Comparative Bacterial Genomics

by

Nicholas James Loman

A thesis submitted to the
University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Biosciences
College of Life and Environmental Sciences
University of Birmingham
May 2012

For Anya Grace

Acknowledgements

I would like to thank all of my colleagues from the Schools of Biosciences and Immunity and Infection, particularly members of the Pallen/Penn research group, including Charles Penn for supervision and advice and Chrystala Constantinidou for sequencing. I would like to acknowledge all those who contributed to the studies presented here: Thomas Lewis, Pauline Jumaa, Debbie Mortiboy, Michael Hornsey, Lewis Bingle, Matthew Ellington, Jane Turton, Anthony Underwood, Tom Gaulton, Claire Thomas, Michel Doumith, David Livermore, Neil Woodford, Holger Rohde, Ruifu Yang, Martin Aepfelbacher, the scientists at the Beijing Genomics Institute, Rebecca Gladstone, Johanna Jefferies, Chrystala Constantinidou, Anna Tocheva, Leigh O'Connor, Jackie Chan, Saul Faust and Stuart Clarke. Special thanks are due to Brendan Wren at the London School of Hygiene and Tropical Medicine and George Weinstock at the Washington University Genomics Centre for funding and resourcing my first forays into high-throughput sequencing.

Thanks to my mum and dad for constant nagging about finishing my PhD and help proof-reading this manuscript. Thanks to Hannah for proof-reading this manuscript and pretty much everything else.

Finally, thanks to Mark Pallen for introducing me to microbial pathogenomics and providing mentorship, support and constructive abuse over the past 14 years.

Abstract

For the most part, diagnostic clinical microbiology still relies on 19th century ideas and techniques, particularly microscopy and laboratory culture. In this thesis I investigate the utility of a new approach, whole-genome sequencing (WGS), to tackle current issues in infectious disease. I present four studies. The first demonstrates the utility of WGS in a hospital outbreak of *Acinetobacter baumannii*. The second study uses WGS to examine the evolution of drug resistance following antibiotic treatment. I then explore the use of WGS prospectively during an international outbreak of food-borne *Escherichia coli* infection, which caused over 50 deaths. The final study compares the performance of benchtop sequencers applied to the genome of this outbreak strain and touches on the issue of whether WGS is ready for routine use by clinical and public health laboratories. In conclusion, through this programme of work, I provide ample evidence that whole-genome sequencing of bacterial pathogens has great potential in clinical and public health microbiology. However, a number of technical and logistical challenges have yet to be addressed before such approaches can become routine.

Contents

List of Papers

* Joint first authors

Chapter 1

Critical review

1.1 Introduction

1.1.1 The first golden age of microbiology

> "Progress in science depends upon new techniques, new discoveries and new ideas, probably in that order." – Sydney Brenner [1]

1.1.1.1 The microbial world

In the 1670s, when Antonie van Leeuwenhoek turned his home-made microscopes on water samples and dental plaque, he found himself staring at the wonderful and varied shapes of his "animalcules", now classified as protists and bacteria [2]. Wherever he looked, whether it was at urine or water, muscle tissue or seminal fluid, he uncovered a microscopic biological world unimagined at this time. His innovative use of the new technique of microscopy earned him the title of "founding father of microbiology" [3]. However, the impact of his discoveries on those dying of infectious diseases remained minimal for two hundred years.

1.1.1.2 Public health and vaccinology

In 1854, when John Snow removed the handle from a water pump on Broad Street, London, he helped stop an outbreak of cholera, but without really understanding its cause. Although there had been suggestions that microbes might cause dis-

ease (by Fracastoro and Bassi, amongst others [4]), the generally accepted view was that miasma, or bad air, was the cause of transmissable diseases. Snow's intervention was informed through his pioneering use of a modern-style epidemiology study. He observed and carefully mappped clusters of cases of cholera in households which drew their water from the Broad Street pump [5]. In removing the handle and ending the outbreak, he provided strong evidence that it was water, not air transmitting the cholera. The very same year, Filippo Pacini, after studying a large cholera outbreak in Padua published his findings of a waterborne comma-shaped bacillus he called *vibrio*, identifying it both as the "specific" cause of cholera as well as demonstrating that it was contagious. However, he was ignored by the scientific community, with his contribution recognised only posthumously [6].

1.1.1.3 The birth of medical microbiology

Hansen made the link between rod-shaped bodies in lepromatous nodules and leprosy in 1873 but, like Pacini, received little initial support for his theory [7]. Instead, Robert Koch did most to popularise the germ theory of disease. Koch made three major discoveries linking microbes to important human diseases: first, the link between a sporulating bacillus and anthrax; second, that the tubercule bacillus caused tuberculosis; and, third, he rediscovered a vibrio as the cause of cholera, unaware of Pacini's earlier work. Koch, working with Henle forged a conceptual framework for evaluating the link between microbes and disease by the formulation of a set of "postulates", which could be used to assess whether a particular agent caused a given disease. Koch, with fellow German Ferdinand Cohn, also pioneered the growth of pure cultures of bacteria on solid media, a line of work which eventually culminated in the present-day use of agar.

In parallel with Koch's studies, Louis Pasteur took on the theory of spontaneous generation as an explanation of the origin of microbial life. Pasteur won a competition sponsored by the French Academy of Sciences, aimed at proving or disproving spontaneous generation with his elegantly effective flask experiment. Sadly, space constraints do not permit a detailed description of Pasteur's many other major achievements, which include his heat-killing treatment ("pasteurisa-

tion") and the development of several vaccines effective against bacterial disease, including chicken cholera.

Pasteur and Koch remain the fathers of medical microbiology, with ideas that continue to influence and inform modern microbiology. At the turn of the 19th century, microbiologists, armed with microscopy, Koch's postulates and pure culture techniques entered the "first golden age of microbiology" [8]. Perkins remarked that "discovery of the principles... resulted in such a sudden burst of investigation that it was a lost month in which a new organism was not described, catalogued, and laid away". An astonishing flurry of discoveries followed, in which the causes of most significant bacterial diseases were determined within a twenty-year period [8]. These discoveries included Theodor Escherich's discovery of *Bacillus coli*, now named *Escherichia coli* and Frankel's discovery of the pneumococcus, subsequently classified as *Streptococcus pneumoniae*.

Concepts	
	Bacterial physiology
	Methods for cultivation and observation of bacteria
	Isolation of bacteria in pure culture
	Bacterial nutrition
	Bacterial classification based on phenotypes
	Medical microbiology
	Germ theory of disease
	Viruses
Applications	
	Clinical identification of microbes
	Antimicrobial chemotherapy
	Vaccines
	Industrial fermentation

Table 1.1: The first golden age of microbiology (adapted from Moloy [8])

1.1.1.4 Bacterial classification

Humans have a fundamental desire to classify things. Bacterial classification rapidly became an obsession for the first generations of medical and environmental microbiologists, aided by an ever-growing battery of tests and features, including morphological characteristics, growth under different conditions and ability to

degrade particular substances (Table 1.2).

In 1872, Ferdinand Cohn proposed a basic taxonomy of microbial life based on morphological criteria, dividing microorganisms into four "tribes" and six genera (*Micrococcus*, *Bacterium*, *Bacillus*, *Vibrio*, *Spirillum* and *Spirochaeta*) [9].

In medical microbiology, discriminatory tests which could distinguish pathogenic strains from harmless ones were highly valued, but this led to what would later be seen as highly unnatural classifications, often based on a single feature. Tests such as Methyl Red were used to differentiate members of the *Enterobacteriaceae*, e.g. *E. coli* from *Enterobacter*. Urea hydrolysis was used to discriminate between *E. coli* and *Proteus* in urinary tract infections. The most important test was Gram's staining method, which we now know divides organisms based on the presence of peptidoglycan content in bacterial cell walls [10]. The Gram stain is often used as the first stage of identification using *dichotomous keys*, a decision flow-chart method which follows a series of Boolean (yes/no) choices leading to a confident identification at the last stage. These keys were, and still are, used extensively in medical microbiology [11]. These early painstaking efforts on the classification of bacteria culminated in the first edition of Bergey's Manual [12], published in 1923.

1.1.1.5 Numerical taxonomy

By the 1960s, the sheer number of observable phenotypes led Robert Sokal and Peter Sneath to propose a system of what they called "numerical taxonomy" [13]. This system, based on methods pioneered by Michel Adanson in the early 19th century, tabulated the results of tests, or "features", against bacterial isolates. The method had several innovative elements. One was that each feature should be considered with equal weight (with care taken not to introduce redundant tests). Importantly, classification of isolates based on the feature table could be carried out by computational methods, allowing the system to be used on large numbers of specimens. Clustering algorithms, such as the neighbour-joining method were applied to the data. By setting appropriate similarity cut-offs, this system gave backwards compatibility with the existing taxonomy and provided a frame-work for new biological insights. Sneath was prescient in predicting that features may

Name	Method
Microscopic morphology	Cell shape, size, colour
Macroscopic morphology	Appearance of colonies
Staining	Examples: Gram's method, Acid-fast reaction
Biochemical assay	Catalase/oxidase, sugar fermentation
Analytical Profile Index (API)	Commercialised, miniaturised test panels. Phenotypic reactions case colour change. Resulting pattern looked up in reference book.
Vitek-2 (bioMérieux)	Plastic card with panel of biochemical tests
Enzyme-linked immunosorbent assay (ELISA)	Solid-phase enzyme immunoassay

Table 1.2: Tests for identification of bacteria in clinical microbiology

be sourced in the future from molecular data, stating "it may be possible in the future to re-define it in terms of genes and perhaps nucleotides; this will not effect the basic concepts of Adansonian methods but may simplify them." [14]

1.1.2 The second golden age of microbiology

1.1.2.1 Genetics and evolution

Darwin corresponded with Cohn and knew of the work of Pasteur. However, his theory of evolution had little impact on microbiological thinking for most of the 19th and 20th centuries. The idea that all life, including bacteria were descended from a common ancestor gained little traction in a medically-dominated anthropocentric viewpoint, in which bacterial species were regarded as fixed entities, often with the sole purpose of causing disease of humans [15].

In the mid-20th century, the "modern synthesis" of evolution and genetics brought together ideas from Mendel and Darwin, resulting in a conceptual framework for understanding and testing evolutionary theory in terms of genes. Theodosius Dobzhansky placed Darwin's ideas in the language of genetics, defining evolution as "a change in the frequency of an allele within a gene pool" and later famously stating that "nothing in biology makes sense except in the light of evolution" [16].

1.1.2.2 Molecular biology

Experimental microbiology had a key role in the birth of molecular biology. Deoxyribonucleic acid (DNA) was isolated for the first time by Miescher in 1868 from surgical pus. The identification of DNA as the hereditary substance was due to two key experiments. First, Griffith showed that the *S. pneumoniae* could be transformed from a rough to a smooth phenotype by the addition of killed cells of smooth phenotype [17]. After exhaustively purifying nucleic acid from the killed cells, Avery subsequently showed it was only this molecule which could cause transformation [18]. A failed attempt by the prolific Linus Pauling to determine the structure of DNA preceded Watson and Crick's double helix structure, which demonstrated elegantly the chemical basis of DNA replication [19].

Following this discovery, many crucial secrets of life were uncovered: one was Crick's "central dogma"–"DNA makes RNA makes protein", with the flow of genetic information in one direction. The genetic code was revealed to be a triple nucleotide system and after much trial-and-error, Nirenberg and Gamow's "RNA tie club" assigned an amino acid or function to each of the possible 64 codons.

The most visible product of the second golden age of microbiology was the development of molecular cloning, harnessing bacterial gene expression and protein synthesis to the needs of biotechnology. The discovery of restriction endonucleases, able to cut DNA at specific sequences, twinned with the ability to join fragments with DNA ligase meant that recombinant DNA molecules could be created within plasmid vectors and then transformed into *E. coli* [20–24].

1.1.2.3 Sequencing

Fred Sanger earned his first Nobel Prize in Chemistry for determining the amino acid sequence of insulin. The first genome to be sequenced was from the RNA virus bacteriophage MS2 in 1976 [25]. Soon afterwards, three methods of DNA sequencing were invented in quick succession; Maxam and Gilbert's method [26], Sanger and Coulson's "plus-minus" method [27] and the chain-termination method now commonly referred to as "Sanger sequencing" [28]. Sanger sequencing employs a chain termination method using di-deoxynucleotide triphosphates (ddNTPs), which prevent extension of nascent chains of DNA. By carrying out

Concepts
Bacterial genetics
DNA as genetic material and its structure
Genetic code
Mechanism of gene expression
Regulation of gene expression
Transposons
Bacterial physiology
Membrane transport and electrochemical gradients
Cellular immunology
Applications
Genetic engineering
Nucleic acid and protein sequencing
Microbial classification based upon genotypes
Monoclonal antibodies

Table 1.3: The second golden age of microbiology (adapted from [8])

four separate reactions, each with only one of the four ddNTPs added, and running the products on a polyacrylamide gel, the sequence of bases can be read. Sanger's method used radio-labelling for detection, but now tagging with a fluorescent dye is most commonly used. Plus-minus sequencing was used to sequence the genome of the DNA phage ϕX174. However, the chain-termination method soon proved the quickest and easiest of the three methods and permitted the sequencing of several landmark genomes – the entire chromosome of human mitochondrial DNA (16.6 kilobase pairs) and bacteriophage λ (49 kb).

1.1.2.4 Molecules as documents of evolutionary history

Comparisons between nucleotide or amino acid sequences of homologous molecules (those sharing a common ancestor) remains the cornerstone of molecular phylogenetics, an approach which has breathed fresh life into Darwin's idea of common descent. Before DNA or even peptide sequences became readily available, Zuckerkandl and Pauling proposed that the information locked in these molecules would enable the construction of molecular phylogenies, derived from comparisons of homologous sequences from different species [29]. They realised, given that the genetic code was degenerate (more than one codons often coding for the

same amino acid), that "isosemantic changes" mean that nucleotide sequences have a higher information content than protein sequences and thus proves a better source of phylogenetic information. Furthermore, they speculated that it might be possible to partition sequence changes into those that had undergone selection and those that had not.

The Luria and Delbrück experiment showed that mutations arose in the absence of selection, rather than as a response to selection [30]. Kimura subsequently proposed that the majority of nucleotide changes were neutral, occurring through genetic drift and that only a few were fixed by positive selection. This suggested the existence of a "molecular clock" permitting measurement of evolutionary distances simply by counting the number of mutations seen between pairs of species. Clustering algorithms from numerical taxonomy, such as neighbour-joining methods [31], permitted the phylogenetic reconstruction of evolutionary history in the form of phylogenetic trees.

Woese showed the ultimate power of these new methods by analysing sequences from the small ribosomal DNA subunit, universal in both bacteria and eukaryotes. Woese made a remarkable discovery; by analysing ribonuclease digestion patterns from 16S rDNA, he found that certain prokaryotes were actually as closely related to eukaryotes as they were to bacteria. These outliers were often "extremophiles", able to withstand extremes of heat, pH or salinity, suggesting they may have been amongst the earliest forms of life. Woese therefore named them the "archaebacteria" (now called "archaea") and proposed that they made up one of three divisions, along with with bacteria and eukaryotes, in a universal tree of life [32–34].

Molecular studies have shown that the classical bacterial taxonomy is often in conflict with phylogenetic data. An example is the taxonomic classification of *E. coli* and *Shigella*. In medical classifications, shigellosis is always caused by the Shiga-toxin producing *Shigella* and is distinct from enterohaemorrhagic disease caused by *E. coli* O157:H7 (classical EHEC). Molecular phylogenetic analysis has revealed that in fact *Shigella* is a member of the B2 phylogroup of *E. coli* [35, 36], making it more closely related to certain *E. coli* strains than some other *E. coli* are from each other. Therefore, in terms of molecular taxonomy *Shigella* is an *E. coli* (or *E. coli* are *Shigellae*). There are examples in other genera, such as

Streptococcus and *Neisseria* where such extensive recombination has occurred to make species boundaries blurred or even meaningless.

1.1.3 Bacterial genomics

The publication of the complete genome sequence of *Haemophilus influenzae* in 1995 ushered in the era of bacterial genomics [37]. The 1.83-megabase chromosome was sequenced at an estimated cost of $0.48 per finished base-pair, giving a total cost of around $900,000 [38].

The process of whole-genome shotgun sequencing pioneered in this study began with the shearing of genomic DNA into short fragments. These fragments were then cloned into plasmid vectors and expressed in *E. coli* to amplify them. This "clone library" was grown on a solid medium, with individual colonies picked for sequencing on capillary sequencing machines, which automate the Sanger sequencing method. This method proved highly successful and became the standard method for sequencing bacterial genomes and was later was used to sequence larger genomes of model organisms: yeast, fruit fly and *Homo sapiens* [39–41].

By 2000, whole-genome sequencing had yielded complete, published sequences for over two dozen biologically and medically important microbial species including *Helicobacter pylori*, *E. coli* K-12, *M. tuberculosis* and *Bacillus subtilis* [42]. Technological and logistical innovations such as library construction using robots, sequencing instruments with increased capacities and the scaling up of workflows in large sequencing centres, such as the Sanger Centre and the Institute for Genomic Research (TIGR), meant that, by 2005, at least one complete genome sequence was available for most bacterial species or pathovars associated with human disease. The availability of multiple strains of the same species led to the first comparative genomics projects, which included a comparison between two strains of H. pylori [43] and pair-wise whole-genome comparisons between *M. tuberculosis* H37Rv, a commonly-used laboratory strain and the "Oshkosh" outbreak strain CDC-1551 [44].

An early "translational" (from the laboratory to the clinic) use of genome sequencing was "reverse vaccinology": an approach to the discovery of vaccine targets that relies on screening whole-genome sequences for potential protective

antigens which are then followed up experimentally [45, 46]. This approach has recently culminated in the creation of an effective vaccine against the meningococcus [47].

1.1.3.1 Bacterial genome dynamics

If genes provide documents of evolutionary history, bacterial genome sequences provide phylogenetic encyclopedias. When compared to related strains, genome sequences often reflect changes in lifestyle or adaptations to particular niches. A notable finding from comparative genomics studies are examples of extreme genome reduction. For example, species of *Buchnera*, phylogenetically closely related to *E. coli*, have genomes a fraction of its size [48, 49], having shed many cellular functions on adopting the endosymbiotic lifestyle. Similarly *Mycobacterium leprae* diverged from the *M. tuberculosis* complex 36-66 million years ago [50, 51]. Since then, it has lost over half of its protein-coding potential through mutations which render genes non-functional, a process termed pseudogenisation, accompanied by a reduction in genome size and a narrowing of its niche and host range: *M. leprae* can only grow in humans, the nine-banded armadillo and the mouse footpad [52, 53]. Recently, a new leprosy-causing species, *M. lepromatosis* was discovered: phylogenetic analysis suggests these two species diverged approximately 10 million years ago [54, 55].

1.1.3.2 Bacterial clonality

Molecular typing methods and genome sequencing have shed light on the population genetics of bacterial species. Spratt used the results of multi-locus enzyme electrophoresis to estimate the rate of change within the genome of a bacterial species [56]. He recognised that clonality might be disrupted by the action of recombination to re-organise the genome or replace segments of the genome from one lineage with those from another. Spratt determined that certain species are highly monomorphic, for example certain pathovars of *Salmonellae*, *M. leprae*, *Y. pestis* and *B. anthracis* [57]. Phylogenetic analysis of such important human pathogens is complicated by the lack of variation; sequencing much less than the whole-genome will not provide sufficient information for the purposes

Monomorphic
Mycobacterium leprae
Mycobacterium tuberculosis
Salmonella enterica serovar Typhimurium
Bacillus anthracis
Yersinia pestis
Intermediate
Acinetobacter baumannii
Escherichia coli
Polymorphic
Streptococcus spp.
Neisseria spp.
Haemophilus influenzae

Table 1.4: Genetically monomorphic and polymorphic pathogens

of typing and epidemiology. These strains are in contrast to strains with high genomic plasticity, where significant gene loss and gain as well as chromosomal rearrangements are seen as a result of recombination. Notable examples are in the ϵ-protobacteria such as *Helicobacter* and *Campylobacter*, *Neisseriaceae* and *Streptococcus* (Table 1.4). The extent of recombination within ϵ-protobacteria is so extreme that inter-species comparisons often reveal a total breakdown of genome synteny (co-linearity along the chromosome) [61, 62].

1.1.4 Clinical microbiology in the 21st century

1.1.4.1 The practice of clinical microbiology

Today, clinical microbiology remains firmly rooted in 19th century techniques, still relying on microscopy and culture to detect and identify potential pathogens. Once obtained in pure culture, identification is made possible by a battery of specific phenotypic assays. Antibiotic sensitivity is assayed by assessing growth in the presence of antimicrobial agents. While the process of performing multiple biochemical tests is faciliated by commercially available semi-automated systems such as Analytical Profile Index (API) and Vitek-2, the principles of diagnostic

microbiology have changed little in over a century.

1.1.4.2 The threat of antibiotic resistance

> "It's time to close the book on infectious diseases, declare the war against pestilence won, and shift national resources to such chronic problems as cancer and heart disease." – William H. Stewart, US Surgeon General

The number dying from infectious diseases has fallen steadily during the 20th century [63]. This is largely due to sanitation, vaccination programmes and antibiotic therapies, as well as improved nutrition [64]. However, the Surgeon General was wrong to think that the war against infection would be nearly over in 1970. Infection is still a leading cause of death world-wide, with an estimated one-third of the world's population infected by tuberculosis [65].

For almost every class of clinically useful antibiotic, antibiotic-resistant strains have been observed within a few years or at most decades after first clinical use [66]. The emergence of antibiotic resistance in microbes poses a major threat to our ability to treat infectious disease [67]. Numerous antibiotic-resistant "superbugs" have attracted attention, including meticillin-resistant *Staphylococcus aureus* (MRSA), vancomycin-resistant *Enterococci* and multidrug-resistant *Pseudomonas aeruginosa*. Space does not present a full description of these organisms. Of particular concern is the emergence of multi-drug resistant Gram-negative bacteria including *Acinetobacter baumannii*, [68].

Antibiotic resistance can arise via a number of molecular mechanisms: mutations affecting target sites of the drug (e.g. *rpoB* bypass mechanisms, mutations conferring resistance to rifampicin in *M. tuberculosis*), antibiotic inactivating or modification enzymes such as β-lactamases, changes in envelope permeability and non-specific systems such as drug efflux pumps. Antibiotic-resistant determinants are commonly found in plasmids which may be transferred between species, these include extended-spectrum β-lactamases (ESBLs) and *Klebsiella pneumoniae* carbapenemases (KPCs). The recent discovery of a novel metallo-β-lactamase (NDM-1) which confers carbapenem resistance has been seen in association with *Klebsiella pneumoniae* and *E. coli* infections [69]. This class is

of particular cause for concern as carbapenems are used as "last-resort" options against ESBL-producing strains [70]. A "post-antibiotic apocalypse" looms with some infections potentially becoming resistant to all known antibiotics. Pan-drug resistant strains of *A. baumannii* and eXtremely-drug resistant (XDR) and totally drug-resistant strains of *M. tuberculosis* have been isolated from patients [71–73]. It is also of concern that there are currently very few antibiotic candidates in the commercial drug-discovery pipeline [74].

1.1.4.3 Bacterial epidemiology and bacterial typing

The aim of bacterial typing is to distinguish between strains within the same species. Some typing schemes work with a number of species, whereas some may be designed for a particular species or subspecies. An ideal bacterial typing scheme would have a number of desirable properties, including speed, low-cost, portability (comparable between laboratories [75]) and reproducibility. However, in the real world, we find a plethora of less-than-ideal schemes. In fact, so many typing schemes have since been proposed that Mark Achtman once proposed the term "YATM"' (Yet Another Typing Method) as a light-hearted response to the number of such schemes being published in the *Journal of Clinical Microbiology* [76].

Although often seen as an arcane adjunct to diagnostic microbiology, epidemiological typing can, if done quickly enough, have an impact on real-world problems by revealing modes of spread of pathogens and informing choice of intervention strategies (e.g. isolate and decontaminate infectious patients, more thorough environmental cleaning, improved hand hygiene, better antibiotic stewardship, removing of environmental source).

Many of the molecular typing methods listed in Table 1.5 target sites in the genome with a high mutation rate, for example microsatellite repeats (VNTR) and repetitive regions (rep-PCR). This gives the advantage of ready discrimination between clones and they have been used with much success in bacterial epidemiology. However many of these methods, however discriminatory, do not provide information useful for phylogenetic reconstructions, making relatedness

Name	Region of genome considered	Method employed
Multilocus enzyme electrophoresis (MLEE) [77]	Whole genome	Gel electrophoresis
Multilocus VNTR analysis (MLVA) [78]	Microsatellite/tandem repeats	analysis of PCR fragment size
Multilocus sequence typing (MLST) [75]	Conserved housekeeping genes	PCR and sequencing
Pulse-field gel electrophoresis (PFGE) [79–81]	genome-wide restriction sites	restriction digest and gel electrophoresis
PCR/multiplex PCR [82]	Specific genomic loci	PCR and optional sequencing
rep-PCR [83]	repetitive elements, outward facing primers	PCR and gel electrophoresis

Table 1.5: Examples of molecular techniques for bacterial typing

between isolates with different profiles hard to assess [84]. Techniques such as MLEE and PFGE, which rely on images produced by gel electrophoresis, are not easily portable between laboratories.

Multilocus sequence typing (MLST) schemes rely on sequencing a number of conserved "house-keeping" genes to generate a profile. Each unique sequence is given an allele number via an on-line database. Each unique combination of alleles gives a "sequence-type" (ST). An appropriate set of genes must be identified for each species under consideration. The success of this scheme relies on choosing genes that are found in all members of the species and evenly spaced around the chromosome. Those wishing to share MLST data therefore must agree on a suitable scheme and use the same set of primers. Multiple schemes may exist for the same species, there are three competing primer sets for *E. coli* and two for *A. baumannii*, creating potential for confusion [85–87]. MLST has proven a highly versatile approach, with schemes available for over fifty taxa. It has been particularly useful in understanding the population structure of recombinogenic species such as *H. influenzae* and the pathogenic *Neisseria*. However, MLST does not work well for genetically monomorphic species such as *M. tuberculosis*, where schemes such as MIRU-VNTR are more discriminatory [88, 89]. Additionally, MLST schemes permit only a limited view of phylogenetic relatedness through cluster analysis of single- or double-locus variants (profiles which differ by one or two alleles and are assumed to be related).

1.1.5 High-throughput sequencing

The era of high-throughput sequencing began with the release of 454 Life Science's GS20 instrument in 2005. Its successor, the GS FLX, was able to produce 200 megabases of sequence each run, enough to sequence several isolates of *E. coli* for around $10,000 in sequencing reagents. In 2008, this technology was used to sequence James Watson's genome, taking just two months [90, 91] at less than 1% of the cost of the original $3bn human genome project. The first-generation Solexa Genetic Analyzer produced a gigabase of sequence data when it debuted in late 2006 [92, 93]. Since then, sequencing throughput has exhibited a hyper-Moore's law increase in throughput, with a reciprocal reduction in costs. As of writing the highest-throughput instrument, the Illumina HiSeq 2500 looks set to be soon able to generate a terabase (1000 gigabases) of sequence data per run. There is no sign of this progress slowing. Table 1.6 summarises currently available high-throughput instruments.

The first generation of high-throughput sequencing technologies differed from traditional Sanger sequencing in a number of important ways. Firstly, amplification of sample relied on the production of "molecular colonies" of clonal DNA template, without the need for cloning into a biological vector and subsequent expression in *E. coli*. These colonies are amplified on beads (454, Ion Torrent) or on a solid-surface (Solexa) and are sequenced in a massively-parallel fashion, between a million and a thousand million at a time, depending on the instrument. The process of reading nucleotides may be light-based: laser-excitation of fluorescently labelled nucleotides (Solexa) or release of photons through the action of luciferase during nucleotide incorporation (454). The Ion Torrent instrument relies on the detection of protons released during nucleotide incorporation. This takes place on a modified silicon chip functioning as a massively-parallel pH meter. Space does not provide a fuller description of the technologies but Metzker provides a comprehensive snapshot of the situation in 2009 [94].

1.1.5.1 Bioinformatics analysis of high-throughput sequencing data

Analysis of molecular sequence data relies on the process of alignment between pairs of sequences. High-scoring alignments suggest the presence of homology and

Technology	Year	Amplification method	Sequencing method	Ref
454 (Roche)	2005	Emulsion PCR on beads	SBS (flow), fluorescence detection	[90]
Solexa (Illumina)	2006	Bridge amplification on solid surface	SBS (reversible blocking)	[93]
SOLiD (Life Technologies)	2008	Emulsion PCR on beads	Sequencing by oligonucleotide ligation and detection	[95]
Helicos	2009	Amplification-free	Single molecule fluorescent sequencing	[96]
Pacific Biosciences	2010	Amplification-free	Monitoring of individual DNA polymerase molecules in zero-mode waveguide detectors	[97]
Ion Torrent (Life Technologies)	2011	Emulsion PCR on beads	SBS (flow), detection of H^+ ions on silicon chip	[98]

Table 1.6: High-throughput sequencing platforms and their year of introduction. SBS: sequencing-by-synthesis.

permit calculation of sequence similarity to be made. The Smith-Waterman and Needleman-Wunsch methods are well-established as "gold-standard" algorithms for global and local alignments respectively [99, 100]. However, when faced with the challenge of aligning millions of reads produced by high-throughput sequencing instruments to a reference genome, these algorithms were found to be too computationally expensive to be of practical use [101]. New aligners, optimised for high-throughput sequencing experiments have been designed for large numbers of short reads.

One of the first short-read aligners was Heng Li's MAQ [102] which was extremely fast, but had drawbacks, particularly an inability to align individual reads across insertions or deletions ("indels"). The BWA and Bowtie short-read aligners subsequently gained popularity due to their speed, thanks to an optimised indexing technique called the Burrows-Wheeler transform [103]. Many of the original algorithms traded sensitivity for speed, to the extent where alignments were often unreliable. Improvements such as BWA-SW, SSAHA2 and Novoalign incorporated a fast "seed" step coupled with the slower, more accurate Smith-Waterman or Needleman-Wunsch algorithms to generate more reliable output [104, 105]. There is now such a variety of short-read aligners that, echoing Acht-

man's YATM, a recently published alignment program was named YOABS (Yet Other Aligner of Biological Sequences) [106]!

Short-read aligners are deployed in re-sequencing projects, where a high-quality reference sequence serves as template. When no reference is available, or when an unbiased method is needed, genome assembly software can be used to attempt to reconstruct genome sequences *de novo*. Initially, *de novo* assembly with reads as short as 20-30 bases was thought to be impossible, as existing methods, such as the overlap-layout-consensus algorithm, which worked with long capillary reads and also with 454 sequencing data, did not work for short-read sequencing data. However, development of new assembly methods permitted useful assemblies to be generated from these data, albeit with large numbers of sequence "gaps", where repetitive sequences were encountered [107]. The most successful *de novo* assembly software now work by constructing *de Bruijn* graphs of overlapping k-mers (short sequence words). Examples of commonly used software packages include Velvet, SOAPdenovo and ABYSS [108–110].

Once whole-genome assemblies have been generated, "downstream" analysis often involves annotation of sequences. Typically this involves an initial stage, where coding sequences are predicted (using software such as Glimmer or GENEMARK, or through homology searches using BLAST [111–113]) and detection of stable RNA species (tRNAScan-SE, RNAmmer [114, 115]). Subsequently coding sequences are assigned a tentative function through homology searches of existing annotation databases, such as the National Center for Biotechnology Information's non-redundant protein database [116]. This process can be performed by automated annotation pipelines such as this author's xBASE-NG [117].

Once annotated, whole-genome assemblies can be viewed through software such as Artemis, or compared to another genome using Artemis Comparison Tool. Multiple whole-genome aligners can build an alignment from many genomes. These alignments can then be used to build whole-genome phylogenies, or to analyse the core and pan-genome of a species.

The choice of whether to analyse data through a re-sequencing approach or *de novo* depends on a number of factors listed in Table 1.7.

Resequencing approach	
	Closely-related reference genome available
	Detection of single nucleotide polymorphisms (SNPs) Detection of small indels
	Detection of sequence absent from the newly-sequenced strain
***de novo* assembly approach**	
	No reference sequence or divergent reference sequence
	Detection of novel genes or sequence in the newly-sequenced strain
	Detection of large-scale genomic rearrangements

Table 1.7: Factors determining choice of sequencing analysis

1.1.5.2 Genomic epidemiology

Whole-genome sequencing has been rapidly adopted as a research tool for molecular evolution studies (Table 1.8).

Scale	Organism	Notes	Ref
Worldwide	*S. aureus* (ST239)	WGS of a historical strain collection demonstrated evolution of this drug-resistant sequence type over four decades. Also demonstrated fine-grained discrimination of isolates from different wards of a Thai hospital.	[118]
	S. pneumoniae (PMEN-1)	WGS identified multiple separate events leading to antibiotic resistance.	[119]
	M. leprae	Confirmed monomorphic nature of leprosy pathogen and demonstrated association of SNPs with early human migrations and trade routes.	[120]
	C. difficile	WGS of diverse clostridial strains reveals a complex population structure.	[121, 122]
	V. cholerae	WGS of isolates from the Haiti epidemic suggest cholera was imported during the emergency aid response; three waves of cholera in seventh pandemic.	[123, 124]
Nationwide	*S. pyogenes*	Investigated three epidemics in Ontario, Canada; identified clonal cluster of invasive infections in San Francisco	[125, 126]
	L. monocytogenes	Demonstrated three distinct strains were involved in nationwide outbreak of listeriosis food poisoning.	[127]
	Salmonella enterica serovar Montevideo	Forensic whole-genome analysis helped trace the origin of an outbreak of food-poisoning. WGS of strains in contaminated food were phylogenetically closely related to those found in a meat processing factory in New England.	[128]
	B. anthracis	Amerithrax investigation helped link anthrax spores sent in the US postal system to a government researcher. Due to the monomorphic nature of this pathogen, forensic investigations relied on presence of low-frequency colonial morphotypes in the samples.	[129]
	S. pneumoniae	Genome sequencing demonstrated five instances of vaccine escape recombination in serotype 19A strains.	[130]
Community	*M. tuberculosis*	Network-based analysis of putative transmission events during an outbreak of tuberculosis in Canada.	[128]

Table 1.8: Notable studies in bacterial whole-genome epidemiology

The use of sequencing in genomic epidemiology was pioneered on viruses. One high-profile example of using phylogenetic recontructions in tracing human-to-human spread of a pathogen was an investigation of patients who contracted HIV without obvious risk factors. Epidemiological analysis revealed they had the same dentist. The dentist, who was infected by HIV was implicated as the likely source. Sequencing of *gp120* and phylogenetic analysis suggested that the virus from the dentist were closely related to the viruses of infected patients [131]. In bacteria, whole-genome sequencing was used to great effect to show that the culture of *Bacillus anthracis*, sent in the US mail to prominent senators and journalists, belonged to a common laboratory strain, the Ames strain. Whole-genome sequencing, rather than conventional molecular typing was required in this case due to the highly genetically monomorphic nature of this pathogen [129].

