryotic genome. bioRxiv 2015.
[8] Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics 2013;29(8):1072–5.
[9] Huddleston J, Ranade S, Malig M, Antonacci F, Chaisson M, Hon L, et al. Reconstructing complex regions of genomes using long-read sequencing technology. Genome Res 2014;24:688–96.
[10] Ikeda H, Ishikawa J, Hanamoto A, Shinose M, Kikuchi H, Shiba T, et al. Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. Nat Biotechnol 2003;21(5):526–31.
[11] Kasianowicz JJ, Brandin E, Branton D, Deamer DW. Characterization of individual polynucleotide molecules using a membrane channel. Proc Natl Acad Sci USA 1996;93(24):13770–3.
[12] Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res 2011;21(3):487–93.
[13] Laszlo AAH, Derrington IM, Ross BC, Brinkerhoff H, Adey A, Nova IC, et al. Nanopore sequencing of the phi X 174 genome. Quant Biol 2014:1–39.
[14] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25(16):2078–9.
[15] Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Genomics 2013:1–3.
[16] Li H. BWA-MEM for long error-prone reads; 2014 [Online] Available from: http://lh3.github.io/2014/12/10/bwa-mem-for-long-error-prone-reads/ [accessed 09.01.15].
[17] Loman NJ, Quinlan AR. Poretools: a toolkit for analyzing nanopore sequence data. Bioinformatics 2014;30(23):3399–401.
[18] Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 2011;27(21):2957–63.
[19] Mikheyev AS, Tin MMY. A first look at the Oxford Nanopore MinION sequencer. Mol Ecol Resour 2014;14(6):1097–102.
[20] Nanoporetech.com [1]. The MinION device: a miniaturised sensing system; 2014 [Online] Available from: http://tinyurl.com/m6uboaj [accessed 26.11.14].
[21] Nanoporetech.com [2]. A guide to MAP; 2014 [Online] Available from: http://tinyurl.com/q86a72v [accessed 26.11.14].
[22] Nurk S, Bankevich A, Antipov D, Gurevich AA, Korobeynikov A, Lapidus A, et al. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. J Comput Biol 2013;20(10):714–37.

[23] Olasagasti F, Lieberman KR, Benner S, Cherf GM, Dahl JM, Deamer DW, et al. Replication of individual DNA molecules under electronic control using a protein nanopore. Nat Nanotechnol 2013;5(11):798–806.

[24] Quick J, Quinlan AR, Loman NJ. A reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer. GigaScience 2014;3(22):1–6.

[25] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26(6):841–2.

[26] Riley M, Abe T, Arnaud MB, Berlyn MK, Blattner FR, Chaudhuri RR, et al. *Escherichia coli* K-12: a cooperatively developed annotation snapshot-2005. Nucleic Acids Res 2006;34(1):1–9.

[27] Satou K, Shiroma A, Teruya K, Shimoji M, Nakano K, Juan A, et al. Complete genome sequences of eight *Helicobacter pylori* strains with different virulence factor genotypes and methylation profiles, isolated from patients with diverse gastrointestinal diseases on Okinawa Island, Japan, determined using PacBio single-molecule real-time technology. Genome Announc 2014;2(2):1–2.

[28] Stoddart D, Heron AJ, Mikhailova E, Maglia G, Bayley H. Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. Proc Natl Acad Sci USA 2009;106(19):7702–7.

[29] Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2013;14(2):178–92.

[30] Todd J, Saltzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 2012;13:36–46.

[31] Utturkar SM, Klingeman DM, Land ML, Schadt CW, Doktycz MJ, Pelletier DA, et al. Evaluation and validation of *de novo* and hybrid assembly techniques to derive high-quality genome sequences. Bioinformatics 2014;30(19):2709–16.