1.2 Present work

1.2.1 Aim of the studies

I present five studies, which explore the potential of high-throughput sequencing in clinical microbiology. These studies spring from several vantage points:

- from an infection control standpoint, looking at transmission chains in a hospital outbreak.
- from the viewpoint of a clinical microbiologist, looking at the impact of antimicrobial therapy on bacteria in a single patient.
- from a pubic health perspective, looking at a colonisation and infection within a local human population.
- from an international perspective during a sudden, serious, large outbreak.
- from the perspective of microbiology laboratory staff faced with a choice of novel technologies and instruments.

1.2.1.1 High-throughput whole-genome sequencing to dissect the epidemiology of *Acinetobacter baumannii* isolates from a hospital outbreak

In this study, we investigated a 2008 outbreak of *A. baumannii* in Selly Oak Hospital in Birmingham. All outbreak isolates processed by the clinical microbiology laboratory had been determined to be clonal through conventional typing techniques. The outbreak was significant because there was a suspicion of transmission from military to civilian patients. Military patients had previously been found to be frequently colonised or infected with *A. baumannii* [132]. We explored in general terms whether whole-genome sequencing could aid our understanding of the outbreak. The series of specific overlapping questions were addressed in this study:

1. Can whole-genome sequencing be used for bacterial typing?
2. Is there variation between isolates within a small outbreak and can this variation be detected reliably?
3. Can the high resolution offered by whole-genome sequencing be used for fine-grained epidemiological typing within short timescales (days, weeks or months).
4. Can such information be used to resolve alternative infection control hypotheses, for example by shedding light on chains of transmission?
5. What are the limitations of this method?

1.2.1.2 Whole-genome comparison of two *Acinetobacter baumannii* isolates from a single patient, where resistance developed during tigecycline therapy

Following an abdominal procedure, a patient was found to have *A. baumannii* in surgical drain fluid, resistant to most antibiotics. Following a course of tigecycline chemotherapy *A. baumannii* was isolated a second time, now resistant to tigecycline but with increased susceptibility to other antibiotics. The specific questions asked in this study are:

1. How do strains evolve during infection of a single patient and during antibiotic treatment?

2. Can whole-genome sequencing provide testable hypotheses as to mechanisms of antibiotic resistance in a case of treatment failure?

1.2.1.3 Open-Source Genomic Analysis of Shiga-Toxin Producing *E. coli* O104:H4

During the spring and summer of 2011, a large outbreak of *E. coli* food poisoning occurred in Germany, causing >4000 infections and ≥40 deaths. Working prospectively, we combined whole-genome sequencing of the strain and distributed "crowd-sourced" analysis to understand the evolutionary origins and pathogen biology of this strain.

1. What is the evolutionary origin of the German *E. coli* O104:H4 outbreak strain?

2. How does this strain differ from classical enterohaemorrhagic *E. coli* (EHEC)?

3. What genetic factors might be responsible for the high levels of mortality in this outbreak?

4. How can whole-genome sequencing be used prospectively during an international outbreak?

5. What advantages does the open-endedness of genome sequencing offer? Limtations of this approach?

6. Crowd-sourcing and prospects for future outbreaks

1.2.1.4 Performance comparison of benchtop high-throughput sequencing platforms

In the previous study, the ability to sequence genomes during an outbreak was made possible by new technologies; Ion Torrent PGM, 454 GS Junior and Illumina MiSeq, all examples of low-cost benchtop sequencers. These instruments

are characterised by a much shorter running time and lower cost than the previous generation of high-throughput sequencers. We wished to determine whether these instruments were fit-for-purpose for use in future outbreaks and if there were remaining challenges needed to be addressed before high-throughput sequencing could become a routine assay in microbiology. Our specific aims were to determine:

1. How do the current benchtop sequencing platforms compare for the purpose of epidemiology and evolution studies in bacteria?
2. What are the limitations in analysing draft genome sequence data?
3. What are the practical limitations of current whole-genome sequencing platforms for genomic epidemiology and evolution?

1.3 Results and discussion

1.3.1 Paper I

1.3.1.1 Can whole-genome sequencing be used for bacterial typing? Is there variation between isolates within a small outbreak and can this variation be detected reliably? Can the high resolution offered by whole-genome sequencing be used for fine-grained epidemiological typing within short timescales? (days, weeks or months)

In this study we demonstrate that whole-genome sequencing of isolates, indistinguishable by routine typing methods such as VNTR and PFGE, can both recapitulate existing typing methods, and detect additional variation as SNPs. Phylogenetic comparisons with other sequenced strains showed that i) the outbreak strains were very closely related and sometimes there was no detectable variation between them, ii) belonged to the European Clone I lineage and iii) had many thousands of SNPs which distinguished them from other, unrelated strains.

The degree of variation we saw was low, with differences between outbreak strains found at only three loci.

A notable feature of our approach is the use of *de novo* assembly of a pooled outbreak strain, followed by mapping alignments of each isolate against the assembly. This contrasts to the approach used in most other genomic epidemiology studies which utilise a re-sequencing approach. This was not appropriate in our case because there was no closely-related reference strain available. It was notable that this approach generated a large number of likely false positive SNP and indel calls, as evidenced by the inspection of the mapping alignment, particularly associated with homopolymeric tracts, contig ends and repetitive regions. A highly stringent SNP filtering technique, adapted from Holt *et al* [133], was used in order to reduce false positives. This stringent filtering resulted in a Sanger sequencing validation rate of 100%. This gives confidence that the 454 sequencing method used, in conjunction with strict SNP filtering, is resilient enough to be used without additional validation methods in future studies.

1.3.1.2 Can such information be used to resolve alternative infection control hypotheses, for example by shedding light on chains of transmission?

The intention was that, should sufficient phylogenetic signal be detected, a tree could be constructed from the available SNPs. The preference is to use synonymous nucleotide substitutions in coding regions, which are more likely to serve as neutral markers of evolution to do this.

However, in this study we detected only three SNPs and only one was synonymous. In such circumstances, reconstructing an accurate phylogenetic tree, excluding homoplasy and recombination, is not possible. Therefore, the three variants were used as a simple genotype (similar to those used in multiplexed PCR typing systems, for example). In this case we need to make the simplifying assumption that no homoplasy was present, and that each locus, once mutated from the ancestral state (inferred by homology to an outlier strain) was unlikely to revert back to this state.

Making these simplifying assumptions, the genotypes were useful in making the case for particular chains of transmission being more parsimonious than others. However, the evidence shown is circumstantial and can only be interpreted

in the context of strong epidemiological information.

1.3.1.3 What are the limitations of this method?

Genome sequencing of bacterial isolates usually depends on culture to generate sufficient DNA template, typically between 500 nanograms and 10 micrograms depending on the instrument used. This additional culture step is potentially an issue for genomic epidemiology studies that rely on detecting such a small amount of variation. There are several potential problems. Firstly, sub-culture in the laboratory may result in mutations, either neutral or selected for by growth on a selective medium, which were not present in the original patient's infection. Secondly, not all bacterial cells in the pathogen population [**in vivo**] need be identical. The patient may have a mixed infection, with two or more quite distinct strain lineages at one or more sites; and even in a clonal population not all genotypes need be the same, as shown by the Amerithrax investigation [129]

One answer to this problem is to pick multiple colonies from a plate and sequence each one separately. However, this will be expensive and may still miss genotypes occurring at low frequency, as seen in the Amerithrax cultures. A better solution may be to employ culture-independent methods such as whole-genome shotgun metagenomics [134]. These methods permit sequencing in an unbiased manner of all DNA present in a sample. These approaches currently suffer from significant complexity in terms of data analysis and in clinical samples, where human DNA outnumbers microbial DNA significantly, the costs of these approaches are currently too high. Additionally, many samples will not have sufficient volume of high-quality DNA required for sequencing. These samples could be subjected to molecular amplification technique such as multiple displacement amplification (MDA) using random hexamer primers [135, 136]. However, these may introduce significant artefacts, including very uneven sequence coverage and the generation of chimeric (hybrid) sequences, which can hamper analysis [137–139].

1.3.2 Paper II

1.3.2.1 How do strains evolve during infection of a single patient and during antibiotic treatment?

In this second study, two isolates were retrieved from the same patient, one before and one after tigecycline therapy. Intriguingly, despite being conducted over similar time-scales as the first study, many more sequence variants were discovered, in the form of SNPs and large-scale genomic deletions. In this case, the numbers of SNPs far exceeded either the predicted nucleotide substitution rate for this genus, or the rate seen in our epidemiological study [140]. The likely explanation is the observation that the gene for an important DNA-repair enzyme was disrupted in the second isolate. This gene, *mutS* encodes one half of the two-component system, *mutRS* and is commonly associated with a hypermutator phenotype in many species, including *Acinetobacter bayli*. This report is the first to suggest that a hypermutator phenotype might be associated with clinical infection in *Acinetobacter baumannii*; this phenotype has subsequently been confirmed experimentally (Hornsey, personal communication).

Disruption of the *mutRS* system is associated with an increased mutation frequency, particularly of transitions (A⇔G, C⇔T). In the *A. bayli* ADP1 strain, disruption of *mutS* resulted in an estimated 54-fold increase in mutation rates [140]. Taken together, these results have important medical implications. *A. baumannii* is intrinsically multi-drug resistant and both extremely-drug resistant and pan-drug resistant strains have been seen. According to Martńez, hypermutable bacteria "are significantly more likely to acquire an antibiotic resistance phenotype when compared to bacteria with lower mutation rates." [141].

1.3.2.2 Can whole-genome sequencing provide testable hypotheses as to mechanisms of antibiotic resistance in a case of treatment failure?

Tigecycline resistance is associated with mutations in *adeS*, which encodes a histidine kinase sensor involved in the regulation of the drug efflux system *adeABC* [142] and a mutation in this gene was the likely cause of the drug-resistant phenotype. Interestingly, I found several regions of difference between the isolates, with

the tigecycline-resistant isolate losing several coding regions associated with antibiotic resistance, including several aminoglycoside-resistance determinants and genes encoding a beta-lactamase gene and a 16S rRNA methylase. An attractive explanation is that the disruption of *mutS* accelerated the loss of these antibiotic-resistance determinants. Whether these mutations have an effect on fitness [**in vivo**] or [**in vitro**] remains a subject for future experiments.

1.3.3 Paper III

1.3.3.1 What is the evolutionary origin of the German *E. coli* O104:H4 outbreak strain?

E. coli food-poisoning resulting in bloody diarrhoea and haemolytic-uraemic-syndrome is usually a result of infection by Shiga-toxin-producing strains of so-called enterohaemorrhagic *E. coli* (EHEC), most commonly belonging to the O157:H7 serotype. This outbreak was caused by a serotype only rarely associated with disease, O104:H4. The outbreak strain was assigned by MLST (real and virtual, based on the whole-genome sequence) to sequence type ST678. This sequence type had only one entry in the online MLST database: strain 01-09591, which had been isolated in 2001, also in Germany. However, it had been seen several times before, but none of the previous examples had been subjected to MLST. Notably, the sequence of *E. coli* 55989 had been deposited in Genbank, but was recognised as an ST678 isolate only through phylogenetic and comparative analysis of whole genomes. This strain had been isolated from an HIV-infected patient with diarrhoea in Africa in 2002.

The German outbreak strain and the two other ST678 strains belong to phylogroup B1, a lineage associated with the entero-aggregative pathovar (EAEC) rather than with EHEC. This was a surprising finding. Its unusual provenance initially hampered microbiological diagnosis, which relies on detection of sorbitol non-fermenters (a phenotype associated with the O157:H7 lineage).

1.3.3.2 How does this strain differ from classical enterohaemorrhagic *E. coli* (EHEC)? What genetic factors might be responsible for the high levels of mortality in this outbreak?

The O104:H4 outbreak strain has a phage-encoded Shiga-like toxin 2 similar to that found in EHEC strains [143], which accounts for the haemorrhagic diarrhoea and haemolytic-uraemic syndrome. However, there are distinct differences in predicted virulence gene repertoires between the outbreak strain and O157:H7 strains. The outbreak strain lacks the locus of enterocyte effacement (LEE), which codes for important virulence determinants including intimin and a type-III secretion system thought to play a key role in attachment to the lumen of the gut and responsible for secreting effector proteins into the cytosol of enterocytes [144]. However, the outbreak strain, consistent with other entero-aggregative strains, exploits alternative adhesins, including the plasmid-encoded AAF/I system. The plasmids also encode for a number of antibiotic resistance genes including *bla*(CTX-M-15).

The conflict between traditional medical classification and molecular phylogeny was significant when categorising this strain. For example, strains may be classified as Shiga-toxin producing *E. coli* (STEC) but this designation is more commonly reserved for EHEC-like strains. However, certainly this strain was STEC. However, the molecular data indicate it belongs to an entero-aggregative lineage and thus should be called an EAEC.

This raises an important issue–if whole-genome data is to be used as the ultimate typing system–can genomic data alone satisfy both the needs of clinical microbiologists to classify isolates for the purpose of diagnosis, and those of molecular epidemiologists who wish to understand evolutionary history. Currently, there is no simple answer to this question.

1.3.3.3 How can whole-genome sequencing be used prospectively during an international outbreak?

As with MLST, whole-genome sequencing data is digital and thus easily portable. The genome sequence of the *E. coli* outbreak strain was generated during the outbreak and released into the public domain, kick-starting a process of public

"crowd-sourced" analysis. This analysis led to an understanding of many important characteristics of this strain, with much analysis completed in a manner of weeks, as documented on the Github repository. A criticism of this approach is that analysis would suffer from low-quality inputs from enthusiastic amateurs. However, this was not the case, and most findings online are consistent with the eventual published literature.

However, genome analysis was confined to a handful of strains during this outbreak. Therefore there was not the opportunity to perform more extensive genomic epidemiology during the outbreak which may have mapped to transmission chains. In a future outbreak, public health laboratories with genome sequencing capacity may be able to deposit sequences for isolates as they are collected. This in turn would permit high-resolution phylogenies to be built "on-the-fly".

1.3.3.4 Crowd-sourcing and prospects for future outbreaks

Crowd-sourcing solutions to scientific problems has been made possible on a vast scale over the past 15 years enabled by the vast growth in internet-connected personal computers. A number of crowd-sourcing projects have had success through appealing to a critical mass of users (Table 1.3.3.4).

During the *E. coli* outbreak, crowd-sourcing of the genome was encouraged by timely release of genome data, from BGI and the Health Protection Agency [147] and the explicit use of licensing to allow free use of the datasets. Subsequent analysis was enabled by freely-available communications tools including blogs, Wikis, Twitter, discussion forums and source code repositories such as Github [148–150]. This was the first example of crowd-sourced analysis for bacterial genomic epidemiology, although such approaches have been used in virology, most notably during outbreaks of influenza.

Name	Type	Aim
SETI @ HOME	Distributed computing	Harness distributed computing power to search for extra-terrestrial life through signal processing of radio transmissions. Two million years total computing time logged.
Phylo [145]	Gaming	A "citizen science" project for improving multiple sequence alignments
FOLDIT	Gaming	Package the process of manual protein structure prediction as an addictive game. This project resulted in an improved crystal structure of the Mason-Pfizer monkey virus retroviral protease protein [146]

Table 1.9: Examples of popular crowd-sourcing projects

1.3.4 Paper IV

1.3.4.1 How do the current benchtop sequencing platforms compare for the purpose of epidemiology and evolution studies in bacteria?

Benchtop sequencers may speed adoption of whole-genome sequencing for clinical microbiology due to their low-cost and fast running times. The three instruments currently available on the market were all used to generate valuable data during the German *E. coli* outbreak. However, it is far from clear whether they perform equally, and whether enthusiastic adoption in public health laboratories is warranted, or whether issues remain which need to be addressed. In the study, comparisons were made between price, read length, throughput, quality and ease of use. Each instruments had strengths and weaknesses, and it was not possible to call a stand-out winner.

1.3.4.2 What are the technical obstacles in analysing draft genome sequence data?

1.3.4.3 What are the practical limitations of current whole-genome sequencing platforms for genomic epidemiology and evolution?

The final study emphasises that whole-genome data is not created equally between sequencing platforms. The main differences between platforms result from variations in read length (from 100 bp to over 500 bp), and in error rates. The 454 GS Junior and Ion Torrent PGM generate systematic errors which hamper the ability to analyse sequence data. These errors affect assembled sequences of genes important for virulence, and genes used for MLST through the introduction of frame-shift mutations. Additionally, short read lengths increase the degree of fragmentation in *de novo* assemblies, with consequent ambiguity when performing certain analyses: for example, it may difficult to determine whether sequences originate from plasmids or the chromosome.

1.4 Concluding statements

Taken together, these four studies demonstrate the promise of high-throughput sequencing in clinical microbiology and public health. Whole-genome sequencing can both recapitulate existing bacterial typing methods as well as laying claim to being the ultimate bacterial typing method; universal, digital, portable and potentially able to discriminate strains which differ by as little as one SNP. Whole-genome sequencing can also give insights into pathogen biology and reveal the underlying mutational processes responsible for the development of antibiotic resistance and immune system evasion.

There is still need to make progress in presenting this wealth of information to the needs of clinicians who currently want to know simply "What is it?" and "How can I treat it?". It may be that clinicians do not ask the best questions, particularly considering the inexorable march of developing antibiotic resistance. Management decisions in infectious disease may be better when clinicians frame their questions from an evolutionary perspective–"Where did it come from?" and "How could it change?", as well as "What else does it live with?". The bacterial species concept, already nebulous, could be discarded and classification could be based on genome data alone.

I believe that the availability of low-cost benchtop sequencing instruments will trigger a shift towards adoption of whole-genome sequencing in clinical microbiology over the next few years. Use by early adopters may be for the purpose of replacing existing bacterial typing techniques such as MLST, PFGE and VNTR. Sequencing demonstrably offers higher resolution as well as backwards-compatibility so this should be an "easy sell". Complete adoption of this technology by all clinical microbiologists is still by no means certain. In the NHS, cost will be a major limiting factor until a bacterial genome can be obtained for a price at least as cheap as existing methods. Other possible constraints to the adoption of whole-genome sequencing are listed in Table 1.10.

Looking further to the future, there is great promise in using these techniques for diagnosis of infectious disease. This is likely to rely on metagenomic approaches. Metagenomics are currently limited to research applications due to technical complexities. However this field is likely on the cusp of becom-

Microbiological
Reliance on pure culture (particularly important for slow-growing organisms)
Sequencing
High input requirements (>500 nanograms DNA)
Complex laboratory work-flows
Analytical
Difficulties due to short read lengths, errors
Data storage
Internet bandwidth
Availability of robust analysis pipelines
Lack of specialist bioinformatics skills
Professional
Inertia
Resistance to change
Proof of non-inferiority
Social
Regulatory approval (particularly in US)
Data sharing and privacy policies, particularly with metagenomics
Commercial
Competing "closed-source" methods

Table 1.10: Impediments to take up of high-throughput sequencing

ing genuinely "translatable". It may be that emerging sequencing technologies, such as those based on nanopore technologies, may allow for direct detection of microbial DNA without any sample preparation [151]. This new field, clinical metagenomics, is the area I wish to explore next, with funding obtained from the Medical Research Council.

Ultimately we may shift from Koch's original model of one bacterial species causing one disease, to a more complete understanding of how the interactions between microbial communities (the "metagenome") modulate states of health and disease. We may be able to model the complex biological interactions of infection, involving host and microbial community, including but not limited to the headline "pathogen" [152]. Such an "eco-evo" view of microbial community dynamics may suggest strategies for treating and preventing infection.

Bibliography

1. Robertson, M. Biology in the 1980s, plus or minus a decade. *Nature* **285,** 358–9 (1980).
2. Porter, J. Antony van Leeuwenhoek: tercentenary of his discovery of bacteria. *Bacteriol Rev* **40,** 260–9 (1976).
3. Guerrero, R. & Berlanga, M. The hidden side of the prokaryotic cell: rediscovering the microbial world. *Int Microbiol* **10,** 157–68 (2007).
4. Wainwright, M. An alternative view of the early history of microbiology. *Adv Appl Microbiol* **52,** 333–55 (2003).
5. Newsom, S. Pioneers in infection control: John Snow, Henry Whitehead, the Broad Street pump, and the beginnings of geographical epidemiology. *J Hosp Infect* **64,** 210–6 (2006).
6. Bentivoglio, M. & Pacini, P. Filippo Pacini: a determined observer. *Brain Res Bull* **38,** 161–5 (1995).
7. Ellis, H. Gerhard Hansen: discoverer of the organism of leprosy. *Br J Hosp Med (Lond)* **73,** 113 (2012).
8. Maloy, S. & Schaechter, M. The era of microbiology: a golden phoenix. *Int Microbiol* **9,** 1–7 (2006).
9. Oren, A. & Thane Papke, R. *Molecular Phylogeny of Microorganisms* (Caister Academic Press, 2010).
10. Brock, T. D. *Milestones in Microbiology 1546-1940* (Milestones in Microbiology, 1999).

11. Noguerola, I. & Blanch, A. Identification of *Vibrio spp.* with a set of dichotomous keys. *J Appl Microbiol* **105,** 175–85 (2008).

12. Bergey, D. *Bergey's Manual of Determinative Bacteriology* (Wiliams and Wilkins, 1923).

13. Sneath, P. & Sokal, R. Numerical taxonomy. *Nature* **193,** 855–60 (1962).

14. Sneath, p. The application of computers to taxonomy. *J Gen Microbiol* **17,** 201–26 (1957).

15. O'Malley, M. What did Darwin say about microbes, and how did microbiology respond? *Trends Microbiol* **17,** 341–7 (2009).

16. Dobzhansky, T. Biology, Molecular and Organismic. *Am Zool* **4,** 443–52 (1964).

17. Griffith, F. The Significance of *Pneumococcal* Types. *J Hyg (Lond)* **27,** 113–59 (1928).

18. Avery, O., Macleod, C. & McCarty, M. Studies on the Chemical Nature of the Substance Inducing Transformation of *Pneumococcal* Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated From *Pneumococcus* Type Iii. *J Exp Med* **79,** 137–58 (1944).

19. Watson, J. & Crick, F. A structure for deoxyribose nucleic acid. 1953. *Nature* **421,** 397–8; discussion 396 (2003).

20. Smith, H. & Wilcox, K. A restriction enzyme from *Hemophilus influenzae.* I. Purification and general properties. *J Mol Biol* **51,** 379–91 (1970).

21. Nathans, D. & Smith, H. Restriction endonucleases in the analysis and restructuring of DNA molecules. *Annu Rev Biochem* **44,** 273–93 (1975).

22. Cohen, S., Chang, A., Boyer, H. & Helling, R. Construction of biologically functional bacterial plasmids in vitro. *Proc Natl Acad Sci U S A* **70,** 3240–4 (1973).

23. Jackson, D., Symons, R. & Berg, P. Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of *Escherichia coli. Proc Natl Acad Sci U S A* **69,** 2904–9 (1972).

24. Lobban, P. & Kaiser, A. Enzymatic end-to end joining of DNA molecules. *J Mol Biol* **78,** 453–71 (1973).

25. Fiers, W. *et al.* Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260,** 500–7 (1976).

26. Maxam, A. & Gilbert, W. A new method for sequencing DNA. *Proc Natl Acad Sci U S A* **74,** 560–4 (1977).

27. Sanger, F. & Coulson, A. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94,** 441–8 (1975).

28. Sanger, F., Nicklen, S. & Coulson, A. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74,** 5463–7 (1977).

29. Zuckerkandl, E. & Pauling, L. Molecules as documents of evolutionary history. *J Theor Biol* **8,** 357–66 (1965).

30. Luria, S. & Delbrück, M. Mutations of Bacteria from Virus Sensitivity to Virus Resistance. *Genetics* **28,** 491–511 (1943).

31. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4,** 406–25 (1987).

32. Balch, W., Magrum, L., Fox, G., Wolfe, R. & Woese, C. An ancient divergence among the bacteria. *J Mol Evol* **9,** 305–11 (1977).

33. Fox, G., Magrum, L., Balch, W., Wolfe, R. & Woese, C. Classification of methanogenic bacteria by 16S ribosomal RNA characterization. *Proc Natl Acad Sci U S A* **74,** 4537–41 (1977).

34. Woese, C. & Fox, G. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A* **74,** 5088–90 (1977).

35. Lan, R. & Reeves, P. *Escherichia coli* in disguise: molecular origins of Shigella. *Microbes Infect* **4,** 1125–32 (2002).

36. Johnson, J. *Shigella* and *Escherichia coli* at the crossroads: machiavellian masqueraders or taxonomic treachery? *J Med Microbiol* **49,** 583–5 (2000).

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

38. Murray, K. DNA sequencing by mass spectrometry. *J Mass Spectrom* **31,** 1203–15 (1996).

39. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274,** 546, 563–7 (1996).

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

41. Adams, M. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287,** 2185–95 (2000).

42. NCBI. *Genome Project Website* (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi, 2012).

43. Alm, R. & Trust, T. Analysis of the genetic diversity of *Helicobacter pylori*: the tale of two genomes. *J Mol Med (Berl)* **77,** 834–46 (1999).

44. Fleischmann, R. *et al.* Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J Bacteriol* **184,** 5479–90 (2002).

45. Pizza, M. *et al.* Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* **287,** 1816–20 (2000).

46. Rappuoli, R. Reverse vaccinology. *Curr Opin Microbiol* **3,** 445–50 (2000).

47. Gossger, N. *et al.* Immunogenicity and tolerability of recombinant serogroup B meningococcal vaccine administered with or without routine infant vaccinations according to different immunization schedules: a randomized controlled trial. *JAMA* **307,** 573–82 (2012).

48. Ochman, H. & Moran, N. Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science* **292,** 1096–9 (2001).

49. Van Ham, R. *et al.* Reductive genome evolution in *Buchnera aphidicola*. *Proc Natl Acad Sci U S A* **100,** 581–6 (2003).

50. Djelouadji, Z., Raoult, D. & Drancourt, M. Palaeogenomics of *Mycobacterium tuberculosis*: epidemic bursts with a degrading genome. *Lancet Infect Dis* **11,** 641–50 (2011).

51. Singh, P. & Cole, S. *Mycobacterium leprae*: genes, pseudogenes and genetic diversity. *Future Microbiol* **6,** 57–71 (2011).

52. Vissa, V. & Brennan, P. The genome of *Mycobacterium leprae*: a minimal mycobacterial gene set. *Genome Biol* **2,** 1023 (2001).

53. Cole, S. *et al.* Massive gene decay in the leprosy bacillus. *Nature* **409,** 1007–11 (2001).

54. Han, X. *et al.* A new *Mycobacterium* species causing diffuse lepromatous leprosy. *Am J Clin Pathol* **130,** 856–64 (2008).

55. Han, X. *et al.* Comparative sequence analysis of *Mycobacterium leprae* and the new leprosy-causing *Mycobacterium lepromatosis*. *J Bacteriol* **191,** 6067–74 (2009).

56. Smith, J., Smith, N., O'Rourke, M. & Spratt, B. How clonal are bacteria? *Proc Natl Acad Sci U S A* **90,** 4384–8 (1993).

57. Parkhill, J. *et al.* Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* **413,** 523–7 (2001).

58. Van Ert, M. *et al.* Global genetic population structure of *Bacillus anthracis*. *PLoS One* **2,** e461 (2007).

59. Han, C. *et al.* Pathogenomic sequence analysis of *Bacillus cereus* and *Bacillus thuringiensis* isolates closely related to *Bacillus anthracis*. *J Bacteriol* **188,** 3382–90 (2006).

60. Read, T. *et al.* The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. *Nature* **423,** 81–6 (2003).

61. Eppinger, M., Baar, C., Raddatz, G., Huson, D. & Schuster, S. Comparative analysis of four Campylobacterales. *Nat Rev Microbiol* **2,** 872–85 (2004).

62. Furuta, Y. *et al.* Birth and death of genes linked to chromosomal inversion. *Proc Natl Acad Sci U S A* **108,** 1501–6 (2011).

63. Armstrong, G., Conn, L. & Pinner, R. Trends in infectious disease mortality in the United States during the 20th century. *JAMA* **281,** 61–6 (1999).

64. Hinman, A. Global progress in infectious disease control. *Vaccine* **16,** 1116–21 (1998).

65. Stewart, G., Robertson, B. & Young, D. Tuberculosis: a problem with persistence. *Nat Rev Microbiol* **1,** 97–105 (2003).

66. Neu, H. The crisis in antibiotic resistance. *Science* **257,** 1064–73 (1992).

67. Bush, K. *et al.* Tackling antibiotic resistance. *Nat Rev Microbiol* **9,** 894–6 (2011).

68. Dijkshoorn, L., Nemec, A. & Seifert, H. An increasing threat in hospitals: multidrug-resistant *Acinetobacter baumannii*. *Nat Rev Microbiol* **5,** 939–51 (2007).

69. Maltezou, H. Metallo-beta-lactamases in Gram-negative bacteria: introducing the era of pan-resistance? *Int J Antimicrob Agents* **33,** 405.e1–7 (2009).

70. Bonomo, R. New Delhi metallo–lactamase and multidrug resistance: a global SOS? *Clin Infect Dis* **52,** 485–7 (2011).

71. Loman, N. & Pallen, M. XDR-TB genome sequencing: a glimpse of the microbiology of the future. *Future Microbiol* **3,** 111–3 (2008).

72. Saleem, A., Ahmed, I., Mir, F., Ali, S. & Zaidi, A. Pan-resistant Acinetobacter infection in neonates in Karachi, Pakistan. *J Infect Dev Ctries* **4,** 30–7 (2010).

73. Udwadia, Z., Amale, R., Ajbani, K. & Rodrigues, C. Totally drug-resistant tuberculosis in India. *Clin Infect Dis* **54,** 579–81 (2012).

74. Wenzel, R. The antibiotic pipeline–challenges, costs, and values. *N Engl J Med* **351,** 523–6 (2004).

75. Maiden, M. *et al.* Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* **95,** 3140–5 (1998).

76. Achtman, M. A surfeit of YATMs? *J Clin Microbiol* **34,** 1870 (1996).

77. Selander, R. *et al.* Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. *Appl Environ Microbiol* **51,** 873–84 (1986).

78. Lindstedt, B. Multiple-locus variable number tandem repeats analysis for genetic fingerprinting of pathogenic bacteria. *Electrophoresis* **26,** 2567–82 (2005).

79. Anderson, D., Kuhns, J., Vasil, M., Gerding, D. & Janoff, E. DNA fingerprinting by pulsed field gel electrophoresis and ribotyping to distinguish *Pseudomonas cepacia* isolates from a nosocomial outbreak. *J Clin Microbiol* **29,** 648–9 (1991).

80. Arbeit, R. *et al.* Resolution of recent evolutionary divergence among *Escherichia coli* from related lineages: the application of pulsed field electrophoresis to molecular epidemiology. *J Infect Dis* **161,** 230–5 (1990).

81. Tenover, F. *et al.* Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. *J Clin Microbiol* **33,** 2233–9 (1995).

82. Saiki, R. *et al.* Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239,** 487–91 (1988).

83. Woods, C., Versalovic, J., Koeuth, T. & Lupski, J. Whole-cell repetitive element sequence-based polymerase chain reaction allows rapid assessment of clonal relationships of bacterial isolates. *J Clin Microbiol* **31,** 1927–31 (1993).

84. Roumagnac, P. *et al.* Evolutionary history of *Salmonella typhi*. *Science* **314,** 1301–4 (2006).

85. Wirth, T. *et al.* Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol Microbiol* **60,** 1136–51 (2006).

86. Bartual, S. *et al.* Development of a multilocus sequence typing scheme for characterization of clinical isolates of *Acinetobacter baumannii*. *J Clin Microbiol* **43,** 4382–90 (2005).

87. Diancourt, L., Passet, V., Nemec, A., Dijkshoorn, L. & Brisse, S. The population structure of *Acinetobacter baumannii*: expanding multiresistant clones from an ancestral susceptible genetic pool. *PLoS One* **5,** e10034 (2010).

88. Baker, L., Brown, T., Maiden, M. & Drobniewski, F. Silent nucleotide polymorphisms and a phylogeny for Mycobacterium tuberculosis. *Emerg Infect Dis* **10,** 1568–77 (2004).

89. Comas, I., Homolka, S., Niemann, S. & Gagneux, S. Genotyping of genetically monomorphic bacteria: DNA sequencing in Mycobacterium tuberculosis highlights the limitations of current methodologies. *PLoS One* **4,** e7815 (2009).

90. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–80 (2005).

91. Wheeler, D. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452,** 872–6 (2008).

92. Bennett, S. Solexa Ltd. *Pharmacogenomics* **5,** 433–8 (2004).

93. Bentley, D. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–9 (2008).

94. Metzker, M. Sequencing technologies - the next generation. *Nat Rev Genet* **11,** 31–46 (2010).

95. McKernan, K. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res* **19,** 1527–41 (2009).

96. Pushkarev, D., Neff, N. & Quake, S. Single-molecule sequencing of an individual human genome. *Nat Biotechnol* **27,** 847–50 (2009).

97. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323,** 133–8 (2009).

98. Rothberg, J. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475,** 348–52 (2011).

99. Smith, T. & Waterman, M. Identification of common molecular subsequences. *J Mol Biol* **147,** 195–7 (1981).

100. Needleman, S. & Wunsch, C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48,** 443–53 (1970).

101. Pop, M. & Salzberg, S. Bioinformatics challenges of new sequencing technology. *Trends Genet* **24,** 142–9 (2008).

102. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18,** 1851–8 (2008).

103. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–95 (2010).

104. Ruffalo, M., LaFramboise, T. & Koyutürk, M. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics* **27,** 2790–6 (2011).

105. Ning, Z., Cox, A. & Mullikin, J. SSAHA: a fast search method for large DNA databases. *Genome Res* **11,** 1725–9 (2001).

106. Galinsky, V. YOABS: yet other aligner of biological sequences–an efficient linearly scaling nucleotide aligner. *Bioinformatics* **28,** 1070–7 (2012).

107. Paszkiewicz, K. & Studholme, D. De novo assembly of short sequence reads. *Brief Bioinform* **11,** 457–72 (2010).

108. Zerbino, D. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18,** 821–9 (2008).

109. Li, R. *et al.* The sequence and de novo assembly of the giant panda genome. *Nature* **463,** 311–7 (2010).

110. Simpson, J. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res* **19,** 1117–23 (2009).

111. Delcher, A., Bratke, K., Powers, E. & Salzberg, S. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23,** 673–9 (2007).

112. Borodovsky, M. & Lomsadze, A. Gene identification in prokaryotic genomes, phages, metagenomes, and EST sequences with GeneMarkS suite. *Curr Protoc Bioinformatics* **Chapter 4,** Unit 4.5.1–17 (2011).

113. Altschul, S. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25,** 3389–402 (1997).

114. Lowe, T. & Eddy, S. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25,** 955–64 (1997).

115. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* **35,** 3100–8 (2007).

116. Sayers, E. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **40,** D13–25 (2012).

117. Chaudhuri, R. *et al.* xBASE2: a comprehensive resource for comparative bacterial genomics. *Nucleic Acids Res* **36,** D543–6 (2008).

118. Harris, S. *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327,** 469–74 (2010).

119. Croucher, N. *et al.* Rapid pneumococcal evolution in response to clinical interventions. *Science* **331,** 430–4 (2011).

120. Monot, M. *et al.* Comparative genomic and phylogeographic analysis of *Mycobacterium leprae*. *Nat Genet* **41,** 1282–9 (2009).

121. He, M. *et al.* Evolutionary dynamics of *Clostridium difficile* over short and long time scales. *Proc Natl Acad Sci U S A* **107,** 7527–32 (2010).

122. Stabler, R. *et al.* Comparative genome and phenotypic analysis of *Clostridium difficile* 027 strains provides insight into the evolution of a hypervirulent bacterium. *Genome Biol* **10,** R102 (2009).

123. Chin, C. *et al.* The origin of the Haitian cholera outbreak strain. *N Engl J Med* **364,** 33–42 (2011).

124. Mutreja, A. *et al.* Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **477,** 462–5 (2011).

125. Beres, S. *et al.* Molecular complexity of successive bacterial epidemics deconvoluted by comparative pathogenomics. *Proc Natl Acad Sci U S A* **107,** 4371–6 (2010).

126. Fittipaldi, N., Olsen, R., Beres, S., Van Beneden, C. & Musser, J. Genomic Analysis of emm59 Group A *Streptococcus* Invasive Strains, United States. *Emerg Infect Dis* **18,** 650–2 (2012).

127. Gilmour, M. *et al.* High-throughput genome sequencing of two *Listeria monocytogenes* clinical isolates during a large foodborne outbreak. *BMC Genomics* **11,** 120 (2010).

128. Lienau, E. *et al.* Identification of a salmonellosis outbreak by means of molecular sequencing. *N Engl J Med* **364,** 981–2 (2011).

129. Rasko, D. *et al. Bacillus anthracis* comparative genome analysis in support of the Amerithrax investigation. *Proc Natl Acad Sci U S A* **108,** 5027–32 (2011).

130. Golubchik, T. *et al. Pneumococcal* genome sequencing tracks a vaccine escape variant formed through a multi-fragment recombination event. *Nat Genet* **44,** 352–5 (2012).

131. Ou, C. *et al.* Molecular epidemiology of HIV transmission in a dental practice. *Science* **256,** 1165–71 (1992).

132. Jones, A. *et al.* Importation of multidrug-resistant *Acinetobacter spp* infections with casualties from Iraq. *Lancet Infect Dis* **6,** 317–8 (2006).

133. Holt, K. *et al.* High-throughput sequencing provides insights into genome variation and evolution in *Salmonella Typhi*. *Nat Genet* **40,** 987–93 (2008).

134. Handelsman, J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* **68,** 669–85 (2004).

135. Raghunathan, A. *et al.* Genomic DNA amplification from a single bacterium. *Appl Environ Microbiol* **71,** 3342–7 (2005).

136. Marcy, Y. *et al.* Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc Natl Acad Sci U S A* **104,** 11889–94 (2007).

137. Walker, A. & Parkhill, J. Single-cell genomics. *Nat Rev Microbiol* **6,** 176–7 (2008).

138. Yilmaz, S. & Singh, A. Single cell genome sequencing. *Curr Opin Biotechnol* (2011).

139. Woyke, T. *et al.* One bacterial cell, one complete genome. *PLoS One* **5,** e10314 (2010).

140. Hall, L. & Henderson-Begg, S. Hypermutable bacteria isolated from humans–a critical analysis. *Microbiology* **152,** 2505–14 (2006).

141. Martínez, J., Baquero, F. & Andersson, D. Predicting antibiotic resistance. *Nat Rev Microbiol* **5,** 958–65 (2007).

142. Ruzin, A., Keeney, D. & Bradford, P. AdeABC multidrug efflux pump is associated with decreased susceptibility to tigecycline in *Acinetobacter calcoaceticus-Acinetobacter baumannii* complex. *J Antimicrob Chemother* **59,** 1001–4 (2007).

143. Fraser, M. *et al.* Structure of Shiga toxin type 2 (Stx2) from *Escherichia coli* O157:H7. *J Biol Chem* **279,** 27511–7 (2004).

144. Pallen, M., Beatson, S. & Bailey, C. Bioinformatics analysis of the locus for enterocyte effacement provides novel insights into type-III secretion. *BMC Microbiol* **5,** 9 (2005).

145. Kawrykow, A. *et al.* Phylo: a citizen science approach for improving multiple sequence alignment. *PLoS One* **7,** e31362 (2012).

146. Khatib, F. *et al.* Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nat Struct Mol Biol* **18,** 1175–7 (2011).

147. Genomic data from *Escherichia coli* O104:H4 isolate TY-2482. *BGI Shenzhen.* doi:10.5524/100001 (2011).

148. Loman, N. J. 2012. ¡http://pathogenomics.bham.ac.uk/blog¿.

149. Li, J. *et al.* SEQanswers: an open access community for collaboratively decoding genomes. *Bioinformatics* **28,** 1272–3 (2012).

150. Li, J. *et al.* The SEQanswers wiki: a wiki database of tools for high-throughput sequencing analysis. *Nucleic Acids Res* **40,** D1313–7 (2012).

151. Schneider, G. & Dekker, C. DNA sequencing with nanopores. *Nat Biotechnol* **30,** 326–8 (2012).

152. Mokili, J., Rohwer, F. & Dutilh, B. Metagenomics and future perspectives in virus discovery. *Curr Opin Virol* **2,** 63–77 (2012).

Chapter 2

High-throughput whole-genome sequencing to dissect the epidemiology of *Acinetobacter baumannii* isolates from a hospital outbreak

Use of high-throughput whole-genome sequencing to dissect the epidemiology of *Acinetobacter baumannii* isolates from a hospital outbreak

*Dr Tom Lewis[a], *Dr Nicholas J. Loman[b], Dr Lewis Bingle[b], Dr Pauline Jumaa[a], Professor George M. Weinstock[c], Dr Deborah Mortiboy[a] and Professor Mark J. Pallen[b,#]

[a]Department of Medical Microbiology, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK

[b]Centre for Systems Biology, University of Birmingham, Birmingham, UK

[c]Department of Genetics, Washington University School of Medicine, St. Louis, Missouri, USA

*These authors contributed equally to this work

Corresponding author. Address: Centre for Systems Biology, University of Birmingham, Birmingham, B15 2TT, UK. Tel: +44 121 414 7163.

E-mail: m.pallen@bham.ac.uk

Running title High-throughput sequencing of *Acinetobacter* isolates

Summary

Shared care of military and civilian patients has resulted in transmission of multi-drug-resistant *Acinetobacter baumannii* (MDR-Aci) from military casualties to civilians. Current typing technologies have proven useful in revealing relationships between *A. baumannii* isolates. However, they are unable to resolve differences between closely related isolates from small-scale outbreaks, where chains of transmission are often unclear. In a recent hospital outbreak in Birmingham, six patients were colonized with MDR-Aci isolates indistinguishable using standard techniques. We have used whole-genome sequencing to identify single nucleotide polymorphisms (SNPs) in these isolates, allowing us to discriminate between alternative epidemiological hypotheses in this setting.

Introduction

In the United Kingdom, military casualties are usually repatriated to Selly Oak Hospital, Birmingham, where they are cared for alongside civilian patients. Military patients from Iraq and Afghanistan are often colonized with strains of multi-drug-resistant *A. baumannii* (MDR-Aci), which can spread to civilian patients and health-care workers.[1-6] Molecular typing systems, such as pulsed-field gel electrophoresis (PFGE) and variable-number tandem repeat (VNTR) analyses, have provided evidence of multiple concurrent or successive clonal outbreaks in hospitals across the UK.[7-8] However, these methods have been unable to provide sufficient resolution to determine chains of transmission within apparently clonal outbreaks. Nor can they provide detailed information on patterns of spread (e.g. the influence of "super-shedders", environmental persistence, staff carriage or infection control practices), even though these questions are important when considering where to focus finite infection control resources.

A recent MDR-Aci outbreak in Selly Oak Hospital illustrates some of these problems. Four military patients, admitted over a five-week period, were each found to be colonised with MDR-Aci a few days after admission. Subsequently, indistinguishable isolates were recovered from two civilian patients on the same unit. Molecular typing distinguished isolates from this outbreak from those recovered from a similar outbreak in 2007 (same PFGE type, different VNTR type). However, such approaches were unable to shed light on transmission events within the 2008 outbreak itself.

Whole-genome sequencing represents the ultimate molecular typing method for bacteria, because it samples the entire collection of genetic information within each isolate. Pioneering work on the "Amerithrax" strain of *Bacillus anthracis* illustrated the utility of this approach[9] as long ago as 2002. However, until recently, the cost and technical complexity of bacterial whole-genome sequencing placed it beyond the reach of the average diagnostic laboratory or academic research group. This has changed in the last couple of years with the advent of "high-throughput sequencing"—an umbrella term for several competing technologies that deliver genome sequences around one hundred times more quickly and more cheaply than conventional sequencing approaches (see recent review by Metzker[10]). All such technologies do away with the need for cloning of DNA in biological systems and instead rely on massively parallel *in vitro* amplification of template molecules attached to a solid surface.

Several recent studies have shown that analysis of single nucleotide polymorphisms (SNPs) in bacterial genomes provides a means of determining relatedness between epidemiologically linked isolates and tracking bacterial evolution over periods of months to years.[11-16] Furthermore, the massive depth of coverage provided by high-throughput sequencing means that, when looking for rare genomic changes, it becomes efficient to pool samples and identity variable loci by polymorphisms within the consensus sequence. SNPs can then be quickly and easily assigned to individual isolates by a small number of confirmatory PCRs.

Despite the promise of these new technologies, at the outset of this study, it remained unclear whether MDR-Aci lineages associated with individual patients harbour SNPs capable of providing useful epidemiological information. We therefore applied one particular high-throughput sequencing technology—454 pyrosequencing—to isolates from our outbreak in the hope of gaining additional epidemiological information.

Methods

Microbiology

A. baumannii isolates were obtained from routine clinical samples. Bacterial identification and antibiotic susceptibility testing was performed on the Vitek 2 system according to the manufacturer's instructions (bioMérieux, Basingstoke, UK), supplemented by CLSI-recommended confirmatory testing. Isolates were frozen on beads and stored at -20°C. Isolates M1, M2, M3, M4 were obtained from wound swabs from military patients. Isolates C1 and C2 were obtained from sputum cultures from civilian patients. Isolates C1-2a and C1-2b were distinct colonies from the same wound specimen, taken two weeks after the initial isolate from civilian patient C1. Isolates C1-3a and C1-3b were isolates from a different wound specimen, taken at the same time as isolates C1-2a and C1-2b. Multidrug resistance was defined as resistance to ≥ 3 classes of antibiotics (quinolones, extended-spectrum cephalosporins, ß-lactam/ß-lactamase inhibitor combinations, aminoglycosides and carbapenems). MDR-Aci isolates were sent to the Laboratory of HealthCare Associated Infection for speciation and PFGE and VNTR analysis. Antibiotic sensitivities were confirmed by agar dilution methods in two isolates (M1 and C1).

Isolation of genomic DNA and 454 sequencing

Genomic DNA was obtained from colony-purified MDR-Aci, using the DNeasy DNA extraction kit (Qiagen, Crawley, UK). DNA was sequenced using 454 Titanium protocols (Roche, Welwyn Garden City, Hertfordshire, UK) at the University of Liverpool's Centre for Genomic Research. DNA samples from isolates M1 and C1 were each sequenced on a quarter plate; the sample from isolate C2 was sequenced on two quarter-plates. Approximately equal quantities of DNA from isolates M2, M3, M4 and C1-2a were pooled and sequenced on a full Titanium plate.

Analysis of genome sequence and of sequence variants

454 sequence reads were assembled using Newbler 2.0.01.14 (Roche, Welwyn Garden City, Hertfordshire, UK). We combined the sequence data from all MDR-Aci isolates to create a consensus outbreak assembly. The gsMapper component of Newbler was then used to map each set of sequence reads against the consensus assembly. False positive variants resulting from sequencing errors were excluded, generating a set of high-confidence variants. This set was subjected to several additional rounds of filtering using xBASE-NG[17] to generate a set of well-trusted SNPs. During this process, we discarded

1. insertions and deletions
2. variants present in < 90% of mapped reads from the runs with single genomes, or in <25% of reads from the pooled sample.
3. variants with excess coverage, > 1 standard deviation from the mean (i.e. in repetitive regions).
4. variants occurring within 200 bases of a contig boundary.
5. variants occurring in clusters (≥3 SNPs in 1000 base pairs)
6. variants not flanked by good-quality coverage for ≥ 20 basepairs.
7. variants where the ancestral state could not be determined by reference to published genomes using BLASTN

Validation of single nucleotide polymorphisms

All well-trusted SNPs were investigated by polymerase chain reaction (PCR) and Sanger sequencing. The sequences of the primers used for this purpose were as follows: SNP1 TAAGGCAGAACAAAGCGTGA/AATCGGTTCTGAGGTTTGGA (product size 222bp); SNP2

GGTGAACCTTGGTGGTGGTA/AGCTTTAATGGCTGCTCGAA (product size 222bp); SNP3 CATTTCCGAAACCCTCTGAT/AGGCGGTATTTGATGATCTTG (product size 218bp).

PCR products were purified by ethanol precipitation and sequenced on a 3730 DNA analyser (Applied Biosystems, Warrington, UK). The sequences of SNP loci from isolates C1-2b, C1-3a and C1-3b were determined only by sequencing of PCR products.

Results

Description of the outbreak

MDR-Aci was isolated from specimens from five male patients, admitted to the Selly Oak Hospital critical care unit over a six-week period in late 2008. Three patients were military [M1, M2, M3]; two were civilian [C1, C2]. MDR-Aci was also isolated from a wound swab from a military patient [M4] in a nearby trauma ward. Figure 1 shows a time line of the MDR-Aci cases on the critical care unit. The critical care unit is split into a ten-bedded unit and a six bedded-unit, separated by a narrow corridor. Most military patients are admitted to the six-bedded unit, which includes a four-bedded bay (beds 1-4) and a two-bedded bay (beds 5-6), separated by an open thoroughway to the main unit. Patient C1 was first found to be colonized while on the main ten-bedded unit, and was subsequently transferred to the six-bedded unit. The other four MDR-Aci-positive critical-care patients (M1, M2, M3, C2) were cared for exclusively in the six-bedded unit.

All MDR-Aci isolates had an identical profile by PFGE and VNTR analyses and all fell within European clone 1 (Turton, personal communication; data not shown).[18] The antibiotic resistance profile was also identical for all isolates (resistant to meropenem, piptazobactam, amikacin; sensitive to gentamicin and tigecycline).

The first MDR-Aci isolate was from patient M1. The last isolate came from patient C2, who yielded a positive sample in week 7. During our initial epidemiological evaluation, we assumed that all military patients were colonised prior to admission. However, the events leading to colonisation of the civilian patients remained unclear. In particular, several epidemiological scenarios could explain the acquisition of MDR-Aci by patient C2 (Figure 1):

1. Transmission from M1. C2 was nursed in a bed next to M1 during week 2.
2. Transmission from M2. C2 was nursed in a bed next to M2 during week 4.
3. Transmission from M3, who occupied a nearby bed space in the six-bedded unit in the two weeks before MDR-Aci was first isolated from C2.
4. Transmission from C1, who occupied a nearby bed space in the six-bedded unit in the week before MDR-Aci was first isolated from C2
5. Acquisition from an unknown source, such as an environmental reservoir or an unidentified patient or health worker.

Transmission from M4 was rendered unlikely by the lack of proximity to C2 in time and space.

Whole-genome sequencing of outbreak strains

Complete genome sequence data was obtained from the initial MDR-Aci isolates from the six patients plus one additional isolate from patient C1 obtained two weeks after colonisation was first detected (C1-2a) (Table 1). The consensus outbreak assembly output by Newbler comprised 4,110,513 base-pairs, containing 107 contigs ≥500 base-pairs. Average contig size was 38,416 base-pairs, with an assembly N50 of 79,057 base-

pairs. The largest contig was 230,667 base pairs. As predicted from multiplex PCR (reference 8) and PFGE analysis, the outbreak strains align most closely with sequences from other representatives of European clone 1 (data not shown).

Detection and interpretation of SNPs

Several hundred variants were discarded (supplementary data) during the filtering of variants to create high-quality informative SNPs associated with three polymorphic loci (Table 2). The sequences at each SNP locus in each isolate (including those that had not been genome-sequenced) were determined by PCR and Sanger sequencing. For the genome-sequenced isolates, there was complete agreement here with the 454 data. The SNPs were placed in a biological and phylogenetic context by reference to the complete genome sequence of the European clone 1 strain AB0057.[19] This comparison identified the M1 isolate as bearing the ancestral genotype at all three SNP loci.

The first SNP distinguishes all the other outbreak isolates (M2, M3, M4, C1, C2) from the ancestral/M1 state. A second SNP separates the C1 isolate from all the other isolates, while a third SNP differentiates the M3 isolate from all the other isolates. No SNPs were detected between isolates from patients C2, M2 or M4. Interestingly, patients M2 and M4 were injured in the same incident and had similar pathways of care until arrival in Selly Oak Hospital. However, once in Birmingham, patient M4 did not come into close contact with patients M2 and C2.

If we assume that all military patients acquired MDR-Aci before arrival in Birmingham, then SNP 1 must have been acquired before admission and so patient M1 cannot be the source of any of the civilian cases (Figure 1). Transmission of MDR-Aci from C1 or M3 to C2 is ruled out by the very low probability of reversion to the ancestral state for SNPs 2 and 3. Therefore, we conclude that the most parsimonious interpretation of the data is that patient C2 acquired MDR-Aci from patient M2.

Discussion

Current typing systems have provided valuable insights into the epidemiology of MDR-Aci. However, we postulated that whole-genome sequencing might provide more detailed resolution between bacterial isolates from a hospital outbreak. Here, we have shown that closely related MDR-Aci lineages contain SNPs that can shed light on transmission events within a small-scale outbreak and can discriminate between alternative epidemiological hypotheses.

Our analyses support transmission of MDR-Aci from the wound of a military patient M2 to the respiratory tract of a civilian patient C2. However, as MDR-Aci was not isolated from C2 until several weeks after M2 left the adjacent bed, we cannot determine when and how transmission occurred. One possibility is that C2 became colonised when the two patients were nursed together, but that colonization did not reach detectable levels in the sputum until much later. Another possibility is that M2 contaminated the local environment and C2 acquired the organism from the environment only after M2 had left the ward. This latter option would be consistent with a significant role of the environment in transmission of MDR-Aci, as suggested by others.[20-24]

Other uncertainties remain in the epidemiology of this outbreak. In particular, we were not able to determine the source and mode of MDR-Aci transmission to patient C1. The isolate from this patient contained a SNP not present in any of the military isolates and prior to detection of MDR-Aci, this patient never came into close proximity with any

other patients known to be colonized. Curiously, one of the five MDR-Aci isolates from this patient (C1-2a) possessed three additional SNPs not found in any of the other isolates (data not shown). The co-existence of two closely related but distinct lineages of MDR-Aci in samples from the same patient remains puzzling and perhaps reflects two separate acquisition events before arrival in Birmingham.

In conclusion, we have highlighted the potential of whole-genome sequencing in the analysis of hospital outbreaks. However, it is worth stressing that in considering only well-validated SNPs, we have been conservative in our analysis and in future studies, additional phylogenetic analyses (e.g. of short repeats, indels or re-arrangements) twinned with genome finishing methods, such as gap closure, might provide further discrimination between closely related MDR-Aci isolates.

It is also clear that additional studies are needed to benchmark genomic variability within populations of MDR-Aci colonizing individual patients, to determine how frequently SNPs arise within a lineage, to dissect the local, national and global population genomics of *A. baumannii* at the highest resolution possible and to optimise use of this technology in hospital infection control. In addition, this technology is certain to illuminate key biological differences between isolates by revealing the genetic determinants associated with virulence or antibiotic resistance. Furthermore, as improvements in high-throughput sequencing result in reduced costs and increased efficiency, whole-genome sequencing will increasingly come within the reach of clinical microbiology laboratories. It is not hard to imagine a time when genome sequencing replaces gel-based methods as the typing method of choice for bacterial nosocomial pathogens.

Sequence Data

454 sequencing reads have been deposited to the NCBI Short Read Archive under reference SRA010038. Supplementary data is available from http://pathogenomics.bham.ac.uk/acinetobacter/

Funding

Genome sequencing was funded by a Small Research Grant from the Hospital Infection Society, London, UK. The xBASE facility and Loman's position are funded by BBSRC grant BBE0111791.

Acknowledgements

We would like to thank the infection control team at UHB, in particular Jane Heron and Jane McGeown, for their help in collecting epidemiological information. We thank Jane Turton and others in the Health Protection Agency's Laboratory of HealthCare Associated Infection for PFGE and VNTR data on clinical isolates.

References

1. Turton JF, Kaufmann ME, Gill MJ, *et al.* Comparison of *Acinetobacter baumannii* isolates from the United Kingdom and the United States that were associated with repatriated casualties of the Iraq conflict. *J Clin Microbiol*. 2006; **44**: 2630-4.
2. Towner KJ. *Acinetobacter*: an old friend, but a new enemy. *J Hosp Infect*. 2009; **73**: 355-63.
3. Peleg AY, Seifert H, Paterson DL. *Acinetobacter baumannii*: emergence of a successful pathogen. *Clin Microbiol Rev*. 2008; **21**: 538-82.
4. Jones A, Morgan D, Walsh A, *et al.* Importation of multidrug-resistant *Acinetobacter* spp infections with casualties from Iraq. *Lancet Infect Dis*. 2006; **6**: 317-8.
5. Whitman TJ, Qasba SS, Timpone JG, *et al.* Occupational transmission of *Acinetobacter baumannii* from a United States serviceman wounded in Iraq to a health care worker. *Clin Infect Dis*. 2008; **47**: 439-43.
6. Hujer KM, Hujer AM, Hulten EA, *et al.* Analysis of antibiotic resistance genes in multidrug-resistant *Acinetobacter* sp. isolates from military and civilian patients treated at the Walter Reed Army Medical Center. *Antimicrob Agents Chemother*. 2006; **50**: 4114-23.
7. Turton JF, Matos J, Kaufmann ME, Pitt TL. Variable number tandem repeat loci providing discrimination within widespread genotypes of *Acinetobacter baumannii*. *Eur J Clin Microbiol Infect Dis*. 2009; **28**: 499-507.
8. Turton JF, Gabriel SN, Valderrey C, Kaufmann ME, Pitt TL. Use of sequence-based typing and multiplex PCR to identify clonal lineages of outbreak strains of *Acinetobacter baumannii*. *Clin Microbiol Infect*. 2007; **13**: 807-15.
9. Read TD, Salzberg SL, Pop M, *et al.* Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. *Science*. 2002; **296**: 2028-33.
10. Metzker ML Sequencing technologies - the next generation. Nat Rev Genet. 2010; **11**: 31-46.
11. Maharjan RP, Gu C, Reeves PR, Sintchenko V, Gilbert GL, Lan R. Genome-wide analysis of single nucleotide polymorphisms in *Bordetella pertussis* using comparative genomic sequencing. *Res Microbiol*. 2008; **159**: 602-8.
12. Holt KE, Parkhill J, Mazzoni CJ, *et al.* High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. *Nat Genet*. 2008; **40**: 987-93.
13. Orsi RH, Borowsky ML, Lauer P, *et al.* Short-term genome evolution of *Listeria monocytogenes* in a non-controlled environment. *BMC Genomics*. 2008; **9**: 539.
14. Garcia Pelayo MC, Uplekar S, Keniry A, *et al.* A comprehensive survey of single nucleotide polymorphisms (SNPs) across *Mycobacterium bovis* strains and *M. bovis* BCG vaccine strains refines the genealogy and defines a minimal set of SNPs that separate virulent *M. bovis* strains and *M. bovis* BCG strains. *Infect Immun*. 2009; **77**: 2230-8.
15. Chaudhuri RR, Ren CP, Desmond L, *et al.* Genome sequencing shows that European isolates of *Francisella tularensis* subspecies *tularensis* are almost identical to US laboratory strain Schu S4. *PLoS One*. 2007; **2**: e352.
16. Barrick JE, Yu DS, Yoon SH, *et al.* Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature*. 2009; **461**: 1243-7.
17. Chaudhuri RR, Loman NJ, Snyder LA, Bailey CM, Stekel DJ, Pallen MJ. xBASE2: a comprehensive resource for comparative bacterial genomics. *Nucleic Acids Res*. 2008; **36**: D543-6.
18. Dijkshoorn L, Aucken H, Gerner-Smidt P, *et al.* Comparison of outbreak and nonoutbreak *Acinetobacter baumannii* strains by genotypic and phenotypic

methods. *J Clin Microbiol*. 1996; **34**: 1519-25.
19. Adams MD, Goglin K, Molyneaux N*, et al.* Comparative genome sequence analysis of multidrug-resistant *Acinetobacter baumannii*. *J Bacteriol*. 2008;**190**: 8053-64.
20. Denton M, Wilcox MH, Parnell P*, et al.* Role of environmental cleaning in controlling an outbreak of *Acinetobacter baumannii* on a neurosurgical intensive care unit. *J Hosp Infect*. 2004; **56**: 106-10.
21. Enoch DA, Summers C, Brown NM*, et al.* Investigation and management of an outbreak of multidrug-carbapenem-resistant *Acinetobacter baumannii* in Cambridge, UK. *J Hosp Infect*. 2008; **70**: 109-18.
22. Markogiannakis A, Fildisis G, Tsiplakou S*, et al.* Cross-transmission of multidrug-resistant *Acinetobacter baumannii* clonal strains causing episodes of sepsis in a trauma intensive care unit. *Infect Control Hosp Epidemiol*. 2008; **29**: 410-7.
23. Simor AE, Lee M, Vearncombe M*, et al.* An outbreak due to multiresistant *Acinetobacter baumannii* in a burn unit: risk factors for acquisition and management. *Infect Control Hosp Epidemiol*. 2002; **23**: 261-7.
24. Getchell-White SI, Donowitz LG, Groschel DH. The inanimate environment of an intensive care unit as a potential source of nosocomial bacteria: evidence for long survival of *Acinetobacter calcoaceticus*. *Infect Control Hosp Epidemiol*. 1989;**10**: 402-7.

Figures and Tables

Figure 1 - Legend

Time line showing bedspaces of individual patients while in the six-bedded bay of the critical care unit. Vertical bars indicate a positive *MDR-Aci* isolate from the patient and their corresponding SNP genotype. Patient C2 had sputum samples sent for microbiological analysis on day 24 and day 42, from which MDR-Aci was not isolated. Patient C1 was initially admitted to the ten bedded main section of the critical care unit, on the same day as patient M2 was admitted to the six bedded section. Patient C1 was first found to be colonized with MDR-Aci two days after patient M2. The arrow shows proposed transmission from M2 to C2.

Table 1

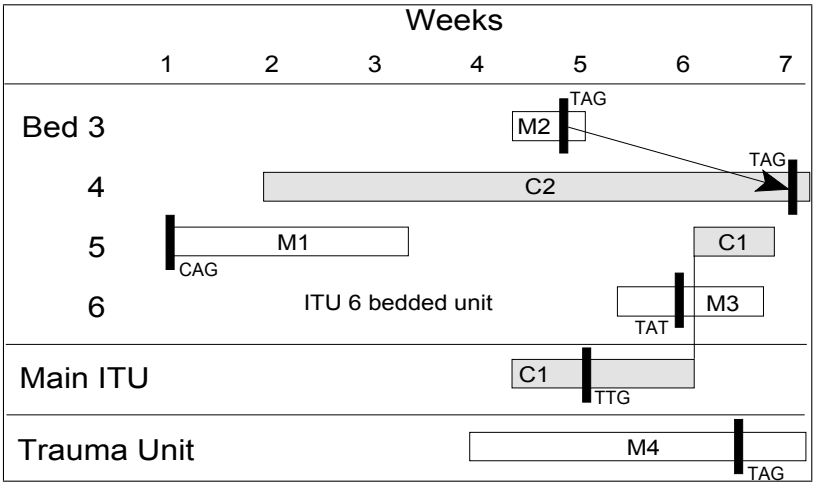
Isolate	*Reads*	*Aligned bases*	*Mean coverage depth*	*SNPs after filtering*
M1	102,493	4115447	9.6	1
C1	68,448	4094862	6.6	1
C2	43,540	3991503	4.3	0
Pool of 4 isolates (M2, M3, M4, C1-2a)	601,802	4117083	53.3 (~13x per isolate)	5

Result of sequencing and mapping alignment showing number of reads generated in each run, the number of nucleotide bases aligned to the reference consensus genome, coverage depth and the number of SNPs detected after filtering.

Table 2

	SNP loci		
	1	2	3
Locus tag	AB57_2551	AB57_2001	AB57_1823
SNP Coordinate	2645863	2093446	1906419
Predicted Product	Two-component heavy metal response regulator	Hypothetical protein	Transcriptional regulator, AraC family
Predicted SNP Effect	Synonymous	Non-synonymous (E to V)	Premature termination at codon 203
	Alleles		
AB0057	C	A	G
M1	C	A	G
M2	**T**	A	G
M3	**T**	A	**T**
M4	**T**	A	G
C1	**T**	**T**	G
C2	**T**	A	G

SNP loci which vary between outbreak isolates are shown. The corresponding annotation for each locus is shown for the ancestral strain AB0057. Alleles in bold demonstrate variation from the ancestral state. The predicted effect of the SNP on each affected protein product is also shown.


Weeks
1
2
3
4
5
6
7
Bed 3
4
5
6
Main ITU
Trauma Unit
M2
TAG
TAG
C2
M1
CAG
C1
ITU 6 bedded unit
M3
TAT
C1
TTG
M4
TAG

Chapter 3

Whole-genome comparison of two *Acinetobacter baumannii* isolates from a single patient, where resistance developed during tigecycline therapy

Whole-genome comparison of two *Acinetobacter baumannii* isolates from a single patient, where resistance developed during tigecycline therapy

Michael Hornsey[1*†], Nick Loman[2†], David W. Wareham[1], Matthew J. Ellington[3], Mark J. Pallen[2], Jane F. Turton[4], Anthony Underwood[4], Tom Gaulton[4], Claire P. Thomas[5], Michel Doumith[4], David M. Livermore[4] and Neil Woodford[4]

[1]*Centre for Immunology & Infectious Disease, Blizard Institute, Barts and The London, Queen Mary's School of Medicine and Dentistry, 4 Newark Street, London, E1 2AT, UK;* [2]*Centre for Systems Biology, University of Birmingham, Edgbaston, B15 2TT, UK;* [3]*Clinical Microbiology and Public Health Laboratory, Health Protection Agency, Addenbrooke's Hospital, Hills Road, Cambridge, CB2 0QQ, UK;* [4]*Health Protection Agency Centre for Infections, 61 Colindale Avenue, London, NW9 5EQ, UK;* [5]*Hammersmith Hospital, Imperial College Healthcare NHS Trust, Du Cane Road, London, W12 0HS, UK*

[*]Corresponding author. Tel: +44-20-7882-2312; Fax: +44-20-7882-2181; E-mail: m.hornsey@qmul.ac.uk

[†]These authors contributed equally to this work.

Running title: Genome sequencing of *A. baumannii* from a patient treated with tigecycline

Keywords: OXA-23 clone 1; glycylcycline resistance; comparative genomics

***Objectives*:** The whole genomes of two *Acinetobacter baumannii* isolates recovered from a single patient were sequenced to gain insight into the nature and extent of genomic plasticity in this important nosocomial pathogen over the course of a short infection. The first, AB210, was recovered before tigecycline therapy and was susceptible to this agent; the second, AB211, was recovered after therapy and was resistant.

***Methods*:** DNA from AB210 was sequenced by 454 GS FLX pyrosequencing according to the standard protocol for whole-genome shotgun sequencing, producing ~250-bp fragment reads. AB211 was shotgun-sequenced using the Illumina Genetic Analyzer to produce fragment reads of exactly 36-bp. Single nucleotide polymorphisms (SNPs) and large deletions detected in AB211 in relation to AB210 were confirmed by PCR and DNA sequencing.

***Results*:** Automated gene-prediction detected 3,850 putative coding sequences (CDS). Sequence analysis demonstrated the presence of plasmids pAB0057 and pACICU2 in both isolates. Eighteen putative SNPs were detected between the pre- and post-therapy isolates, AB210 and AB211. Three contigs in AB210 were not covered by reads in AB211, representing three deletions of approximately 15, 44 and 17 kb.