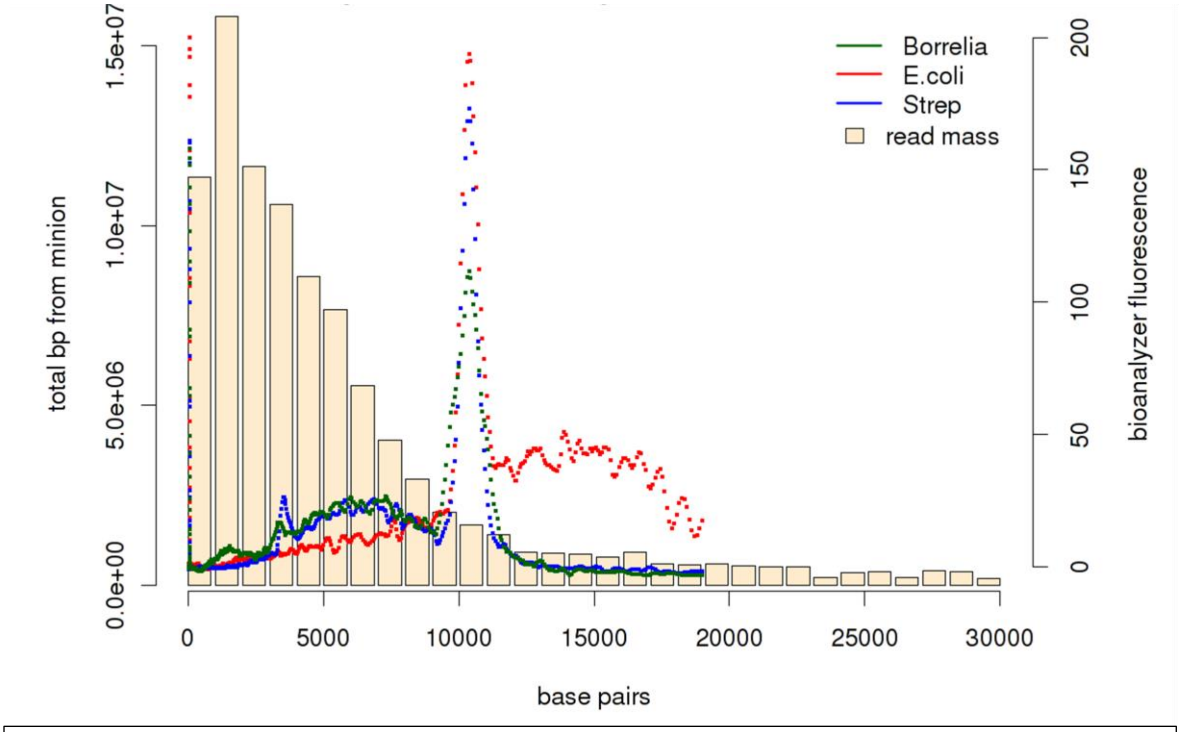
Supplementary Material

Supplementary Table 1 MiSeq sequencing statistics

Species	Raw read pairs	Trimmed read pairs	Genome coverage
S. avermitilis	1232293	1201308	59
E. coli	643253	634237	39
B. burgdorferi	1861953	1851907	585

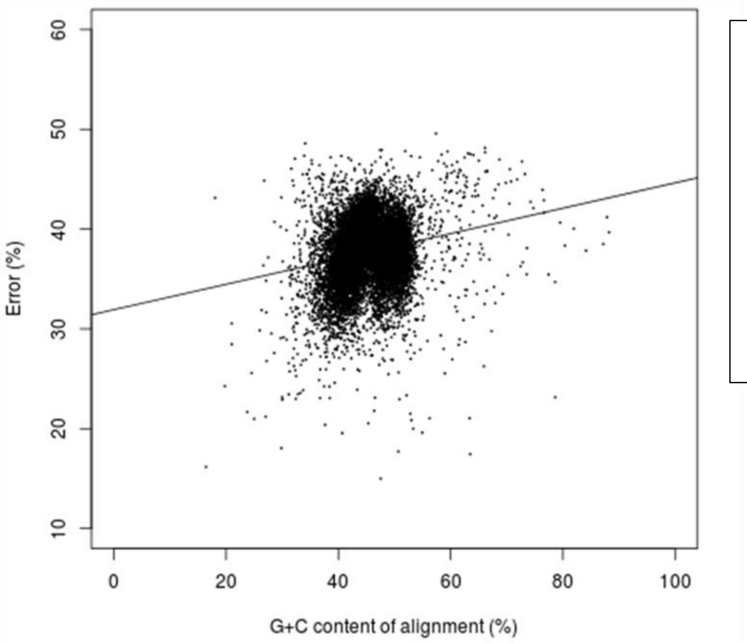
Supplementary Table 2 Alignment rates of alternative software
The alignment rates for alternative alignment software. Bowtie2 was run with --very-sensitive-local setting (the best performing of the pre-sets) while BWA mem was run with default settings and ont2d (the pre-set developed for ONT reads).

Alignment software	Reads aligned	% total reads aligned	Bases aligned	% total bases aligned
Bowtie2 --very-sensitive-local	1303	2.77	1833495	1.95
BWA mem default	2111	4.48	129618	0.138
BWA mem ont2d	15115	32.1	34336255	36.5



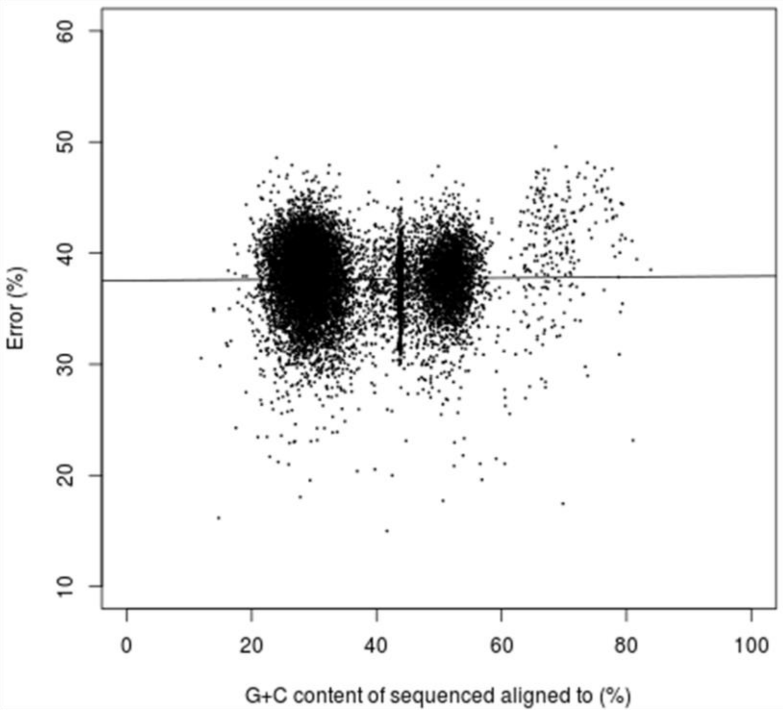
Supplementary Figure 1 Bioanalyser trace compared to read mass