***Conclusions*:** This study demonstrates that significant differences were detectable between two bacterial isolates recovered one week apart from the same patient, and reveals the potential of whole-genome sequencing as a tool for elucidating the processes responsible for changes in antibiotic susceptibility profiles.

Introduction

Acinetobacter baumannii is an important nosocomial pathogen, with multidrug-resistant (MDR) and even pan-drug-resistant strains reported world-wide.[1] In the UK, carbapenem-resistant clonal lineages limit available treatment options. One successful lineage, designated OXA-23 clone 1, belonging to European clone II, has been recovered from over 60 hospitals, clustered mainly in London and South-East England.[2] Representative isolates of this clone are usually susceptible to colistin and tigecycline only. We previously reported the emergence of tigecycline resistance during antibiotic therapy in the OXA-23 clone 1 epidemic lineage, and showed that increased expression of the resistance-nodulation-division (RND) efflux system, AdeABC was responsible for the resistance phenotype.[3]

The recent availability of rapid and inexpensive whole-genome sequencing permits detailed investigation of genetic differences between pairs of bacterial isolates. In *A. baumannii* whole-genome studies have thus far focused either on comparing distinct antibiotic-susceptible and MDR strains,[4,5] or related isolates from different patients.[6] The results of these and other similar studies[7] point to a high degree of genome plasticity, the rapid emergence of antibiotic resistance, and considerable genetic variability even among closely-related isolates.

Tigecycline is used as a treatment of last resort for MDR *A. baumannii* infection, despite a lack of formal trial data and the emergence of resistance is a major concern. We sequenced the genomes of two *A. baumannii* isolates from a single patient, the first recovered before tigecycline therapy and susceptible to this agent, the second after one week of therapy for an intra-abdominal infection and resistant. The study aimed to gain insight into the nature and extent of genomic plasticity over the course of a short infection.

Materials and Methods

Bacterial isolates

Clinical isolates AB210 and AB211 have been described previously.[3] As OXA-23 clone 1 representatives, they belong to the globally successful European clone II group, and were assigned to Group 1 by the multiplex PCR method described by Turton *et al*.[8] They were typed by PFGE of *Apa*I-digested genomic DNA (Figure 1), as described previously,[2] and the presence of $bla_{\text{OXA-23-like}}$ was confirmed by multiplex PCR.[9]

Antimicrobial susceptibility testing and DNA manipulations

MICs were determined by BSAC agar dilution or Etest (AB bioMérieux, Solna, Sweden) on IsoSensitest agar (Oxiod, Basingstoke, UK) with the results interpreted according to BSAC guidelines.[9] Genomic DNA was extracted with the Wizard Genomic DNA Purification Kit (Promega, Southampton, UK) and was used as template for DNA sequencing. Plasmids were isolated from AB210 and AB211 using the PureYield Plasmid Miniprep System (Promega) and analysed by agarose gel electrophoresis.

Whole-genome DNA sequencing and data analysis

DNA from AB210 was sequenced by 454 GS FLX pyrosequencing (Roche, Branford, Connecticut, USA) according to the standard protocol for whole-genome shotgun sequencing, producing ~250 bp fragment reads. AB211 was shotgun sequenced using the Illumina Genetic Analyzer (Illumina, Saffron Walden, UK) to produce fragment reads of exactly 36-bp. All sequencing was performed at GATC Biotech Ltd (Constance, Germany). A draft genome assembly for AB210 was produced from flowgram data, using Newbler 2.5 (Roche). The Newbler command-line option '-rip' was used to ensure reads were aligned to single contigs

only. The resulting contigs were annotated by reference to the related strain *A. baumannii* ACICU[10] (also belonging to European clone II) using the automated annotation pipeline on the xBASE server.[11]

Illumina reads for isolate AB211 were mapped against the draft AB210 assembly using Bowtie 0.12.0.[12] For the purposes of single nucleotide polymorphism (SNP) detection, Bowtie was run with parameter '-m 0' to suppress alignments that map equally to multiple locations in the genome. To detect deletions this setting was not used. A consensus pileup was produced using SAMtools,[13] and putative SNPs were called using Varscan 2.2[14] with the following parameters: minimum coverage (10), min-reads2 (2), min-avg-qual (15), min-var-freq (0.9). To detect microindels (insertion or deletion events) less than 3-bases long, AB211 reads were additionally mapped using Novoalign 2.5.[15] Whole-genome alignments were visualised and SNPs and deletions manually inspected using the output files from the above steps using BAMview. [16]

Confirmation of SNPs and chromosomal deletions

SNPs and deletions detected in AB211 in relation to AB210 were confirmed by PCR and DNA sequencing using the primers listed in Table S1. Nucleotide sequences of the resulting amplicons were determined with an ABI 3730*xl* DNA analyser (Applied Biosystems, Warrington, UK).

Results & Discussion

Antibiotic susceptibilities

MICs of tigecycline, tobramycin, amikacin, gentamicin and azithromycin for the pre-therapy isolate AB210 were 0.5, >32, >64, >32 and >256 mg/L, respectively, while MICs for the post-therapy isolate AB211 were 16, 2, 4, 8 and >256 mg/L, respectively.

Sequencing results

Sequencing produced >128 million and >156 million sequence reads for AB210 and AB211, respectively. The assembly of AB210 resulted in 91 contigs larger than 500-bp, comprising 4.06 megabases of sequence and representing a median 29-fold coverage. Automated gene-prediction detected 3,850 putative coding sequences (CDS), of which 3,504 were homologous (defined as BLASTP e-value ≤ 1e-05) to a sequence in the reference genome of *A. baumannii* ACICU. The vast majority (96.6 %) of the AB211 reads mapped to a region on the AB210 genome. The AB210 draft assembly has been deposited in GenBank (accession number: AEOX00000000) and raw sequence reads for AB210 and AB211 have been submitted to NCBI's Sequence Read Archive under Study Accession Number SRP004860.

Plasmid profile

Plasmid profiles of AB210 and AB211 were identical and showed the presence of two plasmids in each isolate (data not shown). Sequence analysis demonstrated the presence of a 9-kb contig in AB210 which displayed 99.98 % identity to the previously characterised pAB0057 plasmid.[5] This was seen at high sequence read coverage in both AB210 and AB211, suggesting it was present as multiple copies. Three other contigs, totalling 65 kb, were seen at below-average

coverage; taken together these were a full match in length and nucleotide identity to the complete pACICU2 plasmid.[10]

AB210 virulence genes and resistance islands

Resistance islands (RIs) have been detected in all sequenced *A. baumannii* genomes containing multiple resistance determinants. They are composite transposons that are complex in nature and which have been designated AbaR (*A. baumannii* resistance).[4] They share a common insertion site (*comM*) but vary considerably among isolates in terms of the exact genetic composition, with that from ACICU, a representative of European clone II being considerably reduced in size compared to those found in representatives of European clone I.[10,17] Clinical isolates AB210 and AB211 were found to contain an AbaR-type RI. In the former isolate (GenBank accession number HQ700358) this was shown to contain sequence corresponding to nucleotides 587330-599047 of strain AB0057 (GenBank accession number CP001182), with a 2.85 kb section absent; this is an AbaR4-type island, and contains $bla_{\text{OXA-23}}$.

SNPs between AB210 and AB211

Eighteen putative SNPs were detected between the pre- and post-therapy isolates. Only one of these was located outside of coding regions at -35 bp upstream of *ureJ* which encodes a hydrogenase/urease accessory protein (AB210 locus tag: AB210-1_2203). The location of this SNP suggests the possibility of regulatory significance although *ureJ* appears to be part of a urease gene cluster which is co-transcribed as an operon in other species.[18] Of the remaining 17, eight were synonymous mutations whereas nine were non-synonymous including one missense mutation (Table 1). Seventeen (94 %) of the SNPs were transitions. Eight of the nine non-

synonymous SNPs could be confirmed by PCR and sequencing while one was not validated (Table 1 and Table S1). Several of these were located within genes predicted to be involved in core biological functions, including translation (*dusB*), nucleic acid biosynthesis, α-ketoglutarate and arabinose transport, environmental sensing (the signal transduction histidine kinase gene, *adeS* which had previously been identified through a candidate-gene approach[3]), and signalling. The mutation in *adeS* is believed to be responsible for up-regulation of the AdeABC efflux system and hence tigecycline resistance. Two SNPs were located within a gene coding for a GGDEF domain-containing protein, one of which was a non-synonymous mutation whilst the other introduced an internal stop codon, thus giving rise to a truncated product (Table 1). These proteins are enzymes that catalyze the synthesis of cyclic-di-GMP, which has been recognized recently as an important second messenger in bacteria and is implicated in adhesin and extrapolysaccharide biosynthesis.[19]

Large structural changes in the genomes of AB210 and AB211

Three contigs in AB210 were not covered by reads in AB211, these putative deletions were designated ROD1, 2 and 3. The first, ROD1, was approximately 15 kb in length. This deletion disrupted the coding sequence of the DNA mismatch repair gene *mutS* (AB210-1_2445) by eliminating the N-terminal *mutS-I* domain. Aside from encoding this mismatch recognition enzyme, ROD1 also encoded a DMT superfamily permease (AB210-1_2447) and an MFS permease (AB210-1_2451), transcriptional regulators (AB210-1_2450; AB210-1_2453), an EAL domain-containing protein (AB210-1_2448), responsible for the degradation of cyclic-di-GMP.[19] At approximately 44 kb ROD2 was the largest deleted region and comprised of genes encoding for transcriptional regulators (AB210-1_3253; AB210-1_3262; AB210-1_3269; AB210-1_3273), ion channels and transporters (AB210-1_3254; AB210-1_3259; [AB210-1_3275;

AB210-1_3276; AB210-1_3277]), a class A β-lactamase enzyme (AB210-1_3248) and components of a type VI secretion system (AB210-1_3280; AB210-1_3281).[20] Interestingly, part of the type VI secretion locus was missing even in AB210, suggesting that this was a degenerate system in both isolates. ROD1 and ROD2 are contiguous in *A. baumannii* ACICU, suggesting this may be a single deletion, but this could not be confirmed experimentally for AB210 by PCR (data not shown). ROD3, approximately 17 kb in length, included a class 1 integron containing antibiotic resistance genes including macrolide resistance determinants (AB210-1_3691 [phosphotransferase]; AB210-1_3692 [an efflux protein]) and several genes encoding aminoglycoside resistance determinants, namely *aac*(*6'*)-*Ib* (AB210-1_3701), two copies of *aadA* (AB210-1_3699; AB210-1_3700) and *armA* (AB210-1_3695), which encodes a 16S rRNA methylase.

Implications for Acinetobacter *evolution*

The extent of genomic changes detected here are consistent with the marked changes in phenotype, particularly the loss of aminoglycoside resistance in AB211. However, we were unable to determine whether these changes were the result of rapid evolution during the course of infection and treatment, or whether the patient initially had a mixed infection (or re-infection), involving different variants of the same defined clone, with subsequent selection for tigecycline resistance.

The disruption of *mutS*, an important DNA mismatch repair gene, is significant and suggests the possibility of a hypermutator phenotype, which may have contributed to the relatively large number of SNPs. Previous work in *Acinetobacter* sp. ADP1 has shown that *mutS*

196 preferentially recognises and repairs transitions,[21] so its disruption in AB211 is consistent with
197 our observation that 94 % of the SNPs belonged to this class.

198 The absence of ROD3 is consistent with the change in aminoglycoside resistance
199 between AB210 and AB211, with MICs of tobramycin, amikacin and gentamicin reduced at
200 least 8-fold in AB211. It is notable that the development of tigecycline resistance was
201 accompanied by increased susceptibility to other antibiotics through a large genomic deletion.

202 GGDEF and EAL-containing proteins have been implicated in sessile to planktonic
203 shifts. Taken together, the termination in a GGDEF domain-containing protein as well as the loss
204 of an EAL-domain containing protein in ROD1 may be advantageous during the process of
205 infection though this remains to experimentally determined.

206 In this study, whole-genome sequencing gave insight into the nature of genetic changes
207 between isolates under selection pressure through antibiotic therapy and a hostile host
208 environment. This study has demonstrated significant differences between two *A. baumannii*
209 isolates belonging to the same epidemic lineage, collected one week apart from the same patient.
210 Such studies are able to shed light on the relative importance of SNPs and transposon
211 mutagenesis on the evolution of *A. baumannii* and can generate hypotheses into the nature of
212 antibiotic resistance and virulence. Although further studies are needed to assess the extent of
213 genetic diversity among populations of *A. baumannii* in a single patient, we clearly demonstrated
214 the potential of whole-genome sequencing as an important tool for helping elucidate the
215 evolutionary processes responsible for the rapid development of antibiotic resistance in this
216 important nosocomial pathogen.

217

Acknowledgements

We wish to thank Anthony Haines, University of Birmingham for advice on bioinformatic analysis.

Funding

This work was supported by an educational grant from Wyeth, now taken over by Pfizer.

Transparency Declarations

D. M. L. has (i) received research grants from Wyeth and Pfizer, (ii) spoken at meetings organised by Wyeth and Pfizer, (iii) received sponsorship to travel to congresses from Wyeth and Pfizer, as well as from numerous other pharmaceutical and diagnostic companies. He holds shares in GlaxoSmithKline, Merck, AstraZeneca, Dechra and Pfizer; he acts also as Enduring Attorney for a close relative, managing further holdings in GlaxoSmithKline and EcoAnimal Health. N. W. has received research grants from Wyeth. M. E., M. D., J. F. T., A. U., T. G., D. M. L. and N. W. are employees of the HPA and are influenced by its views on antibiotic use and prescribing. M. H., D. W. W. and C. P. T. have received sponsorship to attend conferences from Wyeth. N. L. and M. J. P. : none to declare.

References

1. Towner KJ. *Acinetobacter*: an old friend, but a new enemy. *J Hosp Infect* 2009; **73**: 355-63.

2. Coelho JM, Turton JF, Kaufmann ME, *et al*. Occurrence of carbapenem-resistant *Acinetobacter baumannii* clones at multiple hospitals in London and Southeast England. *J Clin Microbiol* 2006; **44**: 3623-7.

3. Hornsey M, Ellington MJ, Doumith M, *et al*. AdeABC-mediated efflux and tigecycline MICs for epidemic clones of *Acinetobacter baumannii*. *Journal of Antimicrobial Chemotherapy* 2010; **65**: 1589-93.

4. Fournier PE, Vallenet D, Barbe V, *et al*. Comparative genomics of multidrug resistance in *Acinetobacter baumannii*. *PLoS Genet* 2006; **2**: e7.

5. Adams MD, Goglin K, Molyneaux N, *et al*. Comparative genome sequence analysis of multidrug-resistant *Acinetobacter baumannii*. *J Bacteriol* 2008; **190**: 8053-64.

6. Adams MD, Chan ER, Molyneaux ND, *et al*. Genomewide analysis of divergence of antibiotic resistance determinants in closely related isolates of *Acinetobacter baumannii*. *Antimicrob Agents Chemother* 2010; **54**: 3569-77.

7. Adams MD, Nickel GC, Bajaksouzian S, *et al*. Resistance to colistin in *Acinetobacter baumannii* associated with mutations in the PmrAB two-component system. *Antimicrobial Agents and Chemotherapy* 2009; **53**: 3628-34.

8. Turton JF, Gabriel SN, Valderrey C, *et al*. Use of sequence-based typing and multiplex PCR to identify clonal lineages of outbreak strains of *Acinetobacter baumannii*. *Clin Microbiol Infect* 2007; **13**: 807-15.

9. Andrews JM. BSAC standardized disc susceptibility testing method. Version 9.1, March 2010. http://www.bsac.org.uk/Resources/BSAC/Version_9.1_March_2010_final.pdf (22 March 2011, date last accessed).

10. Iacono M, Villa L, Fortini D, *et al*. Whole-genome pyrosequencing of an epidemic multidrug-resistant *Acinetobacter baumannii* strain belonging to the European clone II group. *Antimicrobial Agents and Chemotherapy* 2008; **52**: 2616-25.

11. Chaudhuri RR, Loman NJ, Snyder LA, *et al*. xBASE2: a comprehensive resource for comparative bacterial genomics. *Nucleic Acids Res* 2008; **36**: D543-D546.

12. Langmead B, Trapnell C, Pop M, *et al*. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009; **10**: R25.

13. Li H, Handsaker B, Wysoker A, *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; **25**: 2078-9.

14. Koboldt DC, Chen K, Wylie T, *et al*. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 2009; **25**: 2283-5.

15. Krawitz P, Rodelsperger C, Jager M, *et al*. Microindel detection in short-read sequence data. *Bioinformatics* 2010; **26**: 722-9.

16. Carver T, Bohme U, Otto TD, *et al*. BamView: viewing mapped read alignment data in the context of the reference sequence. *Bioinformatics* 2010; **26**: 676-7.

17. Post V, White PA, Hall RM. Evolution of AbaR-type genomic resistance islands in multiply antibiotic-resistant *Acinetobacter baumannii*. *Journal of Antimicrobial Chemotherapy* 2010; **65**: 1162-70.

18. McMillan DJ, Mau M, Walker MJ. Characterisation of the urease gene cluster in *Bordetella bronchiseptica*. *Gene* 1998; **208**: 243-51.

19. Hengge R. Principles of c-di-GMP signalling in bacteria. *Nat Rev Microbiol* 2009; **7**: 263-73.

20. Bingle LE, Bailey CM, Pallen MJ. Type VI secretion: a beginner's guide. *Curr Opin Microbiol* 2008; **11**: 3-8.

21. Young DM, Ornston LN. Functions of the mismatch repair gene *mutS* from *Acinetobacter* sp. strain ADP1. *J Bacteriol* 2001; **183**: 6822-31.

Table 1. Confirmed SNPs indentified in clinical isolate AB211 resulting in amino acid substitution or termin

SNP	Position in AB210 assembly	Locus tag in AB210 assembly	Protein product	Amino acid identity	
				AB210	AB211
1	159509	AB210-1_0138	tRNA-dihydrouridine synthase, DusB	A	T
2	639321	AB210-1_0587	nucleoside-diphosphate-sugar epimerase	T	A
3	755474	AB210-1_0703	major facilitator superfamily permease	V	A
4	1469178	AB210-1_1405	hypothetical protein	A	V
5	2548057	AB210-1_2423	major facilitator superfamily permease	A	T
6	2852737	AB210-1_2721	Signal transduction histidine kinase, AdeS	A	V
7	3362158	AB210-1_3207	GGDEF domain-containing protein	Q	*
8	3362175	AB210-1_3207	GGDEF domain-containing protein	G	V

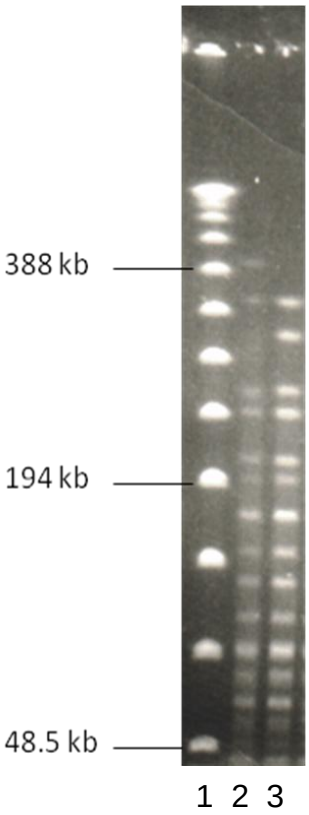


Figure 1.

Figure Legands

Figure 1. PFGE profiles of AB210 (lane 2) and AB211 (lane 3).

Chapter 4

Open-Source Genomic Analysis of Shiga-Toxin Producing *E. coli* O104:H4

The NEW ENGLAND JOURNAL of MEDICINE

BRIEF REPORT

Open-Source Genomic Analysis of Shiga-Toxin–Producing *E. coli* O104:H4

Holger Rohde, M.D., Junjie Qin, Ph.D., Yujun Cui, Ph.D., Dongfang Li, M.E., Nicholas J. Loman, M.B., B.S., Moritz Hentschke, M.D., Wentong Chen, B.S., Fei Pu, B.S., Yangqing Peng, B.S., Junhua Li, B.E., Feng Xi, B.E., Shenghui Li, B.S., Yin Li, B.S., Zhaoxi Zhang, B.S., Xianwei Yang, B.S., Meiru Zhao, M.S., Peng Wang, B.M., Yuanlin Guan, B.E., Zhong Cen, M.E., Xiangna Zhao, B.S., Martin Christner, M.D., Robin Kobbe, M.D., Sebastian Loos, M.D., Jun Oh, M.D., Liang Yang, Ph.D., Antoine Danchin, Ph.D., George F. Gao, Ph.D., Yajun Song, Ph.D., Yingrui Li, B.S., Huanming Yang, Ph.D., Jian Wang, Ph.D., Jianguo Xu, M.D., Ph.D., Mark J. Pallen, M.D., Ph.D, Jun Wang, Ph.D., Martin Aepfelbacher, M.D., Ruifu Yang, M.D., Ph.D., and the *E. coli* O104:H4 Genome Analysis Crowd-Sourcing Consortium*

SUMMARY

An outbreak caused by Shiga-toxin–producing *Escherichia coli* O104:H4 occurred in Germany in May and June of 2011, with more than 3000 persons infected. Here, we report a cluster of cases associated with a single family and describe an open-source genomic analysis of an isolate from one member of the family. This analysis involved the use of rapid, bench-top DNA sequencing technology, open-source data release, and prompt crowd-sourced analyses. In less than a week, these studies revealed that the outbreak strain belonged to an enteroaggregative *E. coli* lineage that had acquired genes for Shiga toxin 2 and for antibiotic resistance.

Escherichia coli is a widespread commensal of the mammalian gut and a versatile pathogen.[1,2] Enterovirulent strains of *E. coli* are classified into a number of overlapping pathotypes, which include Shiga-toxin–producing, enterohemorrhagic, and enteroaggregative varieties.[2] Enteroaggregative *E. coli* strains have been associated with sporadic and epidemic diarrhea and, in the laboratory, show a distinctive pattern of adherence to Hep-2 cells (termed aggregative, or "stacked brick").[3] In Shiga-toxin–producing *E. coli*, the toxin is encoded on a prophage and inhibits protein synthesis within susceptible eukaryotic cells. Strains of enterohemorrhagic *E. coli* produce Shiga toxin and a specific protein secretion system (called a type III secretion system) that is encoded by the locus of enterocyte effacement (LEE) and that is responsible for attachment to the intestine.[2] Shiga-toxin–producing and enterohemorrhagic *E. coli* strains are commonly associated with the hemolytic–uremic syndrome, a combination of renal impairment, thrombocytopenia, and hemolytic anemia that is often accompanied by neurologic and myocardial damage.

The authors' affiliations are listed in the Appendix. Address reprint requests to Dr. Pallen at the Centre for Systems Biology, University of Birmingham, B15 2TT, United Kingdom, or at m.pallen@bham.ac.uk; or to Dr. Yang at BGI-Shenzhen, Shenzhen 518083, China, or at the Beijing Institute of Microbiology and Epidemiology, 20 Dongda St., Beijing 100071, China, or at yangruifu@genomics.org.cn.

The following two groups of authors contributed equally to this article: Drs. Rohde, Qin, Cui, D. Li, and Loman; and Drs. Pallen, J. Wang, Aepfelbacher, and R. Yang.

*Members of the *E. coli* O104:H4 Genome Analysis Crowd-Sourcing Consortium are listed in the Supplementary Appendix, available at NEJM.org.

This article (10.1056/NEJMoa1107643) was published on July 27, 2011, at NEJM.org.

N Engl J Med 2011.

More than 3000 cases of infection with an unusual strain of Shiga-toxin–producing *E. coli* O104:H4 were reported to the Robert Koch Institute in Berlin during a nationwide outbreak in Germany in May and June of 2011.[4] This outbreak resulted in more than 40 deaths, and associated cases were reported in more than a dozen countries in Europe and North America (mostly in travelers returning from Germany). Household transmission was described in the Netherlands, and life-threatening colonic ischemia was reported as a complication in addition to the hemolytic–uremic syndrome and bloody diarrhea.[5,6] Epidemiologic and microbiologic evidence indicated that the O104:H4 strain was distributed throughout Germany on bean sprouts.[7]

The outbreak was characterized by several unusual features: a high incidence in adults (especially women), a greatly increased incidence of the hemolytic–uremic syndrome (in approximately 25% of patients, as compared with 1 to 15% in previous outbreaks of Shiga-toxin–producing *E. coli*), a predominance of female patients among cases of the hemolytic–uremic syndrome, and a rare serotype of Shiga-toxin–producing *E. coli* that had been linked to only two sporadic cases of the hemolytic–uremic syndrome (one in Germany and the other in South Korea).[4,8,9] Recognition of infection during the outbreak was hampered by a laboratory approach that targeted phenotypes associated with the most common lineage of enterohemorrhagic *E. coli* (the non–sorbitol-fermenting O157:H7 serotype) rather than one aimed at finding all strains of Shiga-toxin–producing *E. coli*.[10] Here, we report a local cluster of cases associated with a family from northern Germany and describe an open-source genomic analysis of an isolate from the family cluster.

CASE REPORTS

On May 17, 2011, a 16-year-old girl was admitted to the pediatric emergency ward at the University Medical Center Hamburg–Eppendorf with bloody diarrhea and abdominal pain. Her laboratory values were normal. Later on the same day, her 12-year-old brother was admitted with a 2-day history of malaise and headache and a 1-day history of vomiting and nonbloody diarrhea. The boy presented with acute renal failure (serum creatinine level, 4.1 mg per deciliter [362 μmol per liter]; and potassium level, 6 mmol per liter), thrombocytopenia (22,000 platelets per cubic millimeter), and hemolytic anemia (hemoglobin, 11.6 g per deciliter; bilirubin, 2.8 mg per deciliter [49 μmol per liter]; and lactate dehydrogenase, 2297 U per liter). His hemoglobin level fell to 8.4 g per deciliter within 48 hours after admission, thereby fulfilling the case definition of the hemolytic–uremic syndrome.

The children, their parents, and a teenage friend had eaten a meal together a week earlier. The meal included a freshly prepared salad containing bean sprouts. The children's mother had no symptoms, and no Shiga-toxin–producing *E. coli* was isolated from her stool. However, the hemolytic–uremic syndrome developed in the father, and his stool sample was culture-positive for Shiga-toxin–producing *E. coli*. The teenage friend had diarrhea but was not admitted to the medical center.

Stool samples from the siblings were plated on Sorbitol–MacConkey agar and incubated in a liquid enrichment culture. The next day, supernatants from the liquid cultures tested positive for Shiga toxin on enzyme-linked immunosorbent assay. Uniformly sorbitol-positive colonies were identified as *E. coli* on MALDI-TOF (matrix-assisted laser desorption ionization–time of flight) mass spectrometry. Several single colonies were positive for the *stx2* gene and negative for the *stx1* and *eae* genes on polymerase-chain-reaction (PCR) assay. None of the isolates agglutinated with polyvalent serum samples directed against the serotypes that are most frequently associated with Shiga-toxin–producing *E. coli*. Subsequent analyses showed that the strain belonged to the rare serotype O104:H4 harboring an extended-spectrum beta-lactamase (ESBL) gene of the CTX-M-15 class.

Although our 16-year-old patient had a mild course of disease without the hemolytic–uremic syndrome and was discharged from the hospital on the same day, the clinical picture for her brother was much less benign. The boy's renal function, hemoglobin level, and thromobocytopenia improved after 9 days of peritoneal dialysis, but severe neurologic symptoms, including somnolence, visual impairment, speech disturbances, hemiplegia, and incontinence, developed. He underwent four cycles of plasmapheresis and therapy with the anti–C5-antibody eculizumab. After this treatment, his clinical condition improved, and he was discharged after 24 days with serum creatinine levels just above the normal range. However, he was left with neurologic sequelae and required rehabilitation.

10.1056/NEJMoa1107643 NEJM.ORG

METHODS AND RESULTS

OPEN-SOURCE GENOMICS

To investigate the evolutionary origins and pathogenic potential of the outbreak strain, we set in motion an open-source genomics program of research that incorporated new high-throughput sequencing approaches, public data release, and rapid outsourcing of analyses to bioinformaticians worldwide (crowd-sourcing) (Fig. 1). Initially, we sequenced the genome of the isolate from the 16-year-old girl (TY2482), using the Ion Torrent Personal Genome Machine (PGM), and obtained an initial draft of the genome 3 days after receipt of the DNA sample. Three DNA libraries were prepared and seven sequencing runs performed, following the protocols of the manufacturer (Life Technologies), to generate 79 Mb of sequence data, with an average read length of 101 bp. (For details regarding the sequencing procedures, see the Supplementary Appendix, available with the full text of this article at NEJM.org.)

We released these data into the public domain under a Creative Commons 0 license, which elicited a burst of crowd-sourced, curiosity-driven analyses carried out by bioinformaticians on four continents.[11] Twenty-four hours after the release of the genome, it had been assembled; 2 days after its dissemination, it had been assigned to an existing sequence type. Five days after the release of the sequence data, we had designed and released strain-specific diagnostic primer sequences, and within a week, two dozen reports had been filed on an open-source wiki (a Web site that facilitates collaborative effort) dedicated to analysis of the strain. These analyses provided timely information on the strain's virulence and resistance genes, in addition to its phylogenetic lineage.

We also performed sequencing on the Illumina HiSeq platform in accordance with the manufacturer's instructions. An initial single-end run was used to correct errors in the Ion Torrent sequence, principally in homopolymeric tracts. We later performed paired-end and mate-pair sequencing on this platform, exploiting libraries with insert sizes of 470 bp, 2 kb, and 6 kb, and generated enough data (1 Gb, 576 Mb, and 576 Mb from each library, respectively) to create a high-quality draft genome sequence within 2 weeks after receipt of the DNA samples. (Additional details are provided in the Supplementary Appendix.) The reads were deposited in GenBank's Short Read Archive with accession numbers SRA037315 for Ion Torrent reads and SRA039136 for Illumina platform reads.

PHYLOGENETIC ANALYSIS

The assembled Ion Torrent data provided gene sequences that could be analyzed with an existing multilocus-sequence-typing scheme for *E. coli* that relied on sequence comparisons for seven conserved housekeeping genes (*adk*, *fumC*, *gyrB*, *mdh*, *purA*, *recA*, and *icd*).[12] This analysis revealed a close relationship to a strain, 01-09591, which was isolated in Germany in 2001 and which fell into sequence type ST678. The TY2482 sequences differed from the profile of the 2001 strain by a single base pair in the *adk* gene and a single-base difference in a homopolymeric sequence in the *recA* gene. (We subsequently discovered that the latter difference was a sequencing error generated by the PGM.) The 2001 strain, which produced Shiga toxin and was associated with the hemolytic–uremic syndrome, fell into the O104:H4 serotype but did not have the genes associated with type III secretion in typical enterohemorrhagic *E. coli*.[13,14] Additional scrutiny of the multilocus-sequence-typing database revealed that strains with the broad O104 serotype were scattered across several sequence types, whereas strains with the narrower O104:H4 serotype appeared to be limited to ST678.[10]

Comparisons of the TY2482 genome with all previously sequenced complete genomes of *E. coli* isolates revealed a very close relationship to *E. coli* strain 55989, with an average nucleotide identity of 99.8% (see the Supplementary Appendix). This strain was isolated in the Central African Republic from a stool sample obtained from an adult with human immunodeficiency virus infection who had persistent watery diarrhea.[15] It has been classified as an enteroaggregative *E. coli*, but unlike TY2482, it does not have Shiga toxin genes.[15] However, it is worth noting that Mossoro et al.,[15] who first described *E. coli* strain 55989, also described strains of enteroaggregative *E. coli* with Shiga toxin genes in the same human population.[15]

COMPARISON OF THE CHROMOSOMES OF TY2482 AND 55989

Isolates from the German outbreak were initially described as enterohemorrhagic *E. coli*. However, the close relationship between TY2482 and 55989 led us to consider the likelihood that TY2482 is an enteroaggregative *E. coli*. Our analysis of the gene content of TY2482 showed that it, like 01-09591,

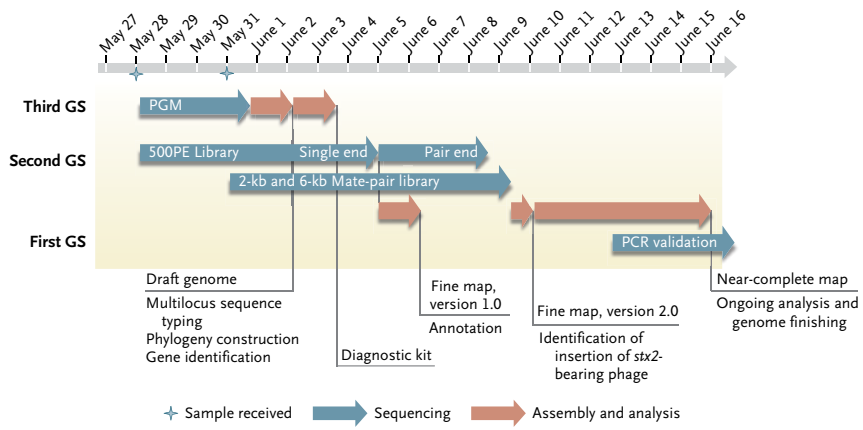


Figure 1. Timeline of the Open-Source Genomics Program.

After receiving the first batch of DNA samples on May 28, 2011, sequencing runs with the use of the Ion Torrent Personal Genome Machine (PGM) and Illumina (small-insert library) were initiated simultaneously. On May 31, the second batch of DNA was received and used for Illumina large-insert sequencing. An assembly of the Ion Torrent reads was released on June 2, which enabled subsequent analyses (multilocus sequence typing, phylogenetic analysis, and genome comparisons). Errors in the Ion Torrent data were corrected with the use of later Illumina data, and a high-quality draft genome sequence was created. GS denotes generation of sequencing technology. The symbols at May 28 and May 31 in the timeline indicate the arrival of DNA samples.

lacked the LEE and genes encoding effectors associated with type III secretion.[16] Instead, we found that the TY2482 genome encodes virulence factors that are typical of enteroaggregative *E. coli*. Other investigators working on the outbreak strain have also observed genes typically found in enteroaggregative strains on PCR assay and have noted a behavioral phenotype that is characteristic of this pathotype on cell-adherence assay.[17]

To identify strain-specific genes, we performed a detailed comparison of the chromosomes of TY2482 and enteroaggregative *E. coli* strain 55989. First, we aligned the TY2482 assembly against the 55989 chromosome (for details, see the Supplementary Appendix). We then adopted the gene predictions and annotation from the 55989 genome for these conserved sequences. Next, we identified several isolate-specific regions of difference (i.e., regions present in the TY2482 chromosome and absent from the 55989 genome or vice versa) that were more than 5 kb (Table 1 and Fig. 2, and the Supplementary Appendix). TY2482-specific regions of difference included prophage remnants or apparently intact prophages, such as the *stx2* prophage, which, like its close relatives in the genomes of O157:H7 strains EDL933 and Sakai, is inserted into the *wrbA* locus. The *stx2* genes differ by only one single-nucleotide polymorphism from the *stx2* allele seen in O157 enterohemorrhagic *E. coli* strain EDL933.

TY2482 PLASMIDS

From our de novo assembly (i.e., assembly without the use of a reference genome), we concluded that the TY2482 genome contains two large conjugative plasmids, pESBL TY2482 and pAA TY2482, and a small plasmid, pG2011 TY2482 (Fig. 2). From scrutiny of copy numbers of sequence reads, it was clear that the two large plasmids were replicating at an approximate ratio of 1:1 with the chromosome, whereas the small plasmid was maintained at a copy number at least nine times that of the other replicons. No phenotype could be ascribed to the small plasmid.

The largest plasmid, pESBL TY2482, was an IncI plasmid similar to pEC_Bactec, which was found in an *E. coli* strain isolated from the joint of a horse with arthritis.[18] The pESBL TY2482 plasmid encodes a CTX-M-15 ESBL, as well as a beta-lactamase from the TEM class. The second large plasmid, pAA TY2482, resembled a plasmid from strain 55989 but carried a gene cluster encoding a

10.1056/NEJMoa1107643 NEJM.ORG

Table 1. Genetic Elements in Strain TY2482 of Shiga-Toxin–Producing *Escherichia coli* O104:H4.

Genetic Element	Notable Features or Functions	Size or 55989 Coordinates*
Plasmid		
pESBL TY2482	IncI1 plasmid, homologous to pEC_Bactec carrying *bla* CTX-M-15	88 kb
pAA TY2482	Plasmid encoding aggregative adherence fimbriae I	76 kb
pG2011 TY2482	Plasmid with no obvious phenotype	1.5 kb
Region of difference		
I-ROD1	Degenerate prophage	296227 (tRNA-*Thr*)
I-ROD2	*Stx2*-encoding prophage	1176265 (*wrbA*)
I-ROD3	Microcin gene cluster; tellurite resistance gene cluster	1207704 (tRNA-*Ser*)
I-ROD4	Prophage	1811905 (*ynfG*)
I-ROD5	Prophage	2102453 (*yecE*)
I-ROD6	Molybdate metabolism regulator; *yehL*	2426442 (IS1)
I-ROD7	Multidrug-resistant gene cluster (*dfA7, sulI, sulII, strA, strB, tetA*); mercury resistance	4211244 (tRNA-*Sec*)
D-ROD1	Prophage	1094587–1140306
D-ROD2	Prophage	1413924–1446834
D-ROD3	Prophage	1754689–1800354
D-ROD4	Prophage	2688656–2701228
D-ROD5	Type VI secretion genes	3401720–3427357
D-ROD6	Prophage	4944269–5004333

* Coordinates from the genome of *E. coli* strain 55989 are given for predicted boundaries of regions of difference, with the gene carrying the insertion site shown in parentheses for a region of difference involving an insertion into 55989 (I-ROD). D-ROD denotes a region of difference involving a deletion.

rare type of aggregative adherence fimbria (AAF/I) instead of the more common type (AAF/III) encoded by genes in the 55989 plasmid. We exploited this AAF/I cluster as a target for strain-specific PCR primers as part of a suite of primers to identify the outbreak isolate.

DISCUSSION

Our genomic analyses suggest that the German outbreak strain evolved from a progenitor that belonged to the enteroaggregative pathotype and resembled strain 55989. The emergence of the outbreak strain depended on the acquisition of a *stx2* prophage and of a plasmid encoding a CTX-M-15 ESBL. Sometime during this process, the strain also appears to have lost one gene cluster, encoding AAF/III fimbriae, and gained another, encoding the rarer AAF/I fimbriae.

Although this outbreak strain has surprised the general public and public health officials, related potential progenitor strains have been reported from three continents. The appearance of an O104:H4 strain associated with the hemolytic–uremic syndrome in Korea in 2005 is unexplained, and its link to the German outbreak is unclear.[9] Also, the O104:H4 strain 01-09591 that was isolated in Germany in 2001 urgently requires further investigation. Both strains should undergo genome sequencing and comparison with TY2482. The link to strain 55989, which was isolated in the Central African Republic in the late 1990s, is also intriguing. Genome sequencing of additional Central African isolates from the study that yielded 55989 is likely to illuminate the evolution of this lineage and of enterovirulent *E. coli* in general (see the article by Rasko et al. elsewhere in this issue of the *Journal*[19]).

Although the genome sequence alone cannot provide a full explanation for the high degree of virulence of this strain, it prompts a reassessment of our assumptions and provides a framework for future hypothesis-driven research. Both commensal and pathogenic varieties of *E. coli* have to survive in the gut. However, mere survival, even if twinned with the production of Shiga toxin, is probably

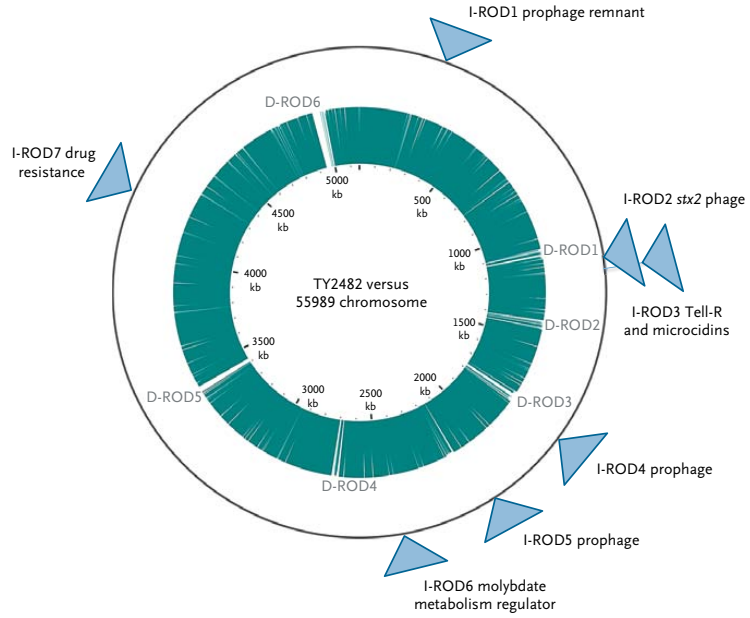


Figure 2. Comparison of the TY2482 and 55989 Genomes.

The outer circle depicts the *Escherichia coli* 55989 chromosome. The inner circle represents the TY2482 assembly mapped against the 55989 chromosome. Regions of difference (ROD) that are present in 55989 but not TY2482 (D-RODs) are shown as gaps in the inner circle. The positions of RODs that were found in TY2482 but not 55989 (I-RODs) are shown as wedges on the outer circle at positions corresponding to the predicted insertion sites. Tell-R denotes tellurite resistance.

not enough to cause the hemolytic–uremic syndrome or bloody diarrhea. For that, the bacteria would probably need to adhere to the gut mucosa. In the past, much research has been concentrated on the adhesion systems of typical enterohemorrhagic *E. coli*, particularly the LEE-encoded type III secretion system.[16,20] This German outbreak strain shows us that Shiga-toxin–producing *E. coli* can exploit alternative adhesion mechanisms, very likely including aggregative adherence fimbriae, to the same end. This strain also shows that pathotypes of *E. coli* can overlap and that they evolve rather than stand as fixed archetypes.

It remains unclear why this strain has proved to be so virulent. As noted, a novel suite of adhesins might provide an explanation. Alternatively, perhaps this strain exploits more efficient mechanisms for toxin release. It is worth remembering that strains of enteroaggregative *E. coli* have caused large sprout-associated outbreaks before, including one outbreak[21] that affected more than 2000 persons in Japan in 1993. Thus, there is clearly an urgent need to understand how the German outbreak strain and other strains of enteroaggregative *E. coli* adhere to and colonize seeds and seedlings.

Our rapid open-source analysis of an outbreak-associated bacterial pathogen was characterized by a propitious confluence of high-throughput genomics, crowd-sourced analyses, and a liberal approach to data release. Although phenotypic or molecular analyses that exploit known virulence, resistance, or epidemiologic targets are useful in diagnostic and public health microbiology, genome sequencing offers the advantages of open-endedness (revealing the "unknown unknowns"), universal applicability, and the ultimate in resolution. Our study shows how benchtop sequencing platforms can generate data with sufficient speed to have an important effect on clinical and epidemiologic problems.

10.1056/NEJMoa1107643 NEJM.ORG

Supported by grants from the State Key Development Program for Basic Research of China (2009CB522600), the National Key Program for Infectious Diseases of China (2008ZX10004-009), Shenzhen Biological Industry Development Special Foundation–Basic Research Key Projects (JC201005250088A), Key Laboratory Project Supported by Shenzhen City (ZD200806180054A), the European Union Microme Program (FP7-KBBE-2007-3-2-08-222886), the Alexander von Humboldt Foundation (to Dr. L. Yang), the Medical Faculty of the University Medical Center Hamburg–Eppendorf, and the British Biotechnology and Biological Sciences Research Council (BB/E011179/1).

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

We thank David Vallenet, Claudine Médigue, Xiaoning Wang, and Jennifer Gardy for their helpful discussions.

APPENDIX

The authors' affiliations are as follows: the Institute of Medical Microbiology, Virology and Hygiene (H.R., M.H., M.C., L.Y., M.A.) and the Department of Pediatrics (R.K., S. Loos, J.O.), University Medical Center Hamburg–Eppendorf, Hamburg, Germany; BGI-Shenzhen, Shenzhen, China (J.Q., Y.C., D.L., W.C., F.P., Y.P., J.L., F.X., S. Li, Yin Li, Z.Z., Z.C., Yingrui Li, H.Y., Jian Wang, Jun Wang, R.Y.), and the School of Bioscience and Biotechnology, South China University of Technology, Guangzhou, China (J.L.); the State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology (Y.C., X.Z., Y.S., R.Y.), CAS Key Laboratory of Pathogenic Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Science (G.F.G.), and the State Key Laboratory for Infectious Disease Prevention and Control and National Institute for Communicable Diseases Control and Prevention, Chinese Center for Disease Control and Prevention (G.F.G., J.X.) — all in Beijing; the Centre for Systems Biology, University of Birmingham, Birmingham, United Kingdom (N.J.L. M.J.P.); and AMAbiotics, Evry, France (A.D.).

REFERENCES

1. Hobman JL, Penn CW, Pallen MJ. Laboratory strains of Escherichia coli: model citizens or deceitful delinquents growing old disgracefully? Mol Microbiol 2007; 64:881-5.
2. Kaper JB, Nataro JP, Mobley HL. Pathogenic Escherichia coli. Nat Rev Microbiol 2004;2:123-40.
3. Weintraub A. Enteroaggregative Escherichia coli: epidemiology, virulence and detection. J Med Microbiol 2007;56:4-8.
4. Frank C, Werber D, Cramer JP, et al. Epidemic profile of Shiga-toxin–producing *Escherichia coli* O104:H4 outbreak in Germany — preliminary report. N Engl J Med 2011. DOI: 10.1056/NEJMoa1106483.
5. Kuijper E, Soonawala D, Vermont C, van Dissel J. Household transmission of haemolytic uraemic syndrome associated with Escherichia coli O104:H4 in the Netherlands. Euro Surveill 2011;16:pii: 19897.
6. Cordesmeyer S, Peitz U, Godde N, Kasper H, Hoffmann M, Allemeyer E. Colonic ischaemia as a severe Shiga toxin/verotoxin producing Escherichia coli O104:H4 complication in a patient without haemolytic uraemic syndrome, Germany, June 2011. Euro Surveill 2011;16: pii:19895.
7. Robert Koch Institute. Information update on EHEC/HUS outbreak, 2011. (http://www.rki.de/nn_217400/EN/Home/PM082011.html.)
8. Scheutz F, Moller Nielsen E, Frimodt-Moller J, et al. 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. Euro Surveill 2011;16:pii:19889.
9. Bae WK, Lee YK, Cho MS, et al. A case of hemolytic uremic syndrome caused by Escherichia coli O104:H4. Yonsei Med J 2006;47:437-9.
10. Chattaway MA, Dallman T, Okeke IN, Wain J. Enteroaggregative E. coli O104 from an outbreak of HUS in Germany 2011, could it happen again? J Infect Dev Ctries 2011;5:425-36.
11. GitHub. *E. coli* O104:H4 genome analysis crowd sourcing, 2011. (https://github.com/ehec-outbreak-crowdsourced/BGI-data-analysis/wiki.)
12. University College Cork. *Escherichia coli* MLST Database, 2011. (http://mlst.ucc.ie/mlst/dbs/Ecoli.)
13. Mellmann A, Bielaszewska M, Kock R, et al. Analysis of collection of hemolytic uremic syndrome-associated enterohemorrhagic Escherichia coli. Emerg Infect Dis 2008;14:1287-90.
14. Creuzburg K, Middendorf B, Mellmann A, et al. Evolutionary analysis and distribution of type III effector genes in pathogenic Escherichia coli from human, animal and food sources. Environ Microbiol 2011;13:439-52.
15. Mossoro C, Glaziou P, Yassibanda S, et al. Chronic diarrhea, hemorrhagic colitis, and hemolytic-uremic syndrome associated with HEp-2 adherent Escherichia coli in adults infected with human immunodeficiency virus in Bangui, Central African Republic. J Clin Microbiol 2002;40: 3086-8.
16. Tobe T, Beatson SA, Taniguchi H, et al. An extensive repertoire of type III secretion effectors in Escherichia coli O157 and the role of lambdoid phages in their dissemination. Proc Natl Acad Sci U S A 2006;103:14941-6.
17. Bielaszewska M, Mellmann A, Zhang W, et al. Characterisation of the Escherichia coli strain associated with an outbreak of haemolytic uraemic syndrome in Germany, 2011: a microbiological study. Lancet Infect Dis 2011 June 22 (Epub ahead of print).
18. Smet A, Van Nieuwerburgh FV, Vandekerckhove TTM, et al. Complete nucleotide sequence of CTX-M-15-plasmids from clinical Escherichia coli isolates: insertional events of transposons and insertion sequences. PLoS ONE 2010;5(6):e11202.
19. Rasko DA, Webster DR, Sahl JW, et al. Origins of the *E. coli* strain causing an outbreak of hemolytic–uremic syndrome in Germany. N Engl J Med 2011. DOI: 10.1056/NEJMoa1106920.
20. Schmidt MA. LEEways: tales of EPEC, ATEC and EHEC. Cell Microbiol 2010;12: 1544-52.
21. Itoh Y, Nagano I, Kunishima M, Ezaki T. Laboratory investigation of enteroaggregative Escherichia coli O untypeable:H10 associated with a massive outbreak of gastrointestinal illness. J Clin Microbiol 1997;35:2546-50.

Supplementary Appendix: Open-source genomics of a Shiga-toxin-producing *Escherichia coli* O104:H4

Crowd-sourcing consortium

The following members of the *E. coli* O104:H4 Genome Analysis Crowd-sourcing consortium made contributions that influenced the analyses reported here: Kathryn E. Holt, David J. Studholme, Michael Feldgarden and Marina Manrique.

A full account of crowd-sourcing efforts can be accessed here: https://github.com/ehec-outbreak-crowdsourced/BGI-data-analysis/wiki/

Methods and Results

Ion Torrent library construction and sequencing

Genomic DNA was extracted and purified using a conventional SDS lysis and phenol-chloroform method. 5µg of DNA (OD260/OD280 = 1.85) was dissolved in TE buffer to a total volume of 100 µl and fragmented by sonication (Covaris S2, Massachusetts, USA) to a size distribution of 50-300 bp. Library preparation and template preparation of live Ion Sphere™ Particles was performed according to the manufacturer's protocol (Ion Torrent, USA). During the library preparation, nick-translation was followed by 5 cycles of PCR amplification. Finally the sequencing was performed on the PGM Sequencer. Seven 314 chips were run to generate 79.1 Mb of sequence, with average length of 101 bp.

Illumina library construction and sequencing

Whole-genome sequencing was performed using Illumina HiSeq 2000 (Illumina Inc. U.S.A) by generating paired-end libraries with an average insert size of 470 bp, 2 kbp and 6 kbp

following the manufacturer's instruction. The read lengths were 90bp, 50bp, 50bp and 1Gb, 576Mb and 576Mb high quality data were generated from each library respectively.

Creation of a draft genome assembly using Ion Torrent PGM data (2nd June 2011)

An assembly was performed using MIRA 3.2.1.17_dev using command-line parameters --*job=denovo,genome,accurate,iontor -GE:not=1*. The Ion Torrent PGM assembly from 5 chips of Ion Torrent 314 data produced an assembly of 3,057 contigs, total bases: 5,491,032 with an N50 value of 3,675.

Creation of a hybrid assembly using Ion Torrent PGM data and Illumina single-end data (6th June 2011)

Ion Torrent and Illumina read data were quality filtered before assembly including removal of adapter contamination. The Ion Torrent PGM assembly from 7 chips of Ion Torrent 314 data were assembled with Newbler 2.0.00.22. Illumina single-end data (taken from the in-progress paired-end run) were assembled using SOAPdenovo 1.06[1] (with *k*-mer of 51 and parameters "-d 1, -R". Assemblies were combined using AMOS minimus2 1.59 with parameters REFCOUNT=0, OVERLAP=50, MINID=94, MAXTRIM=10[2]. The resulting assembly consisted of 451 contigs greater than 200bp with an N50 of 53266bp. The largest contig was 204342bp.

Creation of a draft genome scaffold assembly using Illumina paired-end and mate-pair reads

A draft *de novo* assembly was produced using SOAPdenovo version 1.05. Contigs were first assembled using the 470bp paired-end library initially using a *k*-mer value of 45 for de Bruijn graph construction. These were subsequently scaffolded in a hierarchical fashion using 2kb followed by 6kb mate-pair libraries by way of the rank parameter in the SOAPdenovo configuration file. Other parameters supplied to SOAPdenovo included -F to attempt to fill

gaps in scaffolds. Where possible, in order to fill remaining scaffold gaps, local information available from the abundant mate-pair data was utilised by the GapCloser utility which was run over the assembly output with a *k*-mer size of 23. Both scaffolds and un-scaffolded contigs were used in further analysis, with the exception of contigs smaller than 200bp which were excluded.

De novo assembly produced 24 scaffolds plus 75 un-scaffolded contigs. The largest scaffold was 757969bp, the smallest was 552bp. Scaffold N50 was 403980bp. After gap filling the scaffolds contained 143 distinct stretches of gaps (represented as ambiguous 'N' bases) comprising 94491bp of sequence.

Insert sizes

The estimated insert size with standard deviations predicted by SOAPdenovo are demonstrated in Table S1.

Table S1. The estimated insert size determined by the *de novo* assembly process.

Library	Estimated insert size	Standard deviation
470bp	468	31
2kb	2548	246
6kb	6193	566

Determination of closest reference by average nucleotide identity (ANI)

Average nucleotide identity with all complete *E. coli* genomes available in GenBank was calculated using the ANIb algorithm which uses BLAST as the underlying alignment method[3-4]. Scrutiny of results (Table S2) revealed that *E. coli* 55989 showed the highest nucleotide identity with an ANI of 99.8% between the TY2482 draft chromosome and *E. coli*

55989. The ANIb algorithm shreds sequences into 1kb segments. BLAST alignments needed to be longer than 700bp and have >70% nucleotide identity to count towards ANIb calculation. ANIb parameters to BLAST were "-F F -e 0.001 -v 1 -b 1 -X 150 -q -1".

Table S2. Average nucleotide identities for TY2482 compared against all complete *E. coli* genomes

TY2482 vs	ANIb
Escherichia coli 55989	99.84
Escherichia coli IAI1	99.2
Escherichia coli W	99.14
Escherichia coli E24377A	99.09
Escherichia coli SE11	99.09
Escherichia coli O103:H2 str. 12009	98.95
Escherichia coli O26:H11 str. 11368	98.98
Escherichia coli O111:H- str. 11128	98.85
Escherichia coli HS	98.67
Escherichia coli ATCC 8739	98.55
Escherichia coli str. K-12 substr. W3110	98.54
Escherichia coli str. K-12 substr. MG1655	98.54
Escherichia coli DH1	98.54
Escherichia coli BL21-Gold(DE3)plyss AG	98.53
Escherichia coli BL21(DE3)	98.53
Escherichia coli BL21(DE3)	98.53
Escherichia coli B str. REL606	98.53
Escherichia coli BW2952	98.49
Escherichia coli str. K-12 substr. DH10B	98.5
Escherichia coli H10407	98.5
Escherichia coli ETEC H10407	98.5
Escherichia coli O55:H7 str. CB9615	97.92
Escherichia coli O157:H7 str. TW14359	97.86
Escherichia coli O157:H7 str. Sakai	97.87
Escherichia coli O157:H7 str. EC4115	97.86
Escherichia coli O157:H7 str. EDL933	97.82
Escherichia coli 042	97.45
Escherichia coli UMN026	97.39
Escherichia coli IAI39	97.3
Escherichia coli SMS-3-5	97.21
Escherichia coli SE15	97.06
Escherichia coli CFT073	97.02
Escherichia coli S88	96.97

Escherichia coli O83:H1 str. NRG 857C	96.99
Escherichia coli O127:H6 str. E2348/69	96.95
Escherichia coli UM146	96.94
Escherichia coli 536	96.95
Escherichia coli UTI89	96.93
Escherichia coli APEC O1	96.98
Escherichia coli ED1a	96.82

Annotation of putative regions of difference between TY2482 and 59989

The TY2482 scaffold assembly was aligned against *E. coli 55989* using progressiveMauve[5] (part of Mauve 2.3.1) using default settings. For ease of viewing, scaffolds were moved and where necessary reverse complemented to fit the order of the *E. coli 55989* chromosome using the Mauve contig mover, again run with default parameters. Unaligned regions of the TY2482 >= 5kb were examined as putative regions of difference. Gene prediction within these regions was performed using Glimmer 3.02[6] using the g3-iterated.sh workflow with default options. Genes with a raw score of >= 1.0 were extracted for further analysis. Due to Glimmer mis-predictions when run on the plasmid sequences, plasmids pESBL and pAA instead used Heuristic GeneMark.hmm[7] PROKARYOTIC (version 2.8a) for gene calling. This was run with default settings and model file "heuristic_no_rbs.mat" (http://opal.biology.gatech.edu/GeneMark/heuristic_hmm2.cgi). Putative protein products >= 50 aa in length were searched against the Genbank non-redundant protein database using PHMMER using HMMER (http://hmmer.janelia.org/). Genome visualisation plots were generated using CGview[8].

Manual inspection of scaffolds revealed each plasmid was contained within a single scaffold. Manual curation of pG2011 demonstrated an ~1.5kb plasmid with >99% nucleotide identity to *E. coli* strain H30 plasmid pO26-S1. This plasmid sequence was present as a 2-copy tandem repeat in the assembly, likely an artefact of the mate-pair assembly process (as insert

sizes longer than the plasmid were used) and has been manually edited to form a single copy. The location of the plasmids in the assembly are as follows: pESBL-TY2482 = scaffold19, pAA-TY2482 = scaffold16, pG2011 = scaffold21.

Accession numbers

The sequencing reads have been deposited into NCBI's Short Read Archive with accession numbers SRR227300, SRR227337, SRR227338, SRR227339, SRR227340, SRR231653, SRR231654 (Ion Torrent) and SRX079806 (Illumina mate-pair), SRX079805 (Illumina mate-pair) , SRX079804 (Illumina paired-end).

The scaffolded assembly and annotation has been deposited to Genbank, accession number AFVR00000000 (draft Illumina scaffold assembly), AFVS00000000 (Ion Torrent assembly) and AFOG01000000 (hybrid Ion Torrent and Illumina single-end assembly).

References

1. Li R, Zhu H, Ruan J, et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 2010;20:265-72.
2. Sommer DD, Delcher AL, Salzberg SL, Pop M. Minimus: a fast, lightweight genome assembler. BMC Bioinformatics. 2007;8:64.
3. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol 2007;57:81-91.
4. Richter M, Rossello-Mora R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A 2009;106:19126-31.
5. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 2010;5:e11147.

6. Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 2007;23:673-679
7. Besemer J, Borodovsky M. Heuristic approach to deriving models for gene finding. Nucleic Acids Res 1999;27:3911-20.
8. Stothard P, Wishart DS. Circular genome visualization and exploration using CGView. Bioinformatics 2005;21:537-9.

Table S3. Annotated genes on the RODs and plasmids of TY2482

ROD_ID	CDS_ID	Best hit (NR)	Curated annotation	Best hit (species)
I-ROD1	irod1_orf00001	conserved domain protein		Escherichia coli MS 84-1
I-ROD1	irod1_orf00002	hypothetical protein ERKG_00886		Escherichia coli H252
I-ROD1	irod1_orf00003	transposon Tn21 resolvase		Escherichia coli B7A
I-ROD1	irod1_orf00005	hypothetical protein ECoL_00180		Escherichia coli EC4100B
I-ROD1	irod1_orf00007	conserved domain protein		Escherichia coli MS 187-1
I-ROD1	irod1_orf00008	conserved domain protein		Escherichia coli MS 187-1
I-ROD1	irod1_orf00009	hypothetical protein HMPREF9550_01817		Escherichia coli MS 187-1
I-ROD1	irod1_orf00011	putative acyl-carrier-protein S-malonyltransferase		Escherichia coli B7A
I-ROD1	irod1_orf00012	hypothetical protein EcB7A_3346		Escherichia coli B7A
I-ROD1	irod1_orf00013	hypothetical protein HMPREF9550_01813		Escherichia coli MS 187-1
I-ROD1	irod1_orf00015	hypothetical protein ECoL_00172		Escherichia coli EC4100B
I-ROD1	irod1_orf00016	hypothetical protein HMPREF9542_01440		Escherichia coli MS 117-3
I-ROD1	irod1_orf00017	phage integrase		Escherichia coli H252
I-ROD2	irod2_orf00001	integrase		Escherichia coli O26:H11 str. 11368
I-ROD2	irod2_orf00002	hypothetical bacteriophage protein		Escherichia coli O157:H7 str. EC4501
I-ROD2	irod2_orf00004	conserved hypothetical protein		Escherichia coli O157:H7 str. EC4501
I-ROD2	irod2_orf00006	putative antirepressor		Escherichia coli O103:H2 str. 12009
I-ROD2	irod2_orf00007	hypothetical protein ECO26_1566		Escherichia coli O26:H11 str. 11368
I-ROD2	irod2_orf00008	hypothetical protein ECO26_1570		Escherichia coli O26:H11 str. 11368
I-ROD2	irod2_orf00009	gp43		Escherichia coli B171
I-ROD2	irod2_orf00010	conserved hypothetical protein		Escherichia coli O157:H7 str. EC508
I-ROD2	irod2_orf00011	conserved hypothetical protein		Escherichia coli O157:H7 str. EC508
I-ROD2	irod2_orf00012	hypothetical protein ECO103_2866		Escherichia coli O103:H2 str. 12009
I-ROD2	irod2_orf00013	hypothetical protein ECO103_2865		Escherichia coli O103:H2 str. 12009
I-ROD2	irod2_orf00014	putative exonuclease		Escherichia coli O103:H2 str. 12009
I-ROD2	irod2_orf00015	RecT protein		Escherichia coli O157:H7 str. EC508
I-ROD2	irod2_orf00016	conserved hypothetical protein		Escherichia coli O157:H7 str. EC4501

I-ROD2	irod2_orf00017	hypothetical protein ECH7EC4501_4934		Escherichia coli O157:H7 str. EC4501
I-ROD2	irod2_orf00019	conserved domain protein		Escherichia coli O157:H7 str. EC4501
I-ROD2	irod2_orf00020	conserved hypothetical protein		Escherichia coli O157:H7 str. EC4501
I-ROD2	irod2_orf00021	phage regulatory protein, Rha family		Escherichia coli O157:H7 str. EC4501
I-ROD2	irod2_orf00022	type II restriction enzyme BsuBI		Escherichia coli O157:H7 str. TW14588
I-ROD2	irod2_orf00023	modification methylase BsuBI		Escherichia coli O157:H7 str. TW14588
I-ROD2	irod2_orf00025	repressor protein CI		Escherichia coli O157:H7 str. EC4501
I-ROD2	irod2_orf00026	hypothetical protein SDY_1924		Shigella dysenteriae Sd197
I-ROD2	irod2_orf00027	helicase domain protein		Escherichia coli O157:H7 str. EC4501
I-ROD2	irod2_orf00028	hypothetical protein ECOK1180_4044		Escherichia coli 1180
I-ROD2	irod2_orf00029	hypothetical protein ECO103_2848		Escherichia coli O103:H2 str. 12009
I-ROD2	irod2_orf00030	protein ninG		Escherichia coli 1180
I-ROD2	irod2_orf00033	DNA modification methylase		Stx2-converting phage 86
I-ROD2	irod2_orf00034	Shiga toxin 2 subunit A		Enterobacteria phage 933W
I-ROD2	irod2_orf00036	Shiga toxin 2 subunit B		Enterobacteria phage 933W
I-ROD2	irod2_orf00037	hypothetical protein		Shigella phage 7888
I-ROD2	irod2_orf00038	hypothetical protein ECs2970		Escherichia coli O157:H7 str. Sakai
I-ROD2	irod2_orf00039	conserved domain protein		Escherichia coli O157:H7 str. EC4113
I-ROD2	irod2_orf00040	lysozyme		Escherichia coli O157:H7 str. EC4501
I-ROD2	irod2_orf00041	anti-repressor protein Ant		Enterobacteria phage VT2phi_272
I-ROD2	irod2_orf00044	endopeptidase (Protein gp15)		Escherichia coli S88
I-ROD2	irod2_orf00045	Rha protein		Escherichia coli O157:H7 str. TW14588
I-ROD2	irod2_orf00048	putative terminase small subunit		Stx2-converting phage 86
I-ROD2	irod2_orf00049	hypothetical protein ECOK1180_4067		Escherichia coli 1180
I-ROD2	irod2_orf00050	large subunit terminase		Escherichia coli O157:H7 str. EC4113
I-ROD2	irod2_orf00051	putative phage portal protein		Stx2-converting phage 86
I-ROD2	irod2_orf00052	hypothetical protein 933Wp53		Enterobacteria phage 933W
I-ROD2	irod2_orf00053	hypothetical protein 933Wp54		Enterobacteria phage 933W
I-ROD2	irod2_orf00054	hypothetical protein 933Wp55		Enterobacteria phage 933W
I-ROD2	irod2_orf00055	hypothetical protein 933Wp56		Enterobacteria phage 933W
I-ROD2	irod2_orf00056	hypothetical protein ECs1226		Escherichia coli O157:H7 str. Sakai
I-ROD2	irod2_orf00057	hypothetical protein ECO103_2826		Escherichia coli O103:H2 str. 12009

I-ROD2	irod2_orf00059	tail fiber protein		Escherichia coli O157:H7 str. EC4113
I-ROD2	irod2_orf00061	hypothetical protein Stx2-86_gp25		Stx2-converting phage 86
I-ROD2	irod2_orf00063	conserved hypothetical protein		Escherichia coli O157:H7 str. EC4196
I-ROD2	irod2_orf00065	outer membrane protein Lom precursor		Enterobacteria phage 933W
I-ROD2	irod2_orf00066	conserved hypothetical protein		Escherichia coli O157:H7 str. EC4501
I-ROD2	irod2_orf00067	hypothetical protein 933Wp68		Enterobacteria phage 933W
I-ROD2	irod2_orf00068	conserved hypothetical protein		Escherichia coli O157:H7 str. EC4501
I-ROD2	irod2_orf00069	hypothetical protein 933Wp70		Enterobacteria phage 933W
I-ROD2	irod2_orf00071	hypothetical protein 933Wp71		Enterobacteria phage 933W
I-ROD2	irod2_orf00072	hypothetical protein Stx2-86_gp35		Stx2-converting phage 86
I-ROD2	irod2_orf00073	hypothetical protein		Enterobacteria phage Min27
I-ROD3	irod3_orf00001	predicted integrase		Escherichia sp. 4_1_40B
I-ROD3	irod3_orf00002	unknown		Shigella flexneri 2a
I-ROD3	irod3_orf00003	prophage CP4-57 regulatory protein alpA		Escherichia coli 3431
I-ROD3	irod3_orf00004	unknown		Shigella flexneri 2a
I-ROD3	irod3_orf00005	type III restriction enzyme, res subunit		Pelobacter propionicus DSM 2379
I-ROD3	irod3_orf00006	hypothetical protein E4_08923		Escherichia sp. 4_1_40B
I-ROD3	irod3_orf00007	Transposase		Shigella dysenteriae CDC 74-1112
I-ROD3	irod3_orf00008	IS66 family element, orf2		Shigella boydii CDC 3083-94
I-ROD3	irod3_orf00011	hypothetical protein		Escherichia coli
I-ROD3	irod3_orf00012	hypothetical protein HMPREF9552_03072		Escherichia coli MS 198-1
I-ROD3	irod3_orf00013	hypothetical protein E4_08823		Escherichia sp. 4_1_40B
I-ROD3	irod3_orf00014	hypothetical protein Z1185		Escherichia coli O157:H7 EDL933
I-ROD3	irod3_orf00015	conserved hypothetical protein		Escherichia coli O157:H7 str. TW14588
I-ROD3	irod3_orf00016	hypothetical protein Z1188		Escherichia coli O157:H7 EDL933
I-ROD3	irod3_orf00017	putative glucosyltransferase		Escherichia coli O157:H7 EDL933
I-ROD3	irod3_orf00018	putative ferric enterochelin esterase McmK		Escherichia coli
I-ROD3	irod3_orf00019	MchS2 protein		Escherichia coli
I-ROD3	irod3_orf00020	hypothetical protein p1ECUMN_0112		Escherichia coli UMN026
I-ROD3	irod3_orf00022	MchC protein		Escherichia coli CFT073
I-ROD3	irod3_orf00023	microcin H47 secretion protein		Escherichia coli 042
I-ROD3	irod3_orf00024	MtfB		Escherichia coli