The DNA fragment size distribution of the libraries before mixing was checked on a Bioanalyser 7500 DNA chip (Agilent Technologies). The spike starting at 10000 bases is the ladder. The Bioanlyser trace is overlaid on the read mass for comparison.



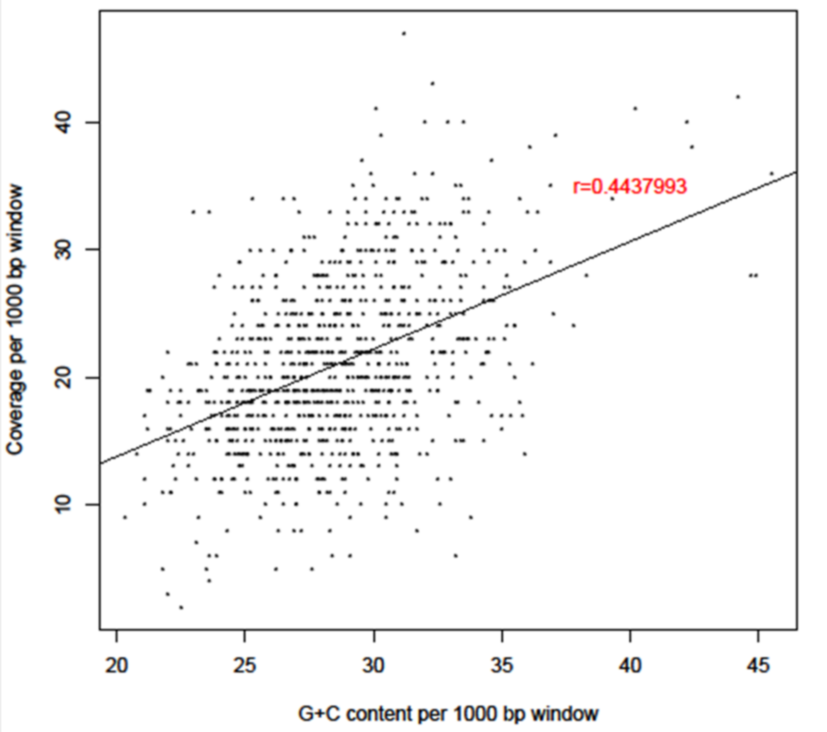
Supplementary Figure 2 G+C content of aligned portions of reads verus their error rate

The G+C content of the aligned portion of a read does not correlate with its error rate (correlation coefficient of 0.198).



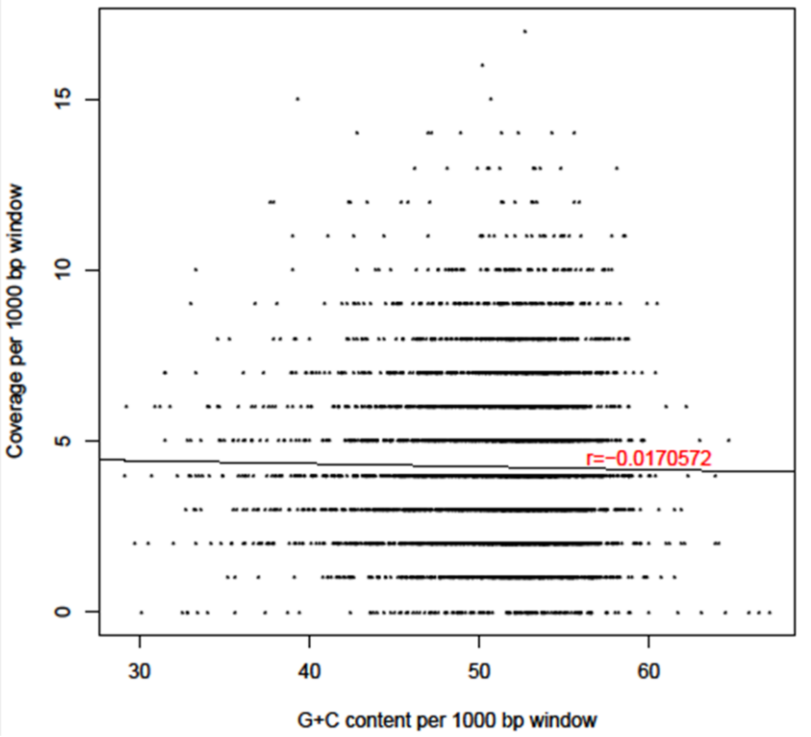
Supplementary Figure 3 GC content of the sections of the reference genome reads align to verus their error rate

The G+C content of the sections of the reference genome to which reads align does not correlate with error rate (correlation coefficient of 0.014).



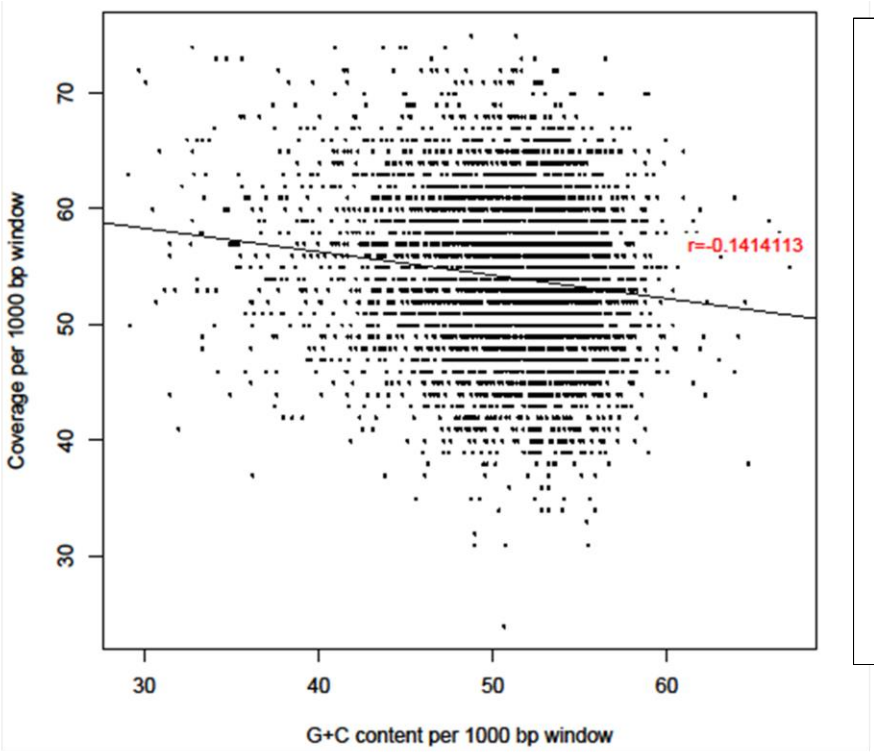
Supplementary Figure 4 G+C content versus coverage depth for *B. burgdorferi*

The *B. burgdorferi* genome was split into windows of 1000 bp across which the coverage depth was compared to the G+C content (r is the correlation coefficient).



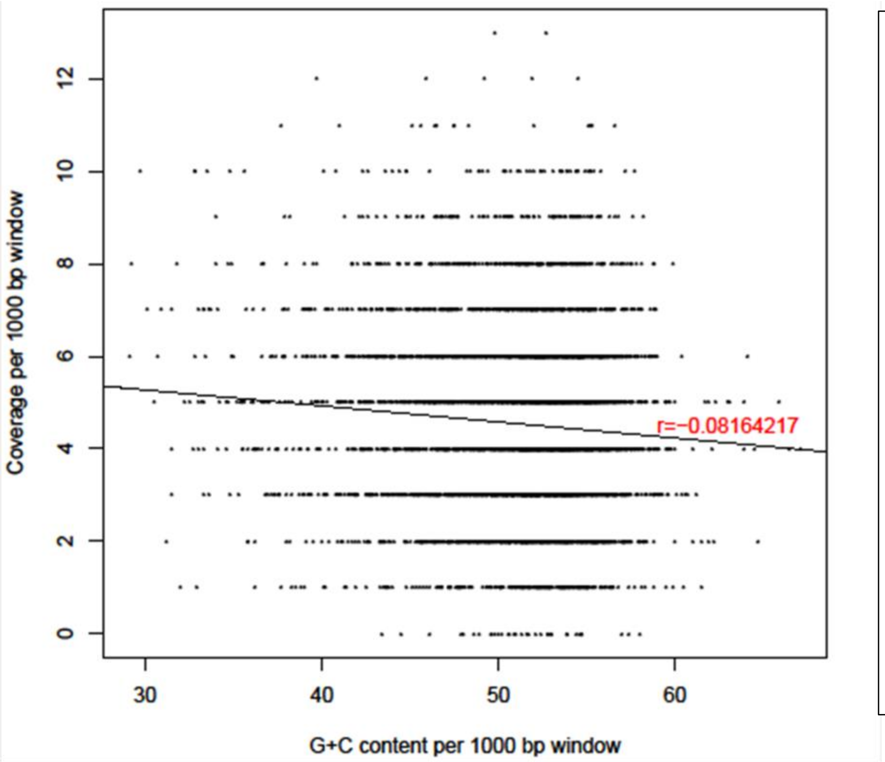
Supplementary Figure 5 G+C content versus coverage depth for E.coli

The *E. coli* genome was split into windows of 1000 bp across which the coverage depth was compared to the G+C content (r is the correlation coefficient).