I-ROD3	irod3_orf00025	conserved hypothetical protein		Escherichia coli O157:H7 str. EC4196
I-ROD3	irod3_orf00026	hypothetical protein ECDG_03856		Escherichia coli B185
I-ROD3	irod3_orf00027	hypothetical protein ROD_49891		Citrobacter rodentium ICC168
I-ROD3	irod3_orf00028	hypothetical protein ROD_49911		Citrobacter rodentium ICC168
I-ROD3	irod3_orf00029	ImpA-related N- superfamily		Escherichia coli M605
I-ROD3	irod3_orf00030	hypothetical protein AHA_1063		Aeromonas hydrophila subsp. hydrophila ATCC 7966
I-ROD3	irod3_orf00031	immunoglobulin-binding regulator A		Escherichia coli M605
I-ROD3	irod3_orf00032	insertion element IS1 7 protein insA		Shigella dysenteriae 1617
I-ROD3	irod3_orf00034	hypothetical protein ECNA114_2538		Escherichia coli NA114
I-ROD3	irod3_orf00035	putative transposase		Shigella flexneri K-671
I-ROD3	irod3_orf00036	putative ATP synthase F0, A subunit		Escherichia coli MS 116-1
I-ROD3	irod3_orf00037	aspartate racemase		Shigella flexneri K-272
I-ROD3	irod3_orf00038	hypothetical protein HMPREF9541_00362		Escherichia coli MS 116-1
I-ROD3	irod3_orf00039	putative transcriptional regulator		Shigella flexneri 2a
I-ROD3	irod3_orf00042	conserved domain protein		Escherichia coli MS 116-1
I-ROD3	irod3_orf00043	predicted protein		Nematostella vectensis
I-ROD3	irod3_orf00044	protein kinase		Yersinia pseudotuberculosis IP 31758
I-ROD3	irod3_orf00045	hypothetical protein ESA_01782		Cronobacter sakazakii ATCC BAA-894
I-ROD3	irod3_orf00046	putative tellurium resistance protein	TerY3	Serratia marcescens
I-ROD3	irod3_orf00047	putative tellurium resistance protein	TerY2	Serratia marcescens
I-ROD3	irod3_orf00049	tellurium resistance protein	TerX	Serratia marcescens
I-ROD3	irod3_orf00050	putative tellurium resistance protein TerY	TerY1	Enterobacter cloacae subsp. cloacae ATCC 13047
I-ROD3	irod3_orf00051	terW	TerW	Citrobacter sp. 30_2
I-ROD3	irod3_orf00052	hypothetical protein SMR0069		Serratia marcescens
I-ROD3	irod3_orf00053	hypothetical protein Z1166		Escherichia coli O157:H7 EDL933
I-ROD3	irod3_orf00054	putative ATP-binding protein		Escherichia coli APEC O1
I-ROD3	irod3_orf00055	hypothetical protein APECO1_O1R65		Escherichia coli APEC O1
I-ROD3	irod3_orf00056	hypothetical protein APECO1_O1R66		Escherichia coli APEC O1
I-ROD3	irod3_orf00057	hypothetical protein APECO1_O1R67		Escherichia coli APEC O1
I-ROD3	irod3_orf00058	putative phage inhibition, colicin resistance and tellurite resistance protein	TerZ	Escherichia coli O157:H7 EDL933

I-ROD3	irod3_orf00059	putative phage inhibition, colicin resistance and tellurite resistance protein	TerA	Escherichia coli O157:H7 EDL933
I-ROD3	irod3_orf00060	putative phage inhibition, colicin resistance and tellurite resistance protein	TerB	Escherichia coli O157:H7 EDL933
I-ROD3	irod3_orf00061	putative phage inhibition, colicin resistance and tellurite resistance protein	TerC	Escherichia coli O157:H7 EDL933
I-ROD3	irod3_orf00063	putative phage inhibition, colicin resistance and tellurite resistance protein	TerD	Escherichia coli O157:H7 EDL933
I-ROD3	irod3_orf00064	putative phage inhibition, colicin resistance and tellurite resistance protein	TerE	Escherichia coli O157:H7 EDL933
I-ROD3	irod3_orf00065	putative tellurium resistance protein TerF	TerF	Escherichia coli O103:H2 str. 12009
I-ROD3	irod3_orf00067	putative GTP-binding protein		Escherichia coli 042
I-ROD3	irod3_orf00068	antigen 43 precursor		Escherichia coli
I-ROD3	irod3_orf00071	putative autotransporter		Shigella sp. D9
I-ROD3	irod3_orf00072	hypothetical protein EscherichiacoliO157_22726		Escherichia coli O157:H7 str. FRIK2000
I-ROD3	irod3_orf00073	hypothetical protein ECS88_2092		Escherichia coli S88
I-ROD3	irod3_orf00074	hypothetical protein ECS88_2092		Escherichia coli S88
I-ROD3	irod3_orf00075	conserved hypothetical protein		Escherichia coli H591
I-ROD3	irod3_orf00077	hypothetical protein SD1617_3951		Shigella dysenteriae 1617
I-ROD3	irod3_orf00078	hypothetical protein ECS88_2094		Escherichia coli S88
I-ROD3	irod3_orf00079	hypothetical protein APECO1_1098		Escherichia coli APEC O1
I-ROD3	irod3_orf00080	hypothetical protein ECO103_3758		Escherichia coli O103:H2 str. 12009
I-ROD3	irod3_orf00081	hypothetical protein ECNA114_2131		Escherichia coli NA114
I-ROD3	irod3_orf00083	toxin of the YeeV-YeeU toxin-antitoxin system		Escherichia sp. 4_1_40B
I-ROD3	irod3_orf00084	conserved hypothetical protein		Escherichia coli ETEC H10407
I-ROD3	irod3_orf00086	hypothetical protein UTI89_C4999		Escherichia coli UTI89
I-ROD3	irod3_orf00087	hypothetical protein Z1226		Escherichia coli O157:H7 EDL933
I-ROD4	irod4_orf00001	AntB		Escherichia coli
I-ROD4	irod4_orf00003	conserved hypothetical protein		Escherichia coli FVEC1302
I-ROD4	irod4_orf00004	valyl-tRNA synthetase		Escherichia coli E110019
I-ROD4	irod4_orf00006	hypothetical protein Stx2-86_gp35		Stx2-converting phage 86
I-ROD4	irod4_orf00008	hypothetical protein SDY_1670		Shigella dysenteriae Sd197
I-ROD4	irod4_orf00010	hypothetical protein ECED1_1152		Escherichia coli ED1a
I-ROD4	irod4_orf00011	hypothetical protein ECED1_1151		Escherichia coli ED1a
I-ROD4	irod4_orf00012	hypothetical protein 933Wp68		Enterobacteria phage 933W
I-ROD4	irod4_orf00013	hypothetical protein Stx2-86_gp30		Stx2-converting phage 86
I-ROD4	irod4_orf00014	putative outer membrane precursor Lom		Escherichia coli O103:H2 str. 12009

I-ROD4	irod4_orf00016	hypothetical protein Stx2Ip034		Stx2 converting phage I
I-ROD4	irod4_orf00018	hypothetical protein Stx2-86_gp25		Stx2-converting phage 86
I-ROD4	irod4_orf00020	putative long tail fiber protein		Stx2-converting phage 86
I-ROD4	irod4_orf00021	hypothetical protein ECED1_1137		Escherichia coli ED1a
I-ROD4	irod4_orf00022	hypothetical protein ECED1_1136		Escherichia coli ED1a
I-ROD4	irod4_orf00023	hypothetical protein ECED1_1135		Escherichia coli ED1a
I-ROD4	irod4_orf00024	hypothetical protein Stx2-86_gp17		Stx2-converting phage 86
I-ROD4	irod4_orf00025	hypothetical protein ECED1_1133		Escherichia coli ED1a
I-ROD4	irod4_orf00026	hypothetical protein ECED1_1132		Escherichia coli ED1a
I-ROD4	irod4_orf00027	putative phage portal protein		Stx2-converting phage 86
I-ROD4	irod4_orf00028	hypothetical protein ECOK1180_4067		Escherichia coli 1180
I-ROD4	irod4_orf00029	large subunit terminase		Escherichia coli O157:H7 str. EC4113
I-ROD4	irod4_orf00030	putative terminase small subunit		Stx2-converting phage 86
I-ROD4	irod4_orf00033	bacteriophage lysis protein		Shigella dysenteriae 1012
I-ROD4	irod4_orf00036	putative endolysin		Shigella dysenteriae Sd197
I-ROD4	irod4_orf00037	protein S		Enterobacteria phage 933W
I-ROD4	irod4_orf00038	conserved hypothetical protein		Shigella dysenteriae 1617
I-ROD4	irod4_orf00039	hypothetical protein SGF_04061		Shigella flexneri CDC 796-83
I-ROD4	irod4_orf00040	YjhS		Shigella boydii CDC 3083-94
I-ROD4	irod4_orf00041	putative NinH protein		Phage BP-4795
I-ROD4	irod4_orf00042	crossover junction endodeoxyribonuclease		Escherichia coli ED1a
I-ROD4	irod4_orf00044	hypothetical protein E2348C_2522		Escherichia coli O127:H6 str. E2348/69
I-ROD4	irod4_orf00045	putative ninB protein		Escherichia coli ED1a
I-ROD4	irod4_orf00046	putative antirepressor protein Ant from prophage		Escherichia coli ED1a
I-ROD4	irod4_orf00047	hypothetical protein ECO26_2262		Escherichia coli O26:H11 str. 11368
I-ROD4	irod4_orf00048	death-on-curing family protein		Escherichia coli STEC_7v
I-ROD4	irod4_orf00049	hypothetical protein ECSTEC7V_1837		Escherichia coli STEC_7v
I-ROD4	irod4_orf00050	hypothetical protein ECO111_1061		Escherichia coli O111:H- str. 11128
I-ROD4	irod4_orf00051	hypothetical protein G2583_1712		Escherichia coli O55:H7 str. CB9615
I-ROD4	irod4_orf00052	hypothetical protein EcE24377A_1426		Escherichia coli E24377A
I-ROD4	irod4_orf00053	hypothetical protein ECO103_1369		Escherichia coli O103:H2 str. 12009
I-ROD4	irod4_orf00055	putative replication protein		Escherichia coli ED1a

I-ROD4	irod4_orf00056	hypothetical protein ECED1_1103		Escherichia coli ED1a
I-ROD4	irod4_orf00057	hypothetical protein ECED1_1102		Escherichia coli ED1a
I-ROD4	irod4_orf00058	regulatory protein CII from prophage		Escherichia coli ED1a
I-ROD4	irod4_orf00059	prophage repressor CI		Enterobacteria phage HK97
I-ROD4	irod4_orf00060	hypothetical protein ECED1_1098		Escherichia coli ED1a
I-ROD4	irod4_orf00061	hypothetical protein ECED1_1097		Escherichia coli ED1a
I-ROD4	irod4_orf00063	monocarboxylate transporter		Culex quinquefasciatus
I-ROD4	irod4_orf00064	hypothetical protein ECED1_1095		Escherichia coli ED1a
I-ROD4	irod4_orf00065	hypothetical protein ECED1_1094		Escherichia coli ED1a
I-ROD4	irod4_orf00067	FtsZ inhibitor protein		Escherichia coli ED1a
I-ROD4	irod4_orf00068	hypothetical protein ECED1_1091		Escherichia coli ED1a
I-ROD4	irod4_orf00069	Exodeoxyribonuclease VIII (putative partial) from phage origin		Escherichia coli ED1a
I-ROD4	irod4_orf00070	putative host-nuclease inhibitor protein Gam		Shigella dysenteriae Sd197
I-ROD4	irod4_orf00071	Recombination protein bet from phage origin		Escherichia coli ED1a
I-ROD4	irod4_orf00072	putative exonuclease encoded by prophage CP-933K		Escherichia coli O157:H7 EDL933
I-ROD4	irod4_orf00074	prophage DLP12 integrase		Escherichia coli 101-1
I-ROD5	irod5_orf00001	hypothetical protein SSON_1273		Shigella sonnei Ss046
I-ROD5	irod5_orf00002	hypothetical protein EC55989_1079		Escherichia coli 55989
I-ROD5	irod5_orf00005	triple helix repeat-containing collagen		Clostridium beijerinckii NCIMB 8052
I-ROD5	irod5_orf00006	hypothetical protein SD15574_2985		Shigella dysenteriae 155-74
I-ROD5	irod5_orf00008	hypothetical protein ECE128010_5420		Escherichia coli E128010
I-ROD5	irod5_orf00011	Putative tail component of prophage		Escherichia coli NA114
I-ROD5	irod5_orf00012	hypothetical protein ECLG_05105		Escherichia coli TA271
I-ROD5	irod5_orf00015	Superoxide dismutase (Cu-Zn)		Escherichia coli O55:H7 str. CB9615
I-ROD5	irod5_orf00019	minor tail protein		Escherichia coli UTI89
I-ROD5	irod5_orf00021	minor tail protein		Escherichia coli UTI89
I-ROD5	irod5_orf00022	putative tail fiber component H of prophage CP-933U		Escherichia coli O157:H7 EDL933
I-ROD5	irod5_orf00024	Phage minor tail protein		Escherichia coli EC4100B
I-ROD5	irod5_orf00025	Phage minor tail protein		Escherichia coli EC4100B
I-ROD5	irod5_orf00026	phage major tail protein		Escherichia coli 042
I-ROD5	irod5_orf00027	hypothetical protein DAPPUDRAFT_279812		Daphnia pulex
I-ROD5	irod5_orf00029	polysaccharide Transporter, PST family		Enterococcus faecium E1679

I-ROD5	irod5_orf00032	hypothetical protein		Arthrospira platensis NIES-39
I-ROD5	irod5_orf00034	terminase large subunit domain protein		Escherichia coli RN587/1
I-ROD5	irod5_orf00035	conserved hypothetical protein		Escherichia albertii TW07627
I-ROD5	irod5_orf00036	phage major capsid protein E		Escherichia coli H489
I-ROD5	irod5_orf00039	Hypothetical protein CBG02325		Caenorhabditis briggsae
I-ROD5	irod5_orf00041	conserved domain protein		Escherichia coli MS 153-1
I-ROD5	irod5_orf00042	hypothetical protein c1457		Escherichia coli CFT073
I-ROD5	irod5_orf00043	Phage minor tail protein		Escherichia coli EC4100B
I-ROD5	irod5_orf00045	hypothetical protein MK0973		Methanopyrus kandleri AV19
I-ROD5	irod5_orf00047	hypothetical protein SCA50_1305		Salmonella enterica subsp. enterica serovar Choleraesuis str. SCSA50
I-ROD5	irod5_orf00049	hypothetical protein ECOK1_1278		Escherichia coli IHE3034
I-ROD5	irod5_orf00052	phage DNA packaging protein Nu1		Escherichia coli MS 21-1
I-ROD5	irod5_orf00053	putative phage protein		Escherichia coli 042
I-ROD5	irod5_orf00055	hypothetical protein ECS88_0566		Escherichia coli S88
I-ROD5	irod5_orf00056	endopeptidase		Escherichia coli 2362-75
I-ROD5	irod5_orf00058	hypothetical protein SBO_1923		Shigella boydii Sb227
I-ROD5	irod5_orf00059	putative membrane-associated lysozyme; Qin prophage		Escherichia coli 55989
I-ROD5	irod5_orf00061	hypothetical protein Stx2-86_gp06		Stx2-converting phage 86
I-ROD5	irod5_orf00062	hypothetical protein Stx2-86_gp05		Stx2-converting phage 86
I-ROD5	irod5_orf00063	lysis protein S		Stx2-converting phage 86
I-ROD5	irod5_orf00066	hypothetical protein DAPPUDRAFT_52038		Daphnia pulex
I-ROD5	irod5_orf00068	heterokaryon incompatibility protein		Glomerella graminicola M1.001
I-ROD5	irod5_orf00070	DNA methylase family protein		Shigella flexneri J1713
I-ROD5	irod5_orf00071	hypothetical protein HMPREF9542_00842		Escherichia coli MS 117-3
I-ROD5	irod5_orf00072	hypothetical protein EcF11_4284		Escherichia coli F11
I-ROD5	irod5_orf00075	hypothetical protein ECRN5871_4170		Escherichia coli RN587/1
I-ROD5	irod5_orf00076	hypothetical protein E4_10746		Escherichia sp. 4_1_40B
I-ROD5	irod5_orf00077	endodeoxyribonuclease RusA family protein		Escherichia coli STEC_7v
I-ROD5	irod5_orf00078	LexA repressor		Escherichia coli S88
I-ROD5	irod5_orf00079	DNA adenine methylase		Escherichia coli UTI89
I-ROD5	irod5_orf00080	hypothetical protein ECS88_0547		Escherichia coli S88

I-ROD5	irod5_orf00081	hypothetical protein PcdtI_gp46		Phage cdtI
I-ROD5	irod5_orf00082	putative antirepressor		Escherichia coli EC4100B
I-ROD5	irod5_orf00083	nucleic acid-binding protein; e14 prophage		Escherichia coli S88
I-ROD5	irod5_orf00084	hypothetical protein ECD227_2469		Escherichia fergusonii ECD227
I-ROD5	irod5_orf00085	regulatory protein cI		Escherichia coli EC4100B
I-ROD5	irod5_orf00086	hypothetical protein ECoL_03975		Escherichia coli EC4100B
I-ROD5	irod5_orf00087	hypothetical protein ECoL_03976		Escherichia coli EC4100B
I-ROD5	irod5_orf00089	Hypothetical protein yfdR		Escherichia coli EC4100B
I-ROD5	irod5_orf00090	hypothetical protein ShiD9_12075		Shigella sp. D9
I-ROD5	irod5_orf00091	conserved hypothetical protein		Escherichia coli E22
I-ROD5	irod5_orf00093	conserved hypothetical protein		Escherichia coli E22
I-ROD5	irod5_orf00095	Phage EaA protein		Escherichia coli EC4100B
I-ROD5	irod5_orf00096	Phage EaA protein		Escherichia coli EC4100B
I-ROD5	irod5_orf00097	Integrase		Escherichia coli EC4100B
I-ROD6	irod6_orf00001	molybdate metabolism regulator		Escherichia coli 536
I-ROD6	irod6_orf00003	hypothetical protein ECP_2154		Escherichia coli 536
I-ROD6	irod6_orf00005	yehL protein		Escherichia coli B088
I-ROD6	irod6_orf00006	hypothetical protein ECP_2157		Escherichia coli 536
I-ROD6	irod6_orf00007	hypothetical protein ECIAI1_2197		Escherichia coli IAI1
I-ROD6	irod6_orf00008	hypothetical protein ECP_2159		Escherichia coli 536
I-ROD7	irod7_orf00001	integrase		Escherichia coli
I-ROD7	irod7_orf00002	Evolved beta-D-galactosidase, beta subunit		Shigella dysenteriae CDC 74-1112
I-ROD7	irod7_orf00003	transposase TnpA		Corynebacterium glutamicum
I-ROD7	irod7_orf00004	resolvase for Tn21		Plasmid R100
I-ROD7	irod7_orf00006	Urf2 protein		Escherichia fergusonii ECD227
I-ROD7	irod7_orf00007	integrase		Plasmid R100
I-ROD7	irod7_orf00008	dihydrofolate reductase type A7		Salmonella enterica subsp. enterica serovar Weltevreden
I-ROD7	irod7_orf00010	putative transposase		Klebsiella pneumoniae subsp. pneumoniae MGH 78578
I-ROD7	irod7_orf00014	3-hydroxyisobutyrate dehydrogenase		Mycobacterium tuberculosis 210
I-ROD7	irod7_orf00016	protein RepC		Salmonella enterica subsp. enterica serovar Enteritidis

I-ROD7	irod7_orf00017	dihydropteroate synthase		Salmonella enterica subsp. enterica serovar Typhi str. CT18
I-ROD7	irod7_orf00018	aminoglycoside/hydroxyurea antibiotic resistance kinase		Escherichia coli MS 200-1
I-ROD7	irod7_orf00019	beta-lactamase		Escherichia coli 3431
I-ROD7	irod7_orf00021	hypothetical protein R100p008		Plasmid R100
I-ROD7	irod7_orf00022	putative mercury resistance protein		Plasmid R100
I-ROD7	irod7_orf00023	transcriptional regulator MerD		Plasmid R100
I-ROD7	irod7_orf00026	RecName: Full=Mercuric reductase; AltName: Full=Hg(II) reductase		
I-ROD7	irod7_orf00027	putative mercury transport protein MerC		Aeromonas salmonicida subsp. salmonicida A449
I-ROD7	irod7_orf00029	Tn501 orf, hypotheical		Shigella flexneri 5a
I-ROD7	irod7_orf00033	InsL		Escherichia coli 53638
I-ROD7	irod7_orf00034	hypothetical protein pFL129_4		Escherichia coli
I-ROD7	irod7_orf00036	TetA		Salmonella enterica subsp. enterica serovar Choleraesuis
I-ROD7	irod7_orf00037	integral membrane protein DUF6		Escherichia coli MS 78-1
I-ROD7	irod7_orf00038	hypothetical protein HMPREF9544_05491		Escherichia coli MS 153-1
I-ROD7	irod7_orf00039	conserved hypothetical protein		Escherichia coli ETEC H10407
I-ROD7	irod7_orf00040	conserved hypothetical protein		Escherichia coli SE15
I-ROD7	irod7_orf00042	hypothetical protein HMPREF9553_03865		Escherichia coli MS 200-1
I-ROD7	irod7_orf00044	putative regulatory protein		Escherichia coli 536
I-ROD7	irod7_orf00045	conserved hypothetical protein		Escherichia coli SE15
I-ROD7	irod7_orf00046	transposase		Escherichia coli SE15
I-ROD7	irod7_orf00048	hypothetical protein		Escherichia coli SE15
I-ROD7	irod7_orf00049	hypothetical protein ECUMN_4880		Escherichia coli UMN026
I-ROD7	irod7_orf00051	putative autotransporter		Escherichia coli 536
I-ROD7	irod7_orf00052	antigen 43 domain protein		Escherichia coli LT-68
I-ROD7	irod7_orf00053	hypothetical protein EcE24377A_4893		Escherichia coli E24377A
I-ROD7	irod7_orf00054	hypothetical protein ECNA114_2131		Escherichia coli NA114
I-ROD7	irod7_orf00056	conserved domain protein		Escherichia coli MS 187-1
I-ROD7	irod7_orf00058	conserved hypothetical protein		Escherichia coli SE15
I-ROD7	irod7_orf00059	putative radC-like protein yeeS		Escherichia coli CFT073
I-ROD7	irod7_orf00060	hypothetical protein c0272		Escherichia coli CFT073

I-ROD7	irod7_orf00061	unknown		Escherichia coli
I-ROD7	irod7_orf00063	DNA repair protein		Escherichia coli MS 78-1
I-ROD7	irod7_orf00064	hypothetical protein c4574		Escherichia coli CFT073
I-ROD7	irod7_orf00065	conserved hypothetical protein		Shigella dysenteriae 1617
I-ROD7	irod7_orf00067	hypothetical protein APECO1_3486		Escherichia coli APEC O1
I-ROD7	irod7_orf00068	hypothetical protein SF3000		Shigella flexneri 2a str. 301
I-ROD7	irod7_orf00069	hypothetical protein ECO103_3594		Escherichia coli O103:H2 str. 12009
I-ROD7	irod7_orf00070	hypothetical protein ECED1_4984		Escherichia coli ED1a
pESBL	scaffold19_orf0002	YciB		Escherichia coli
pESBL	scaffold19_orf0003	hypothetical protein pECBactecp21		Escherichia coli
pESBL	scaffold19_orf0004	hypothetical protein SC121		Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67
pESBL	scaffold19_orf0005	single-stranded DNA-binding protein		Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67
pESBL	scaffold19_orf0006	hypothetical protein pO157p50		Escherichia coli O157:H7 str. Sakai
pESBL	scaffold19_orf0007	plasmid SOS inhibition protein B		Escherichia coli
pESBL	scaffold19_orf0008	plasmid SOS inhibition protein A		Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67
pESBL	scaffold19_orf0009	hypothetical protein SC115		Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67
pESBL	scaffold19_orf0010	antirestriction protein		Escherichia coli MS 107-1
pESBL	scaffold19_orf0011	hypothetical protein ECSE_P1-0063		Escherichia coli SE11
pESBL	scaffold19_orf0012	hypothetical protein HMPREF9542_03988		Escherichia coli MS 117-3
pESBL	scaffold19_orf0013	hypothetical protein SeHA_A0062		Salmonella enterica subsp. enterica serovar Heidelberg str. SL476
pESBL	scaffold19_orf0014	hypothetical protein EcE22_3665		Escherichia coli E22
pESBL	scaffold19_orf0015	CcgAII protein		Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67
pESBL	scaffold19_orf0016	putative transposase		Escherichia coli E22
pESBL	scaffold19_orf0018	hypothetical protein SC107		Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67
pESBL	scaffold19_orf0019	hypothetical protein R64_p076		Salmonella enterica subsp. enterica serovar Typhimurium
pESBL	scaffold19_orf0020	hypothetical protein SC105		Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67

pESBL	scaffold19_orf0021	hypothetical protein pECBactecp14		Escherichia coli
pESBL	scaffold19_orf0022	hypothetical protein LH0067		Escherichia coli
pESBL	scaffold19_orf0023	relaxosome component		Plasmid ColIb-P9
pESBL	scaffold19_orf0024	NikB		Escherichia coli O157:H7 str. Sakai
pESBL	scaffold19_orf0025	hypothetical protein EcE24377A_D0057		Escherichia coli E24377A
pESBL	scaffold19_orf0026	hypothetical protein pECBactecp09		Escherichia coli
pESBL	scaffold19_orf0027	hypothetical protein pECBactecp08		Escherichia coli
pESBL	scaffold19_orf0028	putative protein FinQ		Escherichia coli MS 84-1
pESBL	scaffold19_orf0029	counter protein for PndA		Escherichia coli
pESBL	scaffold19_orf0030	hypothetical protein SC084		Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67
pESBL	scaffold19_orf0031	conserved hypothetical protein		Escherichia coli MS 107-1
pESBL	scaffold19_orf0032	hypothetical protein ECSE_P1-0081		Escherichia coli SE11
pESBL	scaffold19_orf0033	putative regulator protein		Escherichia coli SE11
pESBL	scaffold19_orf0034	exclusion-determining family protein		Escherichia coli MS 84-1
pESBL	scaffold19_orf0035	TraY		Escherichia coli O157:H7 str. EC4486
pESBL	scaffold19_orf0036	F pilin acetylation protein		Escherichia coli
pESBL	scaffold19_orf0037	F pilus assembly		Escherichia coli
pESBL	scaffold19_orf0038	F pilus assembly		Escherichia coli
pESBL	scaffold19_orf0039	TraU		Escherichia coli O157:H7 str. EC4401
pESBL	scaffold19_orf0040	hypothetical protein HMPREF9542_01329		Escherichia coli MS 117-3
pESBL	scaffold19_orf0041	TraR protein		Escherichia coli
pESBL	scaffold19_orf0042	hypothetical protein ColIb-P9_p070		Plasmid ColIb-P9
pESBL	scaffold19_orf0043	hypothetical protein ColIb-P9_p071		Plasmid ColIb-P9
pESBL	scaffold19_orf0044	hypothetical protein ColIb-P9_p072		Plasmid ColIb-P9
pESBL	scaffold19_orf0045	hypothetical protein ColIb-P9_p073		Plasmid ColIb-P9
pESBL	scaffold19_orf0046	hypothetical protein ColIb-P9_p074		Plasmid ColIb-P9
pESBL	scaffold19_orf0047	thick pilus signal peptide		Escherichia coli W
pESBL	scaffold19_orf0048	DNA primase		Escherichia coli O157:H7 str. EC4401
pESBL	scaffold19_orf0049	EDTA-resistant nuclease		Escherichia coli
pESBL	scaffold19_orf0051	ATP-binding protein		Plasmid ColIb-P9
pESBL	scaffold19_orf0052	lipoprotein		Salmonella enterica subsp. enterica serovar

				Typhimurium
pESBL	scaffold19_orf0053	hypothetical protein ColIb-P9_p082		Plasmid ColIb-P9
pESBL	scaffold19_orf0054	hypothetical protein ColIb-P9_p083		Plasmid ColIb-P9
pESBL	scaffold19_orf0055	F pilus assembly		Escherichia coli
pESBL	scaffold19_orf0056	TraE protein		Escherichia coli
pESBL	scaffold19_orf0057	shufflon-specific DNA recombinase		Escherichia coli AA86
pESBL	scaffold19_orf0058	hypothetical protein HMPREF9536_01879		Escherichia coli MS 84-1
pESBL	scaffold19_orf0059	conserved hypothetical protein		Escherichia coli MS 107-1
pESBL	scaffold19_orf0060	hypothetical protein R64_p118		Salmonella enterica subsp. enterica serovar Typhimurium
pESBL	scaffold19_orf0061	shufflon protein C'		Escherichia coli O157:H7 str. EC4486
pESBL	scaffold19_orf0062	conserved hypothetical protein		Escherichia coli MS 107-1
pESBL	scaffold19_orf0063	shufflon protein A		Salmonella enterica subsp. enterica serovar Kentucky str. CVM29188
pESBL	scaffold19_orf0064	peptidase A24A prepilin type IV		Escherichia coli W
pESBL	scaffold19_orf0065	type IV prepilin cluster		Escherichia coli
pESBL	scaffold19_orf0066	type IV prepilin cluster; prepilin		Escherichia coli
pESBL	scaffold19_orf0067	integral membrane protein		Escherichia coli E22
pESBL	scaffold19_orf0068	ATP-binding protein PilQ		Escherichia coli SE11
pESBL	scaffold19_orf0069	IncI1 conjugal transfer protein PilP		Escherichia coli
pESBL	scaffold19_orf0070	IncI1 conjugal transfer protein PilO		Escherichia coli
pESBL	scaffold19_orf0071	lipoprotein PilN		Escherichia coli SE11
pESBL	scaffold19_orf0072	hypothetical protein ColIb-P9_p101		Plasmid ColIb-P9
pESBL	scaffold19_orf0073	IncI1 conjugal transfer protein PilL		Escherichia coli
pESBL	scaffold19_orf0074	predicted protein		Nematostella vectensis
pESBL	scaffold19_orf0075	IncI1 conjugal transfer protein TraC		Escherichia coli
pESBL	scaffold19_orf0076	transcription termination factor NusG		Escherichia coli MS 84-1
pESBL	scaffold19_orf0077	TraA protein		Escherichia coli SE11
pESBL	scaffold19_orf0078	replication initiation protein		Salmonella enterica subsp. enterica serovar Kentucky str. CVM29188
pESBL	scaffold19_orf0079	hypothetical protein ND12IncI1_3		Escherichia coli
pESBL	scaffold19_orf0080	hypothetical protein pECBactecp34		Escherichia coli
pESBL	scaffold19_orf0081	YagA		Escherichia coli O157:H7 str. EC4486

pESBL	scaffold19_orf0082	transposase		Salmonella enterica subsp. enterica serovar Infantis
pESBL	scaffold19_orf0083	conserved hypothetical protein		Escherichia coli MS 21-1
pESBL	scaffold19_orf0084	hypothetical protein		Escherichia coli
pESBL	scaffold19_orf0085	hypothetical protein pC15-1a_016	blaCTX-M-15	Escherichia coli
pESBL	scaffold19_orf0086	ISEcp1 transposase		Escherichia coli
pESBL	scaffold19_orf0087	transposase for transposon Tn3		Escherichia coli
pESBL	scaffold19_orf0088	hypothetical protein pC15-1a_019		Escherichia coli
pESBL	scaffold19_orf0089	TEM-1 beta-lactamase	blaTEM-1	Salmonella enterica subsp. enterica serovar Montevideo
pESBL	scaffold19_orf0090	conserved domain protein		Escherichia coli MS 21-1
pESBL	scaffold19_orf0091	cobyrinic acid a,c-diamide synthase		Escherichia coli
pESBL	scaffold19_orf0093	protein impB domain protein		Escherichia coli 1357
pESBL	scaffold19_orf0094	hypothetical protein ColIb-P9_p029		Plasmid ColIb-P9
pESBL	scaffold19_orf0095	DinI-like family protein		Escherichia coli MS 21-1
pESBL	scaffold19_orf0096	hypothetical protein p026VIR_p092		Escherichia coli
pESBL	scaffold19_orf0097	hypothetical protein ECO103_p71		Escherichia coli O103:H2 str. 12009
pESBL	scaffold19_orf0098	conserved hypothetical protein		Escherichia coli H299
pESBL	scaffold19_orf0099	hypothetical protein ND12IncI1_24		Escherichia coli
pESBL	scaffold19_orf0100	conserved hypothetical protein		Escherichia coli W
pAA	scaffold16_orf0001	putative secreted protein		Streptomyces hygroscopicus ATCC 53653
pAA	scaffold16_orf0002	hypothetical protein c3579		Escherichia coli CFT073
pAA	scaffold16_orf0003	unknown protein encoded in ISEc8		Escherichia coli O157:H7 EDL933
pAA	scaffold16_orf0004	hypothetical protein SbBS512_A0019		Shigella boydii CDC 3083-94
pAA	scaffold16_orf0005	AggA457 protein	AggA	Escherichia coli
pAA	scaffold16_orf0006	RecName: Full=Protein AggB; Flags: Precursor	AggB	
pAA	scaffold16_orf0007	HdaC, HUS-associated diffuse adherence	AggC	Escherichia coli
pAA	scaffold16_orf0008	RecName: Full=Chaperone protein AggD; Flags: Precursor	AggD	
pAA	scaffold16_orf0010	putative resolvase		Escherichia coli
pAA	scaffold16_orf0011	3-hydroxyisobutyrate dehydrogenase		Mycobacterium tuberculosis 210
pAA	scaffold16_orf0012	hypothetical protein ColIb-P9_p027		Plasmid ColIb-P9
pAA	scaffold16_orf0013	StbA protein		Escherichia coli MS 84-1