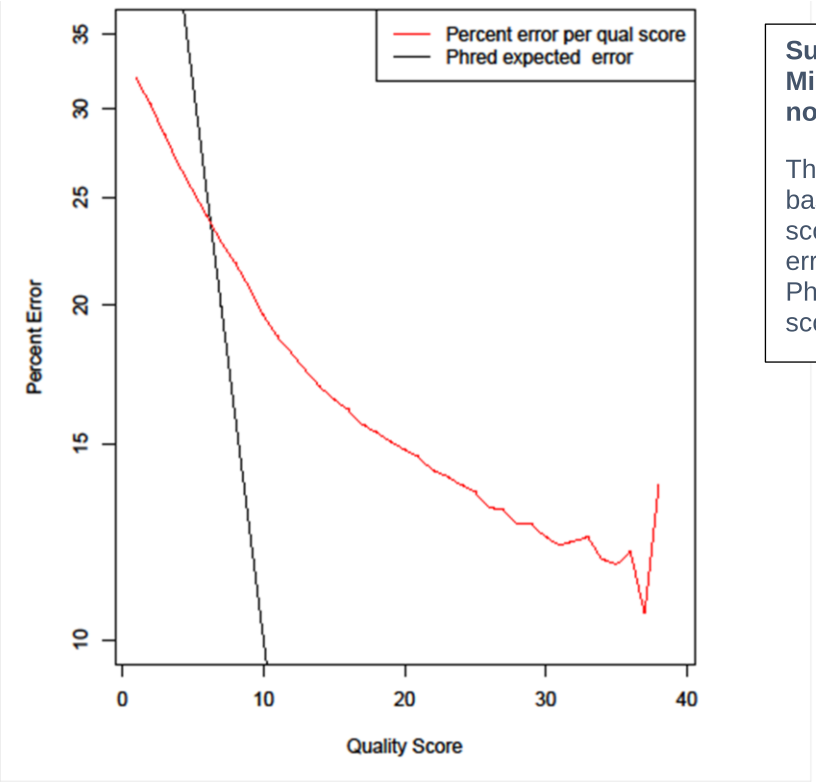
Supplementary Figure 6 G+C content versus coverage depth for Quick et al. Template reads

The *E. coli* genome was split into windows of 1000 bp across which the coverage depth for the Quick et al. Template reads was compared to the G+C content (r is the correlation coefficient).



Supplementary Figure 7 G+C content versus coverage depth for Quick et al. High Quality Two Direction reads

The *E. coli* genome was split into windows of 1000 bp across which the coverage depth for the Quick et al. High Quality Two Direction reads was compared to the G+C content (r is the correlation coefficient).



Supplementary Figure 8 MinION quality scores do not follow Phred

The likelihood of error for a base with a specific quality score does not match the error rate expected for a Phred value of that quality score

Chapter 7:

Discussion

This thesis represents a significant contribution to the field of genomics of bacterial pathogens, a field with potential to have a huge impact on human health and global food security. The aim was to explore the use of NGS as a tool for understanding and combatting bacterial pathogens of both humans and plants.

One of the most important groups of plant pathogens is the genus *Xanthomonas*, which is responsible for devastating infections such as bacterial leaf blight, common bacterial blight and black rot of crucifers. *Xanthomonas* species are a significant problem to global agriculture being responsible large scale crop losses worldwide. Underlining this is the presence of three *Xanthomonas* species in the top 10 bacterial pathogens of plants [1].

In the first two chapters of the thesis we used NGS to investigate the genomics of host adaptation and phenotypic convergence within the xanthomonads. We investigated the evolutionary history of *Xanthomonas* pathogens of bean and sugar cane uncovered evidence of recent horizontal gene transfer events associated with convergent evolution among phylogenetically distant strains sharing a common host.

The motivation for sequencing *Xanthomonas axonopodis* pv. *vasculorum* (*Xav*) was to elucidate its relationship with *X. vasicola* pv. *vasculorum*. Both these taxa were previously grouped together as a single taxon, namely *X. campestris* pv. *vasculorum*. Recent work (reviewed in Studholme *et al*. in press) demonstrated that they belong to two distinct species and have independently converged on life as a sugarcane pathogen. *Xav* NCPPB 900, isolated from sugarcane on Réunion island was sequenced, assembled, annotated and analysed. Multi-locus sequence analysis confirmed that *Xav* NCPPB900 fell within the *X. axonopodis* clade of *X. campestris* pv.

vasculorum. A genomic region was identified closely resembling that found in *X. vasicola* strains also known to infect sugar cane. This suggests that these genomic features have been transferred between *Xanthomonas* species. A T3SS gene cluster with a range of predicted effectors was also identified including TAL effectors which may contribute to the ability of *Xav* to colonise this host. It would be interesting to compare the T3SS effector profile of *Xanthomonas* sugar cane pathogens to assess to what extent these potential virulence factors have contributed to the phenotypic convergence of these phylogenetically distant pathogens. Our NCPPB 900 genome sequence has been cited by two subsequent papers, characterising and classifying novel strains of *Xanthomonas* [2,3].