pAA	scaffold16_orf0015	putative 60 kDa chaperonin		Escherichia coli E24377A
pAA	scaffold16_orf0016	hypothetical protein ColIb-P9_p024		Plasmid ColIb-P9
pAA	scaffold16_orf0017	resolvase		Salmonella enterica subsp. enterica serovar Kentucky str. CVM29188
pAA	scaffold16_orf0018	plasmid maintenance protein CcdB		Escherichia coli
pAA	scaffold16_orf0019	plasmid maintenance protein CcdA		Escherichia coli
pAA	scaffold16_orf0021	hypothetical protein E4_23171		Escherichia sp. 4_1_40B
pAA	scaffold16_orf0022	hypothetical protein p1ECUMN_0160		Escherichia coli UMN026
pAA	scaffold16_orf0024	orf906		Escherichia coli
pAA	scaffold16_orf0026	phage integrase		Escherichia coli M863
pAA	scaffold16_orf0027	COG1506: Dipeptidyl aminopeptidases/acylaminoacyl-peptidases		Magnetospirillum magnetotacticum MS-1
pAA	scaffold16_orf0028	hypothetical protein pECL46p020		Escherichia coli
pAA	scaffold16_orf0029	hypothetical protein pEC55989_0007		Escherichia coli 55989
pAA	scaffold16_orf0030	hypothetical protein IPF_103		Escherichia coli 1520
pAA	scaffold16_orf0031	incFII family plasmid replication initiator RepA		Escherichia coli MS 78-1
pAA	scaffold16_orf0032	replication initiation protein		Escherichia coli E128010
pAA	scaffold16_orf0033	replication protein		Escherichia sp. 4_1_40B
pAA	scaffold16_orf0034	conjugal transfer pilus acetylation protein TraX		Shigella flexneri 2a str. 301
pAA	scaffold16_orf0035	hypothetical protein pYT1_p113		Salmonella enterica subsp. enterica serovar Typhimurium
pAA	scaffold16_orf0036	DNA helicase TraI		Escherichia coli MS 57-2
pAA	scaffold16_orf0037	conserved hypothetical protein		Salmonella enterica subsp. enterica serovar Kentucky
pAA	scaffold16_orf0038	hypothetical protein c3659		Escherichia coli CFT073
pAA	scaffold16_orf0039	hypothetical protein c3661		Escherichia coli CFT073
pAA	scaffold16_orf0040	hypothetical protein pB171_031		Escherichia coli
pAA	scaffold16_orf0041	conserved hypothetical protein		Escherichia coli H299
pAA	scaffold16_orf0042	conjugal transfer fertility inhibition protein FinO		Escherichia coli
pAA	scaffold16_orf0043	conjugal transfer pilus acetylation protein TraX		Salmonella enterica subsp. enterica serovar Kentucky str. CVM29188
pAA	scaffold16_orf0044	hypothetical protein pYT1_p113		Salmonella enterica subsp. enterica serovar Typhimurium
pAA	scaffold16_orf0045	conjugal transfer nickase/helicase TraI		Escherichia coli

pAA	scaffold16_orf0046	conjugal transfer nickase/helicase TraI		Salmonella enterica subsp. enterica serovar Kentucky str. CVM29188
pAA	scaffold16_orf0047	hypothetical protein R100p115.2br		Plasmid R100
pAA	scaffold16_orf0048	Protein traJ		Escherichia coli 55989
pAA	scaffold16_orf0049	TraM		Escherichia coli
pAA	scaffold16_orf0050	putative lytic transglycosylase		Escherichia coli ETEC H10407
pAA	scaffold16_orf0051	conserved hypothetical protein		Escherichia coli MS 185-1
pAA	scaffold16_orf0052	putative recombinase		Escherichia coli
pAA	scaffold16_orf0053	SepA		Escherichia coli 536
pAA	scaffold16_orf0054	putative transposase		Escherichia coli
pAA	scaffold16_orf0057	conserved hypothetical protein		Escherichia coli MS 153-1
pAA	scaffold16_orf0058	AatD		Escherichia sp. 4_1_40B
pAA	scaffold16_orf0059	AatC ATB binding protein of ABC transporter		Escherichia coli 55989
pAA	scaffold16_orf0060	AatB		Escherichia coli 55989
pAA	scaffold16_orf0061	AatA outermembrane protein		Escherichia coli 55989
pAA	scaffold16_orf0062	AatP permease		Escherichia sp. 4_1_40B
pAA	scaffold16_orf0063	serine protease eatA		Shigella dysenteriae 1617
pAA	scaffold16_orf0064	protease IgA1		Escherichia coli
pAA	scaffold16_orf0065	hypothetical protein E4_23001		Escherichia sp. 4_1_40B
pAA	scaffold16_orf0066	Serine protease sat precursor (Secreted autotransporter toxin sat) (fragment)		Escherichia coli 55989
pAA	scaffold16_orf0067	ISPsy2, transposase		Escherichia coli MS 185-1
pAA	scaffold16_orf0069	14 kDa aggregative adherence fimbriae I protein (Fragment) (modular protein)		Escherichia coli 55989
pAA	scaffold16_orf0070	putative transposase domain protein		Escherichia coli 3431
pAA	scaffold16_orf0071	Serine protease sepA precursor (fragment)		Escherichia sp. 4_1_40B
pAA	scaffold16_orf0072	IS186 transposase		Escherichia coli UMNK88
pAA	scaffold16_orf0073	CvaB, IS186 transposase		Escherichia coli BW2952
pAA	scaffold16_orf0074	hypothetical protein		Escherichia coli
pAA	scaffold16_orf0075	hypothetical protein Mtub2_09757		Mycobacterium tuberculosis 210
pAA	scaffold16_orf0076	putative IS639 ORF1		Escherichia coli ETEC 1392/75
pAA	scaffold16_orf0077	putative transcriptional activator aggR (AAF-III) regulatory protein)		Escherichia coli 55989
pAA	scaffold16_orf0078	transposase ORF A, IS1		Escherichia coli 55989

pAA	scaffold16_orf0079	transposase		Escherichia coli M863
pAA	scaffold16_orf0080	hypothetical protein Mtub2_09757		Mycobacterium tuberculosis 210
pAA	scaffold16_orf0081	hypothetical protein E4_23056		Escherichia sp. 4_1_40B
pAA	scaffold16_orf0083	putative transposase (fragment)		Escherichia coli 55989
pAA	scaffold16_orf0084	putative Isopentenyl-diphosphate delta-isomerase (IPP isomerase) (Isopentenyl pyrophosphate isomerase) (IPP:DMAPP isomerase)		Escherichia coli 55989
pAA	scaffold16_orf0085	hypothetical protein pEC55989_0080		Escherichia coli 55989
pAA	scaffold16_orf0086	conserved hypothetical protein		Escherichia coli MS 119-7
pAA	scaffold16_orf0087	transposase		Escherichia coli M863
pAA	scaffold16_orf0088	putative transposase insK for insertion sequence element IS150		Shigella sonnei 53G
pAA	scaffold16_orf0089	putative protein encoded within IS		Shigella sonnei Ss046

Chapter 5

Performance comparison of benchtop high-throughput sequencing platforms

Performance comparison of bench-top high-throughput sequencing platforms

Nicholas J. Loman[1], Raju Misra[2], Tim Dallman[2], Chrystala Constantinidou[1], Saheer Gharbia[2], John Wain[2,3*], and Mark J. Pallen[1*]

[1]Centre for Systems Biology, University of Birmingham, Birmingham, B15 2TT, United Kingdom

[2]Health Protection Agency, 61 Colindale Avenue, London, NW9 5HT, United Kingdom

[3]School of Medicine, University of East Anglia, Norwich, NR4 7TJ, United Kingdom

* Joint corresponding authors, e-mail for correspondence m.pallen@bham.ac.uk and j.wain@uea.ac.uk

Abstract

Three bench-top high-throughput sequencing instruments are now available. The 454 GS Junior (Roche), MiSeq (Illumina) and Ion Torrent PGM (Life Technologies) are laser-printer sized and offer modest set-up and running costs. Each instrument can generate a draft bacterial genome sequence in days, making them attractive for use in the identification and characterization of pathogens in the clinical setting. We compared the performance of these instruments by sequencing isolates of *Escherichia coli* O104:H4 from the German outbreak of 2011. We compared performance of the platforms, analysing throughput, read length, read error profile and rate, *de novo* assembly quality and completeness. MiSeq had the highest throughput and lowest error rate. The 454 Junior generated the longest reads and best assemblies. The Ion Torrent PGM produced intermediate throughput with the shortest reads. The Ion Torrent PGM and 454 GS Junior both suffer from errors in homopolymers.

Over the past decade and a half, genome sequencing has transformed almost every corner of the biomedical sciences, including the study of bacterial pathogens [1]. In the last five years, high-throughput (or "next-generation") sequencing technologies have delivered a step change in our ability to sequence genomes, whether human or bacterial [2, 3]. Since arriving in the market place, these technologies have experienced sustained technical improvement, which, twinned with lively competition between alternative platforms, has placed genome sequencing in a state of permanent revolution.

Although high-throughput sequencing has seen extensive use in bacteriology, e.g. in the genomic epidemiology of bacterial pathogens [4], until recently sequencing platforms were tailored chiefly towards large-scale applications, focused on the race to the "$1,000 human genome", with footprints, workflows, reagent costs and run times poorly matched to the needs of small laboratories studying small genomes. However three different bench-top high-throughput sequencing instruments are currently available, all roughly the size of a laser printer, with modest set-up and running costs and all capable of sequencing bacterial genomes in a matter of days (Table 1).

The 454 GS Junior from Roche was released in early 2010 and is a smaller, lower-throughput version of the 454 GS FLX machine, exploiting similar emulsion PCR and pyrosequencing approaches, but with lower set-up and running costs. The Ion Torrent Personal Genome Machine (PGM) was launched in early 2011 [5]. Like the 454 GS Junior, this technology exploits emulsion PCR. This platform also incorporates a sequencing-by-synthesis approach, but uses native dNTP chemisty and relies on a modified silicon chip to detect hydrogen ions released during base incorporation by DNA polymerase (making it the first "post-light" sequencing instrument). The MiSeq (Illumina) was announced in January 2011 and began to ship to customers in the fourth quarter of 2011. The MiSeq is based on the existing Solexa sequencing-by-synthesis chemistry [6] but has dramatically reduced run times compared to the Illumina HiSeq (fastest run 4 hours versus 1.5 for 36-cycle sequencing or 16 hours versus 8.5 days for 200-cycle sequencing) made possible by a reduced size flow cell, reduced imaging time and faster microfluidics.

We wished to compare the performance of these three sequencing platforms by analysing data with commonly used assembly and analysis pipelines. We therefore benchmarked these platforms by using them to genome-sequence isolates from the recent outbreak of Shiga-toxin-producing *E. coli* (STEC) O104:H4 that struck Germany between May and July 2011. This outbreak was responsible for over 4000 infections and more than 40 deaths [7]. Previous whole-genome sequencing efforts applied to isolates from the outbreak yielded novel diagnostic reagents and provided important clues as to the nature, origins and evolution of the outbreak strain [8–12]. These efforts also demonstrated the utility of an "open-source" approach to outbreak genomics that included rapid sequencing, a liberal approach to data release and use of crowdsourcing [10]. Although all infections during the outbreak were acquired in Germany, travellers took their infections back to other countries in Europe and North America, including the United Kingdom [7]. Here, we have focused on a single *E. coli* isolate of serotype O104 from the United Kingdom epidemiologically linked to the German outbreak.

Results

Creation of reference assembly

To permit comparisons of bench-top sequencing data we generated a reference assembly for *E. coli* O104:H4 280 (HPA materials identifier H112160280) using established high-throughput sequencing platforms. This strain was recovered from a female traveller returning from Germany who had developed hemolytic uremic syndrome and thrombotic thrombocytopenic purpura. The strain was confirmed as typical of an outbreak strain (ST678, *stx-2* ositive and intimin negative) [13].

We used the Roche 454 GS FLX+ system to generate very long fragment reads (modal read length 812bp bases, maximum read length 1170 bases) to an estimated 32-fold mean coverage. Additionally, Roche 454 GS FLX was used to sequence an 8kb insert paired-end library using Titanium chemistry. The reads were assembled into contigs, which were scaffolded to produce a draft reference assembly. The use of abundant long reads and long-insert paired-end information plus error correction from a complementary sequencing technology resulted in a very high quality draft genome assembly consisting of three scaffolds. 99.42% of the bases in the assembly are Q64 bases (the highest quality assigned by Newbler, representing accuracy of one miscall around every 2.5m bases), 99.6% are Q30 or higher. Lower quality bases were masked with a lower-case letter. The largest scaffold corresponded to the chromosome (5,340,022 bp), the two smaller scaffolds corresponded to two large plasmids (pESBL and pAA). The 1.5kb plasmid sequence was present in a single contig. Although each scaffold represented a single circular replicon, 153 gaps remained within the scaffolds. These gaps represent repetitive regions longer than the mean read length and shorter than the paired-end insert library and which cannot be resolved by this sequencing strategy.

Characteristics of reads from bench-top sequencers

Genome depth, evenness of coverage, read length and read quality are the four major factors which determine the ability to reconstruct genome sequences from sequence data. There were large differences in the number, predicted quality and length of reads obtained from the three platforms (Table 2, Figure 1). 454 Junior produced the longest reads, with a mean length of 522 bases, but had the lowest throughput of the three instruments. Ion Torrent PGM runs generated over four times the throughput of 454 Junior but generated the shortest reads (mean 121 bases). MiSeq produced the greatest throughput with reads slightly longer than Ion Torrent PGM, permitting the multiplexing of seven *E. coli* strains on a single run. MiSeq reads were paired-end: that is, fragments were sequenced in both directions. Across the reference chromosome, coverage was generally even although in the MiSeq data we saw a peak associated with the Shiga-toxin producing phage, a smaller peak was detectable in the Ion Torrent PGM data (Supplementary Figure 3). Differences in relative coverage levels were also seen in the pESBL and pAA plasmids between instruments.

Because each manufacturer uses a unique software implementation to generate base quality score predictions, direct comparison of these scores between platforms is difficult. We recalibrated quality scores for each instrument by first aligning reads to the reference genome. By observing the counts of matched and mismatched bases in each aligned read a new quality score can be calculated

(alignment quality, AQ). We used the scoring system of Ewing and Green which scores both substitutions, insertions and deletions. Mismatches resulting in deletions are assigned randomly to the position of one of the adjacent bases in the read. Alignment quality scores predicted in this way generally had good agreement with predicted scores, with the Ion Torrent PGM generally underestimating accuracy and the other instruments slightly overestimating (Figure 2). The MiSeq produced the highest quality reads, due to a low substitution error rate and the near absence of indel errors compared to the other platforms. The Ion Torrent PGM showed a steadily decreasing accuracy across the read to 100 bases. The accuracy seems to improve after this point due to the aligner soft-clipping trailing bases. Comparison of the frequency of indels through alignment to the reference demonstrated Ion Torrent PGM reads had 1.5 indels per 100 bases (1.72 indels per read). The 454 Junior had 0.38 indels per 100 bases (1.74 indels per read). In contrast, indels were detected very infrequently in MiSeq data with $<$0.001 indels per 100 bases. These results were confirmed by alignment to two other reference genomes sequenced with other sequencing technologies (see Supplementary Materials). As with 454 sequencing, the major source of indels in Ion Torrent PGM data are runs of identical bases (homopolymers). Comparison of homopolymer accuracy between Ion Torrent PGM and 454 Junior demonstrated that Ion Torrent PGM was less accurate when calling homopolymers of any length (Figure 3). The dominant source of error were deletions, with accuracy rates as low as 60% for homopolymers of length six or grater.

Comparison of *de novo* assemblies

The use of high-throughput sequencing for the discovery of differences in gene content and arrangement relies on the generation of accurate *de novo* assemblies. We compared draft, *de novo* assemblies from bench-top instruments using a variety of metrics. Assembly metrics such as total assembly size and N50 [14] give a guide to assembly completeness or fragmentation but not accuracy. An ideal assembly produces a single accurate contig for each replicon but this is rarely possible due to the presence of long repeat sequences. When comparing bench-top *de novo* assemblies we saw two major groupings of assembly quality. Heavily fragmented assemblies were obtained from with Ion Torrent data (single runs or combined), 454 Junior (single runs) and MiSeq contigs. Less heavily fragmented assemblies were obtained when reads from two 454 Junior runs were combined to increase depth of coverage and when paired-end information was used to scaffold contigs generated from the MiSeq data. However, runs of ambiguous bases were seen in the scaffolded MiSeq assemblies, unlike the assemblies obtained from the 454 Junior data.

The number of contigs that can be mapped unambiguously to the reference gives a measure of genome coverage. Differences in genome coverage were seen when comparing assemblies from each platform (Table 3). No platform delivered data that aligned unambiguously to 100% of the reference. Contigs obtained from the 454 Junior data aligned to the largest proportion of the reference, with 5.4% of the reference unmapped. This compared to 6.5% for Ion Torrent PGM and 5.9% for MiSeq.

The Ion Torrent PGM assemblies had large numbers of gaps (Figure 5), compared to assemblies obtained from 454 Junior and MiSeq data. Increasing sequence coverage by combining assemblies from the two Ion Torrent PGM runs reduced the numbers of gaps in the assembly. However this had little effect on the miscalls in long homopolymeric tracts, so that even in this

combined Ion Torrent PGM assembly, around 10% of the coding sequences (as predicted from the reference assembly) were disrupted either by contig breaks or apparent frameshifts. Of the 1,864 gaps seen in the combined Ion Torrent PGM assembly around a quarter were due to gaps associated with ends of contig or unmapped sequence, the rest being associated with homopolymeric tracts. Manual inspection of assembly alignments revealed that many of the indels associated with short homopolymeric tracts demonstrated strand bias, with the correct call predominantly associated with either forward or reverse reads and the erroneous sequences associated with the opposite strand (Supplementary Figure 2). While problems with homopolymers are known to result from flow-based chemistries, it is unclear why this strand bias should occur with Ion Torrent technology. However, scrutiny of other public data sets from this instrument (http://mira-assembler.sourceforge.net/docs/DefinitiveGuideToMIRA.html) suggests it is a pervasive problem.

How useful are bench-top assemblies for public health microbiology?

A key test for a genome-sequencing technology is whether it can deliver trustworthy new insights into the biology of the organism under scrutiny. We therefore evaluated how *de novo* assemblies from each platform performed in reporting features of biological interest in the outbreak strain. For some features, all platforms did well–for example all documented the presence and accurate full-length sequence of the genes encoding the Shiga toxin type-2 subunits. However, at the other extreme, all instruments did badly–for instance, in all assemblies the two larger plasmids were broken into multiple contigs, which could not be readily assigned to chromosome or plasmid without alignment to the reference genome.

We used 31 protein sequences linked to pathogen biology as queries in translated BLAST searches of the assemblies obtained from the bench-top sequencing platforms (Supplementary Table 1 and Supplementary Files **??**. No assembly contained a full set of full-length sequences. The best MiSeq assembly captured 28/31 full-length sequences; the best 454 Junior assembly found 26 and the best Ion Torrent PGM assembly found 22. Perhaps the most challenging targets in the survey were the four serine protease autotransporters encoded in the genome of the outbreak strain. These genes code for multiple-domain proteins. None of the platforms managed to recover all four genes as full-length fragments: the Ion Torrent PGM assembly recovered only one of them. This is because the SPATEs are multiple-domain proteins and some domains exist as multiple copies in the genome which are assembled into repeat consensus contigs which cannot be unambiguously placed in the genome.

Integration of whole-genome sequencing into existing practice in a public health laboratory requires backwards compatibility with existing typing methods. We therefore attempted to generate multi-locus sequence typing (MLST) profiles from each assembly. An accurate MLST profile was generated for the outbreak strain by the 454 Junior and MiSeq. However, all Ion Torrent PGM assemblies generated indel errors in at least one housekeeping gene.

Discussion

Sequencing and Public Health 2.0

In our evaluation, all three benchtop sequencing platforms generated useful draft genome sequences of the German *E. coli* outbreak strain. All could be judged "fit for purpose" in producing assemblies that mapped to 93% or more of the reference genome and recovered the vast majority of coding sequences. However, no instrument could on its own generate completely accurate one-contig-per-replicon assemblies that might equate to a finished genome. Thus, for each technology there is a trade-off between advantages and disadvantages. In our survey, the MiSeq generated the highest throughput per run and lowest error rate of the instruments, without significant indel or substitution errors (although accuracy does drop off toward the ends of reads). However, the MiSeq delivered shorter read lengths, and thus worse assemblies, than the 454 Junior. Even with paired-end sequencing, the single scaffold assemblies from the MiSeq are interrupted by unfillable gaps, representing difficult-to-resolve repeats. Furthermore, paired-end 150 base sequencing on a pre-release instrument took over 27 hours (60 megabases per hour). The 454 Junior delivered the longest read length but the lowest throughput (eight megabases per hour during a nine-hour run) and suffered from errors in homopolymeric tracts, even at high coverage. The Ion Torrent PGM produced intermediate throughput with the shortest reads and the worst performance with homopolymers. However, it delivered the fastest throughput (80-100 megabases per hour) and shortest run time (around 3 hours). This platform has also shown the greatest improvement in performance in recent months-an assembly for the outbreak strain generated in May 2011 from data from the original Ion Torrent 314 chip contained >3000 contigs [10], whereas, in this study, data from the recently available 316 chip assembled into <600 contigs.

Speed, set-up and running costs and ease of workflow are also important factors when comparing these platforms. However, as these may vary from one time or place to another and may be subject to rapid changes, it is harder to make objective durable evaluations on these criteria. Nonetheless, whatever the setting, the cost per base of generating sequence data appears to be an order of magnitude higher for the 454 Junior than the other two platforms. The MiSeq workflow has the fewest manual steps due to the bridge amplification occurring on the instrument as the initial step of sequencing, whereas Ion Torrent PGM and 454 GS Junior require a sequence-ready library which has been amplified through emulsion PCR and subsequently enriched. All three platforms have protocols for generating and sequencing long mate-pair libraries (templates with ends a fixed distance apart in the genome). Since this study was performed, a paired-end protocol for the Ion Torrent PGM has been announced similar to that on the MiSeq which requires a second sequencing reaction to be carried out immediately after the first which also has the effect of doubling the run-time (http://www.iontorrent.com/lib/images/PDFs/pe_appnote_v12b.pdf).

One important conclusion from this evaluation is that saying that one has "sequenced a bacterial genome" means different things on different benchtop sequencing platforms. Potential users of these technologies need to be sensitive to these differences, particularly when comparing or combining data generated on different platforms. Other important questions include how far can errors be corrected by comparison to reference data, when is it safe to use a mapping approach that makes assumptions that a novel sequence is like an existing reference sequence and how much should one have to rely on human insight rather than automated analyses and pipelines?

In this study, we set a tough test by evaluating algorithmically generated *de novo* assemblies. However, during the real-world test case of the German *E. coli* outbreak, even the first-generation Ion Torrent platform, with its low throughput and high error-rate, delivered useful insights into the biology and evolution of the outbreak strain [9, 10]. For example, a homopolymer error in an MLST profile was easily corrected by expert opinion. We are thus confident that benchtop high-throughput sequencing platforms are poised to make a decisive impact on diagnostic and public health microbiology in the near future.

Author contributions

N.J.L, J.W, S.G and M.J.P conceived the experiments, J.W. and S.G. supplied the strains, N.J.L, R.M. and T.D. performed the bioinformatics analysis, C.C. performed the Ion Torrent sequencing, S.G. and R.M. perfomed the 454 GS Junior sequencing. N.J.L. and M.J.P. wrote the manuscript. All authors commented on the manuscript.

Accession codes

454 sequences have been deposited into the Short Read Archive under study number SRA048574, with run accessions SRR388806 (454 GS Junior run 1), SRR388807 (454 GS Junior run 2), SRR388808 (454 FLX+), SRR388809 (454 Titanium 8kb paired-end). Ion Torrent PGM sequences have been deposited under study number SRA048511, with accessions SRR389193 (Ion Torrent PGM run 1), SRR389194 (Ion Torrent PGM run 2). The multiplexed MiSeq reads have been deposited under study number SRA048664. Assembly files and analysis scripts have been uploaded to a public Github repository (https://github.com/nickloman/benchtop-sequencing-comparison).

Acknowledgements

We gratefully acknowledge the blogging community for helpful discussion in the comments section of our blog (http://pathogenomics.bham.ac.uk/blog), and in particular to Bastien Chevreux, Justin Johnson, Keith Robison and Lex Nederbragt. We are grateful to Colin Hercus at Novocraft for help with the Novoalign software and to Aaron Darling for help with Mauve Assembly Metrics. We thank Roche Diagnostics, UK for 454 GS FLX+ and 454 FLX paired-end sequencing, technical support and helpful discussion. We thank Life Technologies for early access to 316 chips and instrument fluidics upgrade. We thank Geoff Smith and Illumina UK for early access to the MiSeq platform and public release of *E. coli* outbreak strain data.

Tables

Table 1: Price comparison of bench-top instruments and sequencing runs. Note pricing may vary between countries/sales territories. Instrument prices do not include service contracts. Sample prices do not include cost of generating the initial fragmented genomic DNA library with adapters (an additional cost of between $50-200 depending on method used). Cost per megabase assumes one sample and one sample sequencing kit per run.

Platform	List price	Approximate cost per run	Minimum throughput (read length)	Run time	Cost per megabase	Megabases per hour
454 GS Junior	$108,000	$1100	35Mb (400 bases)	8 hours	$28	4.375
Ion Torrent PGM[1]	$80,490[2]	$225[3] (314 chip)	10Mb (100 bases)	3 hours	$22.5	3.33
		$425 (2) (316 chip)	100Mb* (100 bases)	3 hours	$4.25	33.3
		$625 (318 chip)	1000Mb (100 bases)	3 hours	$0.63	333.3
MiSeq	$125,000	$750 (2 x 150 bases)	1500 Mb	27 hours	$0.5	55.5

Unless stated, pricing information is from the online supplement of [3].

(1) Ion Torrent PGM pricing from Invitrogen US territory website (www.invitrogen.com, accessed 21st February 2012).

(2) Price includes Ion Torrent PGM, server, OneTouch and OneTouch ES sample automation systems.

(3) Ion Torrent PGM prices includes chip and sample preparation kit.

* Configuration used in this study.

Table 2: Bench-top sequencing results. Metrics for each sequencing run are shown as well as results of alignment against the reference sequence. Depth of coverage for the chromosome and two large plasmids are shown with the percentage of reads which align. For the MiSeq run the sequence metrics are shown for the entire run as well as the results of de-multiplexing the seven *E. coli* strains.

					Alignment coverage		
Run	Reads	Total bases	Modal length	Mean length (s.d.)	Chromosome	Large plasmids	Reads aligned (%)
454 GS Junior (1)	135,992	70,999,968	518	522 (46)	11.50	5.66	99
454 GS Junior (2)	137,528	71,710,564	516	521 (47)	11.54	5.39	99
Ion Torrent PGM (1)	2,483,868	303,579,279	123	122 (11)	46.60	53.33	90
Ion Torrent (2)	2,154,577	260,017,346	123	120 (16)	39.33	43.80	89
MiSeq (1)	11,708,156	1,652,529,000	150	141 (22)	-	-	
			Demultiplexed MiSeq data				
... strain 280	1,766,516	250,356,566	150	141 (21)	22.11	625.46	99

Table 3: Comparison of assembly metrics and quality. Breakpoints relate to the number of putative misassemblies indicating where different parts of the same contig align to different parts of the reference genome. Unmappable bases are from contigs which cannot be aligned unambiguously to the reference genome. Broken CDS is the number of coding sequences in the reference which are split into two or more fragments either due to contig ends or indel errors.

Name	Contigs	N50	Breakpoints	Gaps in reference	Gaps in assembly	% unmappable bases	Broken CDS
454 Junior (1)	362	50603	3	350	455	6.5404	141
454 Junior (2)	415	39823	1	453	595	6.0885	196
454 Junior (1+2)	216	119080	3	162	270	5.433	83
Ion Torrent (1)	553	58782	1	431	1831	6.5738	597
Ion Torrent (2)	589	51211	4	445	1863	6.4979	634
Ion Torrent (1+2)	577	55431	1	376	1488	6.5965	516
MiSeq (contigs)	612	46119	9	284	289	5.902	134
MiSeq (scaffolds)	505	111638	17	343	337	5.9097	146

List of figures

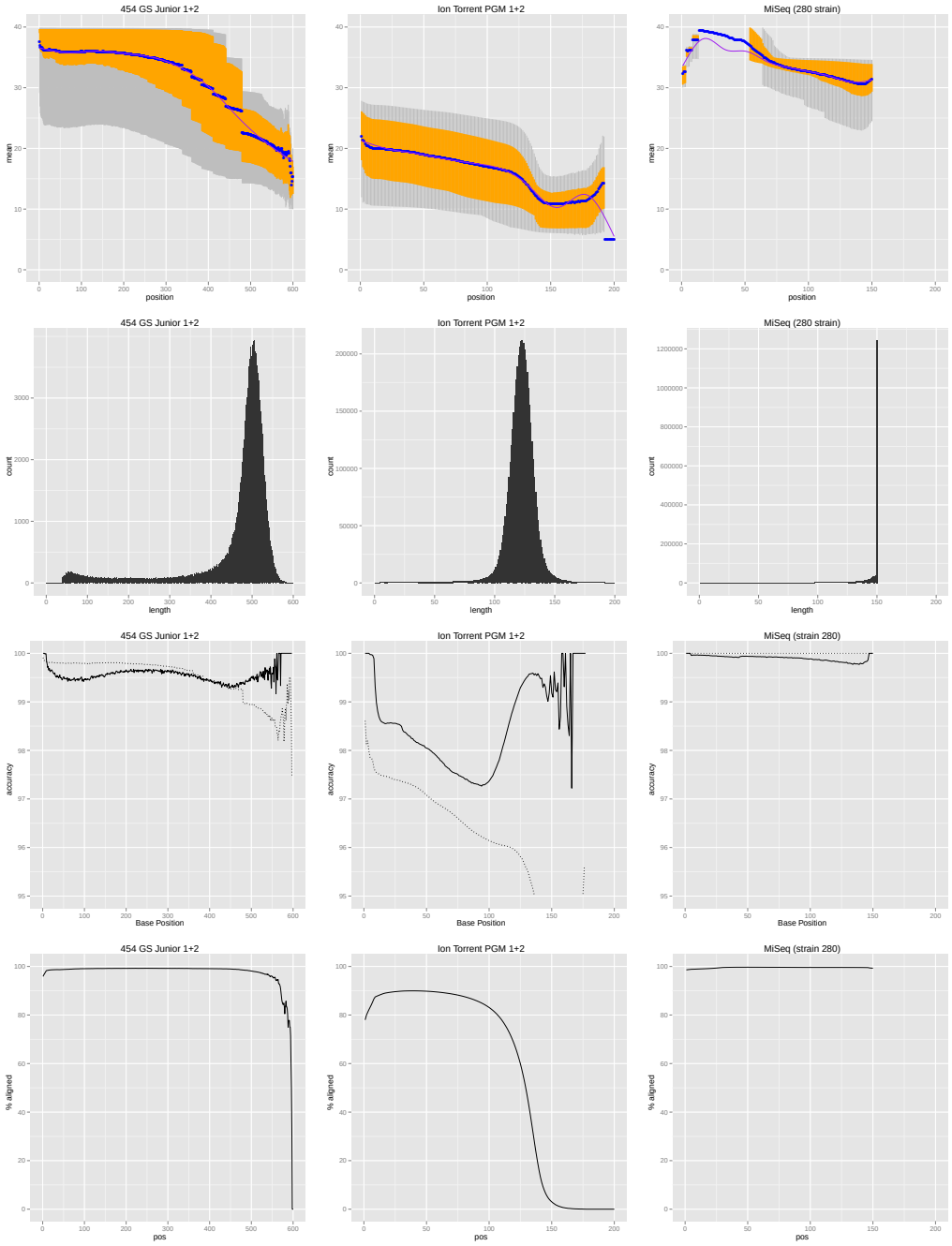


Figure 1: Evaluation of read length and quality from bench-top sequencers. Row A) Boxplots showing the predicted per-base quality score for combined sequencing runs for each bench-top instrument at each read position created by the qrqc package. Grey lines indicate the 10% and 90% quantiles, orange lines indicate the lower and upper quartiles, the blue dot is the median, and the green dash the mean. A purple smooth curve is fit through the distributions [15]. Quality scores are given as Phred-scaled quality values where $Q = -10\ \log_{10} P$ (P being the probability of the base call being correct). Row B) Histograms showing read lengths produced by each instrument. Row B) Comparison of the predicted and measured accuracy for each benchtop sequencer. Predicted accuracy is determined by multiplying the number of alignments of bases of each quality score by the probability of an incorrect base call ($10^{\frac{-Q}{10}}$). Row C) The percentage of reads aligned at any

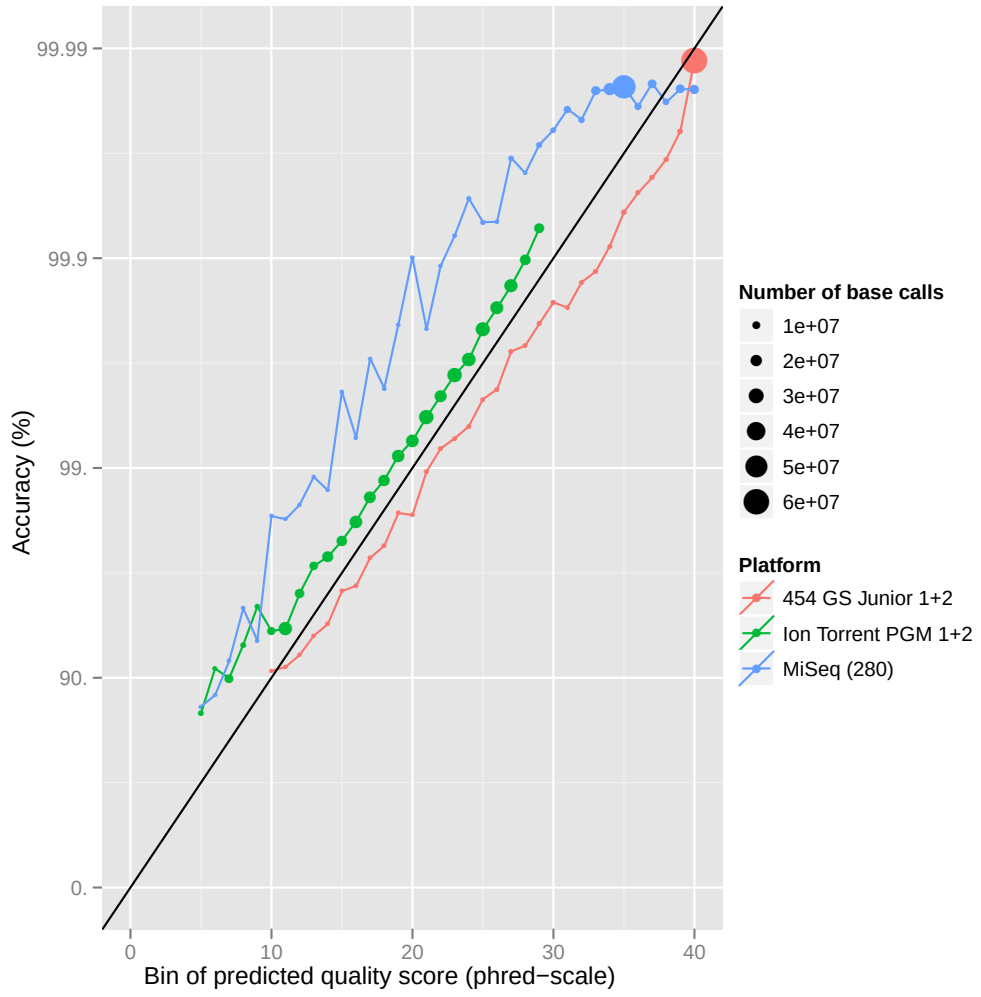


Figure 2: Chart showing the relationship between predicted quality scores and measured base accuracy. The area of each point shows the number of aligned bases in the predicted quality score bin. The diagonal slope indicates the relationship between base quality scores and accuracy.