To further investigate the genomics of host adaptation and phenotypic conversion we focused on two species of *Xanthomonas* bean pathogens: *X. axononpodis* pv. *Phaseoli* and *X. fuscans* subsp. *fuscans*. We sequenced and analysed 26 strains known to cause common bacterial blight on three important species of bean: common bean (*Phaseolus vulgaris*), Lima bean (*Phaseolus lunatus*) and lablab bean (Lablab purpureus). The genomic analysis of these strains uncovered a high degree of genetic variation within both taxa, including single nucleotide changes and variable gene content. The results of our analysis also suggested a recent acquisition of over 100 genes by *X. axonopodis* pv. *phaseoli* from *X. fuscans* subsp. *fuscans* which may have a role in the phenotypic convergence of these strains. Interestingly, the four strains isolated from Lablab bean were shown to represent a previously undescribed phylogenetically distinct genetic lineage closely related to *X. axonopodis* pv. *glycines*. These novel findings contribute to the knowledge of the causes of this devastating bacterial disease and provide markers which could prove useful in identifying outbreaks and contributing to the surveillance and tracking of the pathogen spread, helping to manage the disease.

The work presented in this chapter has been cited a number of papers since its publication, including several exploring the genomics of host specificity and adaptation [4,5,6] in xanthomonads and others investigating genomic and phenotypic diversity within the xanthomonads[7,8].

The second two chapters of the thesis introduce a related but distinct concept: the use of NGS to investigate and inform the surveillance and tracking of newly emerging plant and human pathogens. In chapter 4 we used NGS to discover an association between a newly identified strain of the human pathogen *C. jejuni* showing severe symptoms and acquisition of genes encoding a T6SS*.* The T6SS was recently discovered in bacteria and was predicted to be an important virulence factor similar to other secretion systems and their effectors and although it had been previously identified in the important food-borne pathogen *C. jejuni*. It was unknown if the T6SS was a common feature in *C. jejuni* strains and whether it was associated with more virulent forms of infection. This study surveyed all sequenced strains of *C. jejuni* for T6SS gene clusters. Evidence of a T6SS was indeed identified in a group of these strains. The *Hcp* gene was identified as a potential molecular marker for an intact T6SS and using this marker, it was shown that presence of the T6SS was significantly associated with the a more serious form of campylobacteriosis. Further to this it was found that the T6SS was significantly more prevalent in Asian isolates than in isolates from the UK. These findings will inform the surveillance of possible infectious *C. jejuni* strains during future import of chicken from the Far East.

Since its publication, the work from this chapter has had a significant impact - it has been cited by 22 papers Including several identifying further T6SS positive *C. jejuni* strains [9–12] and several further investigating the influence of the T6SS on virulence in *C. jejuni* [13–16]. Latterly, advances have even been made in the reduction of virulence of T6SS carrying *C. jejuni* [17,18] and most recently the *Hcp* gene has been used as a marker to survey the presence of the T6SS in *Helicobacter pullorum* [19] from chicken and suggesting that similar to the study presented here that the T6SS is indicative of a more virulent form of infection.

A further example of the value of rapid and cheap genome sequencing is in the characterisation of emerging pathogens. This is exemplified by previous studies such as the crowd-sourced analysis of an *Escherichia coli* outbreak in Germany in 2011 [20] and by our study on the mysterious Nyagatare strain that recently appeared in Rwanda, causing unusual symptoms on common bean. Genomic sequencing identified this strain as being quite unrelated to previously known bean pathogens and ultimately as a member of the species *X. cannabis*, which includes a range of pathogens, weakly pathogenic strains and non-pathogens. Unlike some members of the species, it appears to encode a potentially functional T3SS and virulence effectors. Further investigation of the effector complement of this pathogen could reveal insight into the adaptation of this pathogen to its host. Several unusual genomic features were identified including a 100 kb sequence with little or no similarity to other xanthomonads and a unique LPS synthesis gene cluster. These features could potentially be used molecular markers to track the spread of this pathogen and inform molecular diagnostics and detection. This study will aid epidemiological investigations of *Xanthomonas* outbreaks which have the potential to seriously impact bean crop production which is vital in many parts of the globe.

The work shown here provides important insight into the organisms concerned where Identification of horizontally acquired virulence-associated genes has applications in basic research. But, this work also demonstrates the utility of NGS to contribute to the investigation, detection and surveillance of emerging plant and human pathogens

Work from this chapter has been cited by several publications; contributing to both a review of the ecology, physiology and host specificity of *Xanthomonas* [5] and a paper concerning the evolution of pathogenicity in the *Xanthomonas* species [21]. The work in this chapter has also contributed to another publication concerning the characterisation of newly identified *Xanthomonas* pathogens of watercress [22]. Finally, this work has been built upon by a paper introducing two further strains which are shown to be in the same species level clade as the xanthomonad presented in this chapter. These newly sequenced strains are pathogenic on the cannabis plant and have no evidence of T3SS effectors, but MLSA analysis confirms they are closely phylogenetically related [23].

Finally, the bulk of the work presented in this thesis is based on data generated from NGS technologies. Whilst NGS offers a wealth of possibilities which are still being explored, chapter 6 of this thesis presents an assessment of the performance of the third generation sequencing platform the ONT MinION. This new technology offered portable, financially viable real time sequencing based on nanopore technology with advertised read lengths into the 100's of kilobases. The results presented are based on the pre-release access program and offer a comprehensive evaluation of the performance of the technology at an early stage of release. In order to assess the MinIONs performance a mixture of three bacteria with varying genome sizes and G + C content were sequenced. It was shown even at this early stage of development of both the MinION and the sequencing chemistry that read lengths of up to 100 kb could be generated. However, the error rate was shown to be ~38% but, even with the relatively high error rate it was still possible to align these reads to a reference genome and that there were single reads covering the entire rRNA operon which would have been unheard of prior to third generation sequencing. There did however appear to be limitations particularly in high G + C content sequences. Despite these limitations, the MinION presents a fantastic resource for bacterial genomics going forward. With the long reads generated from single molecule sequencing offering the possibility to unambiguously determine the sequence of repetitive regions such as TAL effectors and repeat regions which are known to be common in bacterial genomes. Further the portable nature of this technology makes it ideal for use in the field, investigating bacterial disease outbreaks and providing immediate data to use molecular analysis to inform epidemiology and track spread. Interest in this exciting new technology and the important nature of this work was subsequently demonstrated by the ~200 citations generated.

In conclusion, the highly cited work presented in this thesis has contributed important findings to the field of bacterial genomics. It reveals novel insights into the evolution of pathogenicity and potential molecular markers for taxonomic classification and epidemiological tracking of both human and plant pathogens. The methods presented here have been used subsequently several times to quickly characterise emerging bacterial disease outbreaks, track their spread and assess the threat posed. Examples of this can be found in both human and plant pathogens. This work has enormous potential to inform both human health and global food security. It also introduces and evaluates a novel technology for the exploration of bacterial genomics in the future.

References

1. Mansfield, J. *et al.* Top 10 plant pathogenic bacteria in molecular plant pathology. *Mol. Plant Pathol.* **13**, 614–629 (2012).

2. Studholme, D. J. *et al.* Transfer of *Xanthomonas campestris* pv. *arecae*, and *Xanthomonas campestris* pv. *musacearum* to *Xanthomonas vasicola* (Vauterin) *as Xanthomonas vasicola* pv. *arecae* comb. nov., and *Xanthomonas vasicola* pv. *musacearum* comb. nov. and description of *Xanthomonas* va. *bioRxiv* 571166 (2019).

3. Lang, J. M. *et al.* Detection and Characterization of *Xanthomonas vasicola* pv. *vasculorum* (Cobb 1894) comb. nov. Causing Bacterial Leaf Streak of Corn in the United States. *Phytopathology* **107**, 1312–1321 (2017).

4. Ruh, M., Briand, M., Bonneau, S., Jacques, M. A. & Chen, N. W. G. *Xanthomonas* adaptation to common bean is associated with horizontal transfers of genes encoding TAL effectors. *BMC Genomics* **18**, 1–18 (2017).

5. Jacques, M.-A. *et al.* Using Ecology, Physiology, and Genomics to Understand Host Specificity in *Xanthomonas*. *Annu. Rev. Phytopathol.* **54**, 163–187 (2016).

6. Rai, K. K., Rai, N. & Rai, S. P. Recent advancement in modern genomic tools for adaptation of *Lablab purpureus L* to biotic and abiotic stresses: present mechanisms and future adaptations. *Acta Physiol. Plant.* **40**, 1–29 (2018).

7. Midha, S. *et al.* Population genomic insights into variation and evolution of *Xanthomonas oryzae* pv. *oryzae*. *Sci. Rep.* **7**, 1–13 (2017).

8. Tugume, J. K., Tusiime, G., Sekamate, A. M., Buruchara, R. &

Mukankusi, C. M. Diversity and interaction of common bacterial blight disease-causing bacteria (*Xanthomonas* spp.) with *Phaseolus vulgaris L*. *Crop J.* **7**, 1–7 (2019).

9. Ungureanu, V. A. *et al.* Virulence of a T6SS *Campylobacter jejuni* chicken isolate from North Romania. *BMC Res. Notes* **12**, 1–7 (2019).

10. Siddiqui, F. *et al.* Molecular detection identified a type six secretion system in *Campylobacter jejuni* from various sources but not from human cases. *J. Appl. Microbiol.* **118**, 1191–1198 (2015).

11. Clark, C. G. *et al.* Comparison of genomes and proteomes of four whole genome-sequenced *Campylobacter jejuni* from different phylogenetic backgrounds. *PLoS One* **13**, 1–28 (2018).

12. Ugarte-Ruiz, M. *et al.* Prevalence of Type VI Secretion System in Spanish *Campylobacter jejuni* Isolates. *Zoonoses Public Health* **62**, 497–500 (2015).

13. Singh, A., Nisaa, K., Bhattacharyya, S. & Mallick, A. I. Immunogenicity and protective efficacy of mucosal delivery of recombinant hcp+ of *Campylobacter jejuni* Type VI secretion system (T6SS)in chickens. *Mol. Immunol.* **111**, 182–197 (2019).

14. Agnetti, J. *et al.* Clinical impact of the type VI secretion system on virulence of *Campylobacter* species during infection. *BMC Infect. Dis.* **19**, 237 (2019).

15. Corcionivoschi, N. *et al.* Virulence characteristics of hcp+ *Campylobacter jejuni* and *Campylobacter coli* isolates from retail chicken. *Gut Pathog.* **7**, 1–11 (2015).

16. Iglesias-Torrens, Y. *et al.* Population structure, antimicrobial resistance, and virulence-associated genes in *Campylobacter jejuni* isolated from

three ecological niches: Gastroenteritis patients, broilers, and wild birds. *Front. Microbiol.* **9**, 1–13 (2018).

17. Sima, F. *et al.* A novel natural antimicrobial can reduce the in vitro and in vivo pathogenicity of T6SS positive *Campylobacter jejuni* and *campylobacter coli* chicken isolates. *Front. Microbiol.* **9**, 1–11 (2018).

18. Bokhari, H. Exploitation of microbial forensics and nanotechnology for the monitoring of emerging pathogens. *Crit. Rev. Microbiol.* **44**, 504–521 (2018).

19. Javed, K. *et al.* Prevalence and role of Type six secretion system in pathogenesis of emerging zoonotic pathogen *Helicobacter pullorum* from retail poultry. *Avian Pathol.* 1–26 (2019).

20. Rohde, H. *et al.* Open-Source Genomic Analysis of Shiga-Toxin–Producing *E. coli* O104:H4. *N. Engl. J. Med.* **365**, 718–724 (2011).

21. Meline, V. *et al.* Role of the acquisition of a type 3 secretion system in the emergence of novel pathogenic strains of *Xanthomonas*. *Mol. Plant Pathol.* **20**, 33–50 (2019).

22. Vicente, J. G., Rothwell, S., Holub, E. B. & Studholme, D. J. Pathogenic, phenotypic and molecular characterisation of *xanthomonas nasturtii* sp. Nov. And xanthomonas floridensis sp. Nov., new species of xanthomonas associated with watercress production in Florida. *Int. J. Syst. Evol. Microbiol.* **67**, 3645–3654 (2017).

23. Jacobs, J. M., Pesce, C., Lefeuvre, P. & Koebnik, R. Comparative genomics of a cannabis pathogen reveals insight into the evolution of pathogenicity in *Xanthomonas*. *Frontiers in Plant Science* **6**, 431 (2015).

Appendix

This appendix contains a list of published papers based on work carried out by the authour during course of this project but no directly relevant to the work presented in this thesis.

1. Wagley, S. *et al.* Galleria mellonella as an infection model to investigate virulence of Vibrio parahaemolyticus. *Virulence* **9**, 197–207 (2017).
2. McDonagh, L. M., West, H., Harrison, J. W. & Stevens, J. R. Which mitochondrial gene (if any) is best for insect phylogenetics? *Insect Syst. Evol.* **47**, 245–266 (2016).
3. Harrison, J., Grant, M. R. & Studholme, D. J. Draft Genome Sequences of Two Strains of Xanthomonas arboricola pv. celebensis Isolated from Banana Plants . *Genome Announc.* **4**, e01705-15 (2016).
4. Harrison, J. & Studholme, D. J. Recently published Streptomyces genome sequences. *Microb. Biotechnol.* **7**, 373–380 (2014).
5. Harrison, J., Dornbusch, M. R., Samac, D. & Studholme, D. J. Draft Genome Sequence of Pseudomonas syringae pv. syringae ALF3 Isolated from Alfalfa . *Genome Announc.* **4**, 2015–2016 (2016).
6. Harrison, J. *et al.* A Draft Genome Sequence for Ensete ventricosum, the Drought-Tolerant “Tree Against Hunger”. *Agronomy* **4**, 13–33 (2014).
7. Quinn, L. *et al.* Genome-wide sequencing of Phytophthora lateralis reveals genetic variation among isolates from Lawson cypress (Chamaecyparis lawsoniana) in Northern Ireland. *FEMS Microbiol. Lett.* **344**, 179–185 (2013).