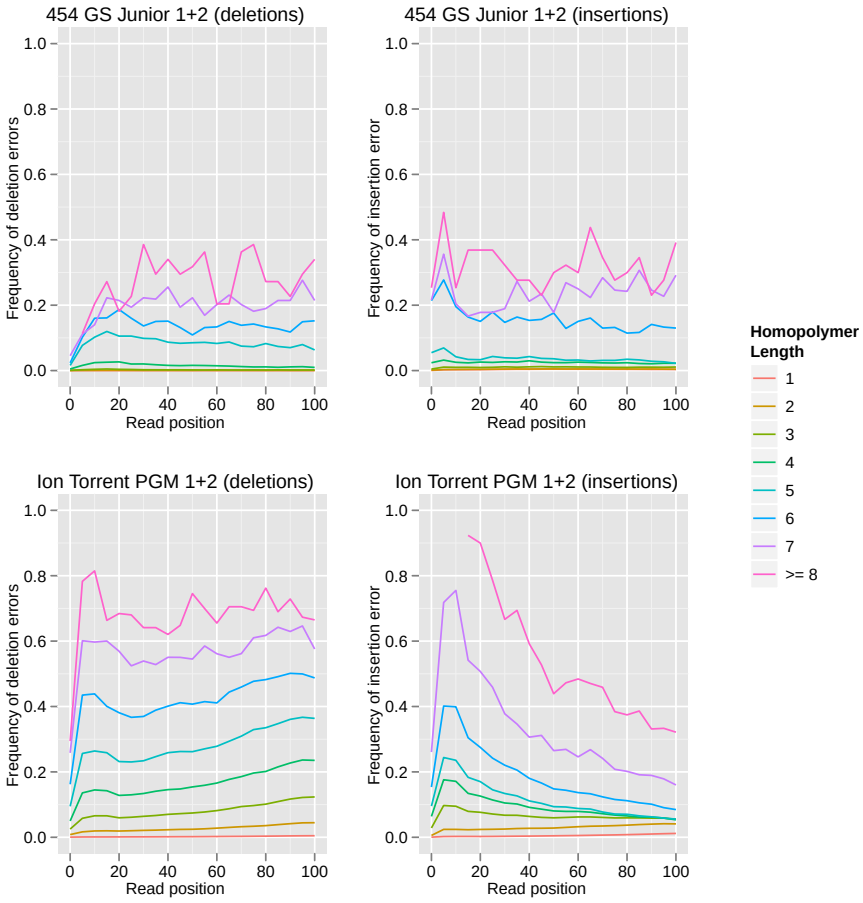


Figure 3: Comparison of homopolymer tract accuracy between 454 Junior and Ion Torrent. Charts show the frequency of erroneous insertions or deletions associated with homopolymeric tracts in the reference genome of lengths 1-7, and 8 or greater.

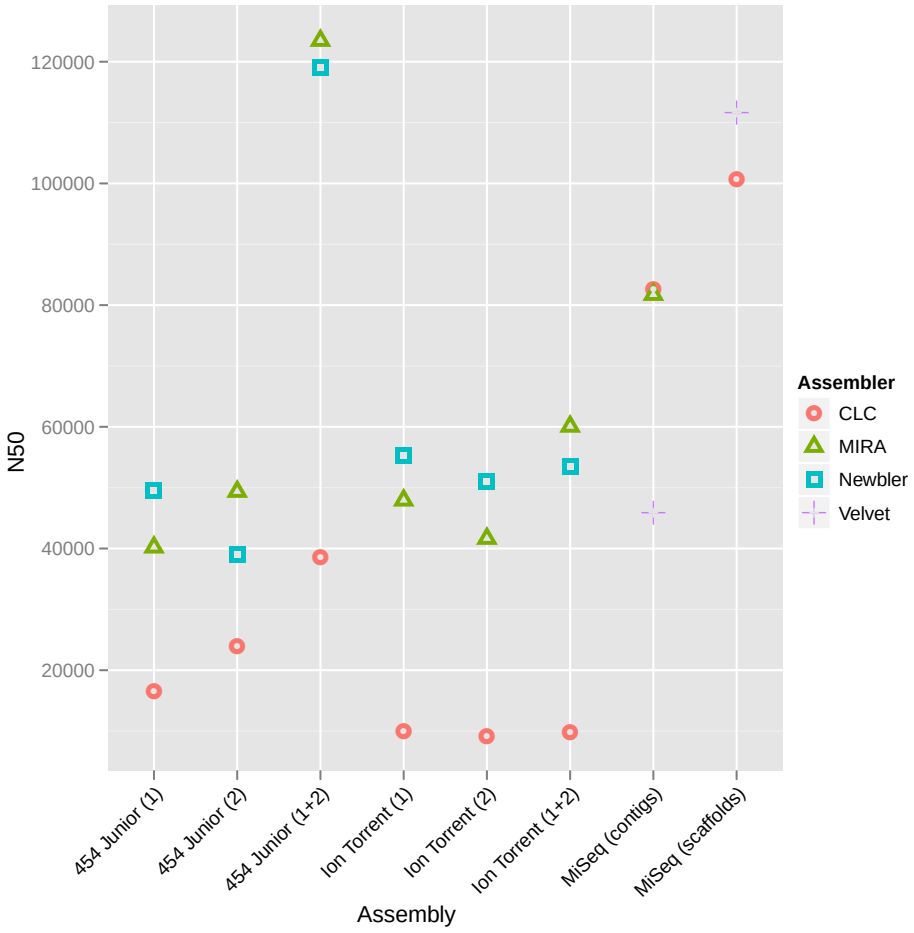


Figure 4: This plot shows N50 values from assemblies generated from sequence data for each sequencing platform. A selection of popular assembly software has been used. The N50 is calculated from the total genome length of the *E. coli* strain 280 reference sequence, rather than the sum total of contig lengths.

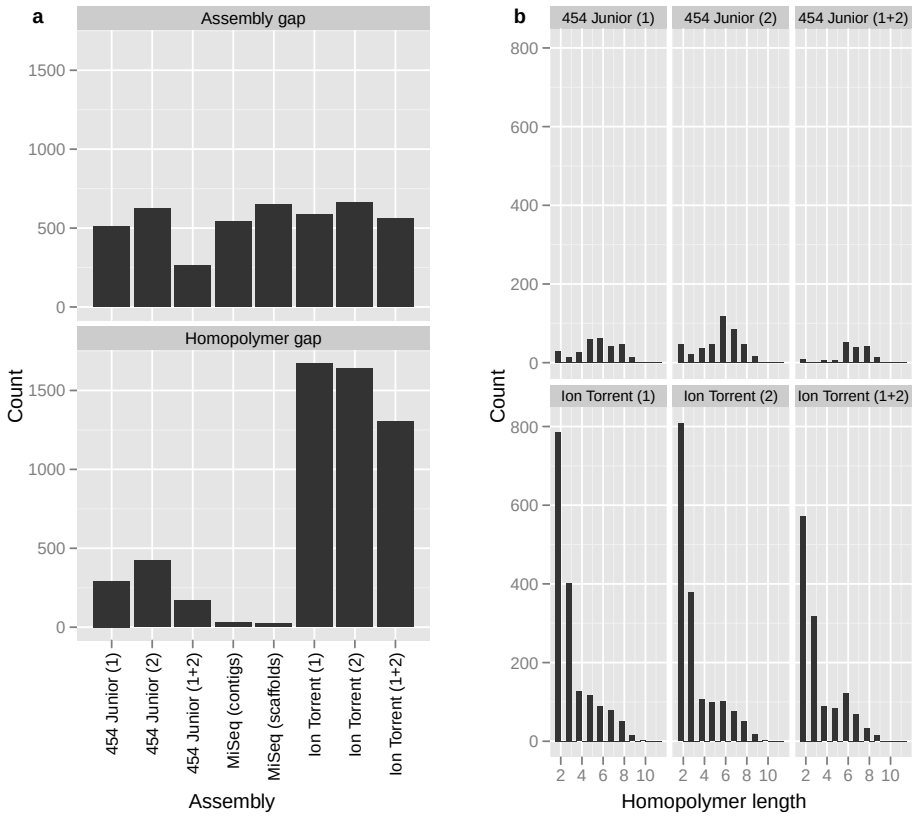


Figure 5: An analysis of gaps when aligning draft *de novo* assemblies to the reference genome. A - top panel) The number of gaps which are not associated with homopolymeric tracts, e.g. contig breaks, misassemblies, missing sequence. A - bottom panel) The number of gaps which are associated with homopolymeric tracts for each draft assembly. B) The length of erroneously called homopolymeric tracts for each 454 Junior and Ion Torrent assembly.

Methods

Collection of isolates

Five UK isolates, all with epidemiological links to the German outbreak were included in the study, with strain 280 being sequenced by each of the bench-top instruments (Supplementary Table **??**). Two *E. coli* O104 isolates were not linked to the German outbreak and were included as comparators. Isolates were grown according to the protocol described in Chattaaway *et al* [13]. To generate enough DNA for sequencing, isolates were grown on multiple occasions.

Sequencing workflow

A general, simplified workflow for library preparation, amplification and sequencing is shown (Supplementary Figure 1) with approximate timings for each stage. These stages comprise library preparation from genomic DNA, amplification and sequencing. Library preparation steps are similar for each instrument, involving extraction and purification of genomic DNA, fragmentation through either enzymatic or physical means, fragment size selection and ligation of sequencing adapters.

Ion Torrent Sequencing

Ion Torrent sequencing was performed at the University of Birmingham according to the Ion Torrent protocol (Life Technologies, Gaithersburg, MD). Total DNA from *E. coli* O104:H4 280 was isolated. 10 µg of this DNA was fragmented with a Bioruptor instrument (Diagenode, Liège, Belgium) using the protocol recommended by Life Technologies. A broad profile of fragment sizes (75-500 bp, peak at 255 bp) were obtained which were end-repaired, ligated with Ion Torrent A and P1 adapters and size selected using E-Gel EX 2% Gel (Invitrogen, Carlsbad, CA) for 150-250 bp fragments. The size-selected fragments were amplified and DNA was purified with Agencourt AMPure XP beads (Beckman Coulter Genomics, High Wycombe, UK). The median fragment size of the final library was 200 bp (assessed by a BioAnalyzer High Sensitivity LabChip, Agilent). Library was diluted to 40pM and two emulsion PCR reactions were set up at two templates per sphere. Sequencing primer and polymerase were added to the final enriched spheres prior to loading onto the 316 chip. Two 316 chips were run in total. Base calls were generated using version 1.5 of the Ion Torrent software suite and for further analysis the resulting flowgram files (assembly) or FASTQ files (alignment) were used.

454 Junior Sequencing

454 Junior sequencing was performed on an instrument at the Health Protection Agency, Colindale, UK. *E. coli* O104:H4 280 DNA was prepared following the Roche Rapid Library protocol (Roche, Welwyn Garden City, UK), whereby 5 ng/µl was taken from each sample and libraries prepared. Briefly, samples were subjected to the following key steps: DNA fragmentation by nebulization, fragment end-repair, AMPure XP bead preparation (Amersham International, Buckinghamshire, UK), adaptor ligation, small fragment removal, quality assessment using the Agilent 2100 Bioanalyzer, library quantitation and finally preparation of working aliquots at a final concentration

of 1×10^7 molecules/µl (500 ng total). Emulsions PCR, enrichment and 454 GS Junior sequencing were carried out as per manufacturer's protocols. The resulting flowgram files were used for downstream analysis.

454 GS FLX+ and 454 GS FLX 8kb Titanium sequencing

454 GS FLX 8kb Titanium paired-end and 454 FLX+ (long read) library construction and sequencing was performed at Roche Diagnostics (Burgess Hill, UK) according to their standard protocols.

Illumina MiSeq Sequencing

Illumina MiSeq sequencing was performed at Illumina UK, Little Chesterford, UK, on a pre-release, prototype MiSeq instrument. The seven *E. coli* samples were quantified with a Qubit High Sensitivity kit and the total amount of DNA for each sample varied between 523ng and 954ng. Samples were sheared with Covaris followed by end repair, A-tailing and the ligation of Truseq adapters containing indexes. Samples were run on a 2% agarose gel (2 samples per gel) and DNA was size selected at 600-700 bp. 10 cycles of PCR were carried out and samples run out on a second 2% agarose gel (2 samples per gel). Samples were excised from the gel and quantified with a Qubit high sensitivity kit. Libraries were diluted to 2nM in EB+0.1% tween and a pool containing an equimolar concentration of each library was prepared. MiSeq instrument was prepared following routine procedures. Briefly, a standard MiSeq flowcell was inserted into the flowcell chamber. Next, the DNA sample containing the pool of seven *E. coli* libraries was diluted to 6.2 pmol and pipetted into the sample well on the MiSeq Consumable Cartridge before loading in the chiller section of the MiSeq instrument. A sample sheet was prepared on the MiSeq instrument to provide run details. The run was initiated for 2x151 bases of SBS sequencing, including on-board clustering and paired-end preparation, the sequencing of the seven barcode indices, and analysis. On the completion of the run, data was basecalled and demultiplexed on the instrument (provided as Illumina FASTQ files, phred+64 encoding). FASTQ format files in Illumina 1.5 format were considered for downstream analaysis. Although MiSeq produces reads of fixed lengths, tails of these reads may be designated as uncallable as indicated by the read segment quality control indicator, noted a quality score of two ('B'). In these cases these low quality tails were trimmed and not used for further analysis.

Bioinformatics

Construction of reference assembly

A high-quality reference sequence for *E. coli* strain 280 was constructed by assembling 454 FLX+ long read data and 454 Titanium paired-end data (8kb insert) using Newbler 2.6. Newbler was run with parameters `-scaffold -tr -cpu 8 -siom 28 -rip`. The resulting scaffolds were used for further analysis. Newbler masks certain bases in the assembly regarded as uncertain by assigning it a lower-case nucleotide. These masked bases correspond with bases with a low quality score. In bacterial genomes these bases are seen predominantly in consensus contigs resulting from long repeat regions, long homopolymeric tracts and contig ends. The resulting assembly was annotated

using the automated xBASE annotation pipeline [16] which utilises Glimmer for coding sequence prediction [17] and tRNAScan-SE and RNAmmer for stable RNA prediction [18, 19].

De novo assembly of individual strains

Assemblies were generated from data generated by each of the bench-top sequencing platforms separately. All data were assembled by MIRA 3.4.0 using default parameters in `genome,denovo,accurate` mode and the appropriate setting for each instrument type (`454,iontor,solexa`). Ion Torrent and 454 Junior data were additionally assembled with Newbler 2.6 with default parameters. Illumina MiSeq data were additionally assembled using Velvet and CLC Assembly Cell (both de Bruijn graph assemblers). Velvet was run using a k-mer value of 55 and `exp_cov` and `cov_cutoff` set to `auto`. The program was run again with `-scaffolding off` to generate a separate assembly without scaffolds. CLC Assembly Cell version 4.0.6 beta was run with default parameters. *De novo* assemblies were compared for chromosomal coverage, broken genes, etc. using Mauve (mauve_snapshot_2011-08-19) and the Mauve Assembly Metrics package [20]. Assemblies were manually examined using the Tablet viewer [21].

Read mapping

For substitution and indel detection, reads from each platform were aligned to the reference assembly using the bwasw module of BWA (version 0.5.9rc1) [22]. The reference genome was indexed with `bwa index -a is`. Bwasw was run with default parameters (gap open penalty 5, gap extension penalty 2) using FASTQ files as input. Output BAM files were post-processed using the `calmd` module of samtools which adds MD tags to each alignment. The MD tag describes the positions of base substitutions. Reads which align to masked bases in the reference genome were excluded from analysis. Read group information was added to the output BAM files using Picard (http://picard.sourceforge.net/). Read accuracy was determined a custom Python script (`calculate_accuracy.py`, available in the Github repository) which utilises the pysam module (http://code.google.com/p/pysam/) to read the BAM alignment. The calculate_accuracy script counts mismatches using the method of Ewing and Green [23] which counts mismatches resulting from substitutions, insertions and deletions. In the case of deletions, mismatches are assigned to one of the adjacent reads in the alignment at random. Depth of coverage reports were generated using `DepthOfCoverage` module of GATK [24]. Reads were additionally mapped against *E. coli* strain c236-11 (PacBio and Illumina sequenced) and *E. coli* strain 55989 (Sanger sequenced) [12][25].

For generation of homopolymer accuracy plots reads for each of the bench-top sequencing platforms were mapped to the reference assembly using Novoalign (version V2.07.13, Novocraft, Malaysia, registered version). Gap penalties were adjusted with parameters as recommended by the documentation `-g 20 -x 5`. Novoalign was set to align its maximum supported read length of 300 using `-n 300`. Homopolymeric tract statistics were enabled using the `-hpstats` option. Quality score recalibration was enabled using the -K option. Only reads that aligned without indels and with a mapping quality of greater than 60 were included in quality score recalibration.

References

1. Pallen, M. J., Nelson, K. & Preston, G. M. *Bacterial Pathogenomics* (ASM Press, 2007).
2. Metzker, M. Sequencing technologies - the next generation. *Nat Rev Genet* **11,** 31–46 (2010).
3. Glenn, T. Field guide to next-generation DNA sequencers. *Mol Ecol Resour* **11,** 759–69 (2011).
4. Pallen, M., Loman, N. & Penn, C. High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. *Curr Opin Microbiol* **13,** 625–31 (2010).
5. Rothberg, J. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475,** 348–52 (2011).
6. Bentley, D. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–9 (2008).
7. Frank, C. *et al.* Epidemic profile of Shiga-toxin-producing Escherichia coli O104:H4 outbreak in Germany. *N Engl J Med* **365,** 1771–80 (2011).
8. Brzuszkiewicz, E. *et al.* Genome sequence analyses of two isolates from the recent Escherichia coli outbreak in Germany reveal the emergence of a new pathotype: Entero-Aggregative-Haemorrhagic Escherichia coli (EAHEC). *Arch Microbiol* **193,** 883–91 (2011).
9. Mellmann, A. *et al.* Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. *PLoS One* **6,** e22751 (2011).
10. Rohde, H. *et al.* Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4. *N Engl J Med* **365,** 718–24 (2011).
11. Rasko, D. *et al.* Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany. *N Engl J Med* **365,** 709–17 (2011).
12. Grad, Y. *et al.* Genomic epidemiology of the Escherichia coli O104:H4 outbreaks in Europe, 2011. *Proc Natl Acad Sci U S A* **109,** 3065–70 (2012).
13. Chattaway, M., Dallman, T., Okeke, I. & Wain, J. Enteroaggregative E. coli O104 from an outbreak of HUS in Germany 2011, could it happen again? *J Infect Dev Ctries* **5,** 425–36 (2011).
14. Kingsford, C., Schatz, M. & Pop, M. Assembly complexity of prokaryotic genomes using short reads. *BMC Bioinformatics* **11,** 21 (2010).
15. Buffalo, V. *qrqc: Quick Read Quality Control* R package version 1.9.1 <http://bioinformatics.ucdavis.edu> (2012).
16. Chaudhuri, R. *et al.* xBASE2: a comprehensive resource for comparative bacterial genomics. *Nucleic Acids Res* **36** (2008).
17. Delcher, A., Bratke, K., Powers, E. & Salzberg, S. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23,** 673–9 (2007).
18. Lowe, T. & Eddy, S. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25,** 955–64 (1997).

19. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* **35,** 3100–8 (2007).

20. Darling, A., Tritt, A., Eisen, J. & Facciotti, M. Mauve assembly metrics. *Bioinformatics* **27,** 2756–7 (2011).

21. Milne, I. *et al.* Tablet–next generation sequence assembly visualization. *Bioinformatics* **26,** 401–2 (2010).

22. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–95 (2010).

23. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8,** 186–94 (1998).

24. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20,** 1297–303 (2010).

25. Touchon, M. *et al.* Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS Genet* **5,** e1000344 (2009).

Supplement

Supplementary Tables

Table 1: Full-length identical matches of clinically important proteins against draft assemblies. Protein coding sequences were searched against draft assemblies for each bench-top instrument using translated BLAST (tblastn, part of the BLAST 2.2.22 package). The results show the number of matches which are identical to the sequence in the reference assembly. For MLST sequences the nucleotide sequences and nucleotide BLAST (blastn) was used.

	454 Junior			Ion Torrent			MiSeq	
	(1)	(2)	(1+2)	(1)	(2)	(1+2)	contigs	scaffolds
Adhesins	1/4	2/4	3/4	1/4	1/4	1/4	2/4	2/4
Antibiotic resistance genes	5/6	6/6	6/6	6/6	5/6	6/6	6/6	6/6
Microcins	0/1	0/1	0/1	0/1	0/1	1/1	1/1	1/1
O104 serotype antigens	0/2	0/2	0/2	2/2	0/2	2/2	2/2	2/2
Serine protease autotransporters of Enterobacteriaceae (SPATEs)	2/4	3/4	3/4	1/4	3/4	3/4	3/4	3/4
Shiga toxin subunits	2/2	2/2	2/2	2/2	2/2	1/2	2/2	2/2
Tellerium resistance	12/12	10/12	12/12	8/12	8/12	8/12	12/12	12/12
MLST housekeeping genes	7/7	6/7	7/7	6/7	6/7	5/7	7/7	7/7
Totals	29/38	29/38	33/38	26/38	25/38	27/38	35/38	35/38

Table 2: Indel summary for benchtop reads against *E. coli* 280 - 454 + Illumina reference

	insertions	deletions	indels_per_100	indels_per_read	total_reads
454 GS Junior (1)	176502	62427	0.38	1.75	136876
454 GS Junior (1+2)	354946	121344	0.38	1.74	275437
454 GS Junior (2)	178444	58917	0.38	1.72	138561
Illumina MiSeq (280)	657	1828	0.00	0.00	1769608
Ion Torrent (1)	1905499	1886049	1.45	1.68	2484481
Ion Torrent (1+2)	3535011	3687275	1.50	1.72	4639731
Ion Torrent (2)	1629512	1801226	1.56	1.77	2155250

Table 3: Indel summary for benchtop reads against *E. coli* 55989 (Sanger sequenced) reference

	insertions	deletions	indels_per_100	indels_per_read	total_reads
454 GS Junior (1)	162292	56866	0.37	1.71	137457
454 GS Junior (1+2)	327199	110863	0.37	1.70	276545
454 GS Junior (2)	164907	53997	0.37	1.69	139088
Illumina MiSeq (280)	2598	5128	0.01	0.01	1772571
Ion Torrent (1)	1720816	1699829	1.44	1.66	2485113
Ion Torrent (1+2)	3195610	3326520	1.49	1.71	4640859
Ion Torrent (2)	1474794	1626691	1.55	1.76	2155746

Table 4: Indel summary for benchtop reads against *E. coli* c236-11 - Illumina + PacBio reference

	insertions	deletions	indels_per_100	indels_per_read	total_reads
454 GS Junior (1)	175391	62434	0.38	1.75	137059
454 GS Junior (1+2)	352620	121470	0.38	1.73	275821
454 GS Junior (2)	177229	59036	0.38	1.72	138762
Illumina MiSeq (280)	909	5338	0.00	0.00	1771145
Ion Torrent (1)	1898486	1879639	1.45	1.68	2484567
Ion Torrent (1+2)	3522207	3674905	1.50	1.72	4639926
Ion Torrent (2)	1623721	1795266	1.56	1.77	2155359

Supplementary Excel File 1. `Assembly_comparison_supplemental.xlsx`

Supplementary Excel File 2. `Assembly_summary_supplemental.xlsx`

Supplementary Excel File 3. `BLAST_searches.xlsx`

Supplementary Figures

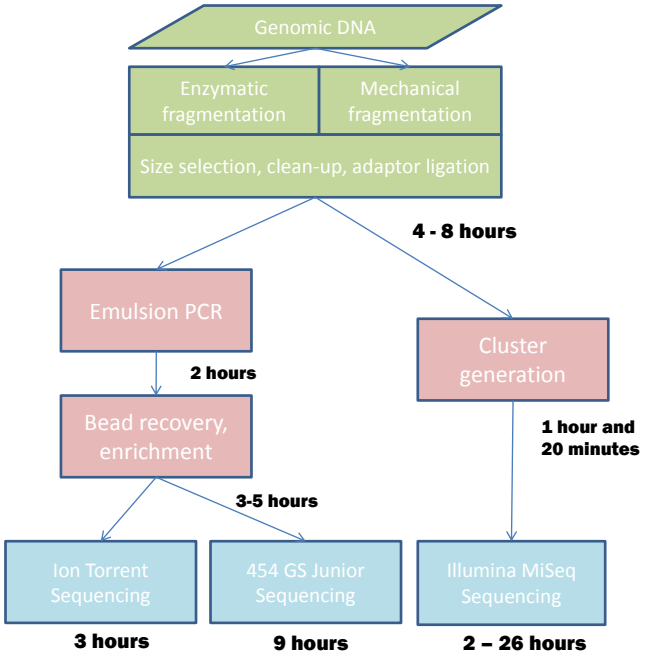


Figure 1: Simplified workflow for bench-top sequencing

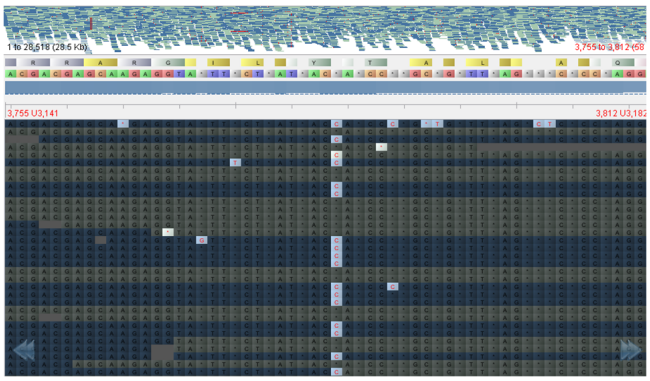


Figure 2: Homopolymeric tract error demonstrating strand bias (light blue is forward strand, dark blue is reverse strand)

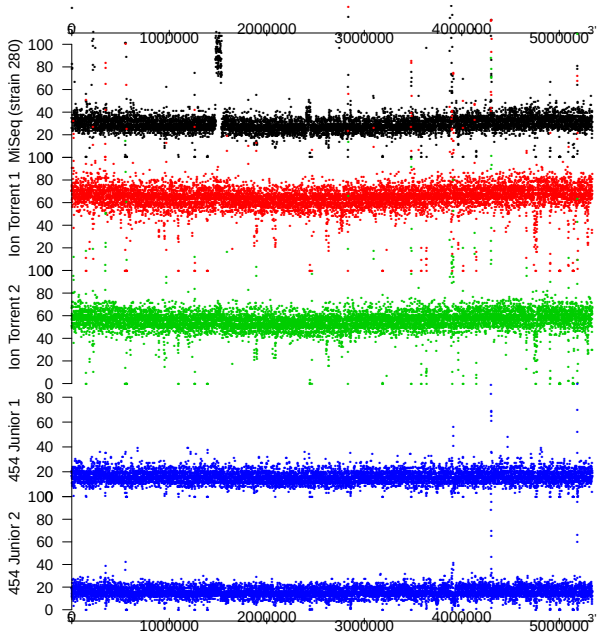


Figure 3: Depth of coverage plot for reads from each benchtop instrument against the *E. coli* strain 280 reference chromosome. In the MiSeq plot the large peak at 1.5 megabases corresponds to the Shiga-toxin producing phage, indicating the phage was likely undergoing lysis when DNA was being prepared. A smaller peak can be seen at the same position in the Ion Torrent PGM data.)

Platform	Reference	Assembler	Number of contigs	Assembly size	Largest contig	Assembly N50	Locally colinear blocks	Total gaps reference	Total gaps assembly	Total gaps	% genome not covered	Assembly gap	Homopolymer gap
454 GS Junior (1)	280	MIRA	188	5122871	195066	42287	38	1177	480	1657	7.5437	926	728
454 GS Junior (1)	280	Newbler	362	5274907	214937	50603	8	364	477	841	5.8339	538	297
454 GS Junior (1+2)	280	MIRA	98	5364151	340118	123481	30	345	292	637	3.7242	375	252
454 GS Junior (1+2)	280	Newbler	216	5326447	385415	119080	3	170	273	443	4.374	271	169
454 GS Junior (2)	280	MIRA	150	5124508	193239	51086	28	811	405	1216	7.2748	604	600
454 GS Junior (2)	280	Newbler	415	5276110	230383	39823	8	457	595	1052	5.4645	628	421
Ion Torrent PGM (1)	280	MIRA	366	5352778	144998	48188	95	459	1785	2244	4.519	725	1513
Ion Torrent PGM (1)	280	Newbler	553	5256982	234450	58782	10	511	1915	2426	5.8175	742	1681
Ion Torrent PGM (1+2	280	MIRA	385	5379335	225465	60073	69	345	1472	1817	4.5922	593	1218
Ion Torrent PGM (1+2	280	Newbler	577	5254254	155086	55431	10	451	1568	2019	5.865	707	1310
Ion Torrent PGM (2)	280	MIRA	376	5342081	196317	41831	77	436	2501	2937	4.7057	752	2178
Ion Torrent PGM (2)	280	Newbler	589	5254958	224797	51211	16	498	1927	2425	5.871	777	1642
MiSeq (contigs)	280	CLC	311	5292732	227129	82564	15	272	259	531	5.101	497	30
MiSeq (contigs)	280	MIRA	214	5353451	341174	81730	53	203	201	404	3.947	371	25
MiSeq (contigs)	280	Velvet	612	5333187	170725	46119	19	358	355	713	4.9435	664	46
MiSeq (scaffolds)	280	CLC	200	5298061	288834	100763	13	267	214	481	4.7433	437	39
MiSeq (scaffolds)	280	Velvet	505	5339506	289526	111638	20	411	398	809	5.0727	765	39
454 Junior (1)	55989	MIRA	188	5122871	195066	42287	63	1208	580	1788	8.7287	not calc	not calc
454 Junior (1)	55989	Newbler	362	5274907	214937	50603	38	425	529	954	7.5219	not calc	not calc
454 Junior (1+2)	55989	MIRA	98	5364151	340118	123481	45	420	369	789	6.5567	not calc	not calc
454 Junior (1+2)	55989	Newbler	216	5326447	385415	119080	38	296	383	679	7.1745	not calc	not calc
454 Junior (2)	55989	MIRA	150	5124508	193239	51086	58	895	523	1418	8.1431	not calc	not calc
454 Junior (2)	55989	Newbler	415	5276110	230383	39823	37	484	611	1095	7.5227	not calc	not calc
Ion Torrent (1)	55989	MIRA	366	5352778	144998	48188	59	511	1661	2172	7.2288	not calc	not calc
Ion Torrent (1)	55989	Newbler	553	5256982	234450	58782	37	489	1758	2247	7.9091	not calc	not calc
Ion Torrent (1+2)	55989	MIRA	385	5379335	225465	60073	52	390	1337	1727	7.2274	not calc	not calc
Ion Torrent (1+2)	55989	Newbler	577	5254254	155086	55431	33	474	1489	1963	7.8686	not calc	not calc
Ion Torrent (2)	55989	MIRA	376	5342081	196317	41831	46	471	2268	2739	7.3526	not calc	not calc
Ion Torrent (2)	55989	Newbler	589	5254958	224797	51211	34	499	1795	2294	7.8792	not calc	not calc
MiSeq (contigs)	55989	CLC	311	5292732	227129	82564	43	356	340	696	7.558	not calc	not calc
MiSeq (contigs)	55989	MIRA	214	5353451	341174	81730	52	324	331	655	7.0667	not calc	not calc
MiSeq (contigs)	55989	Velvet	612	5333187	170725	46119	47	372	379	751	7.5342	not calc	not calc
MiSeq (scaffolds)	55989	CLC	200	5298061	288834	100763	44	396	349	745	7.2292	not calc	not calc
MiSeq (scaffolds)	55989	Velvet	505	5339506	289526	111638	50	398	397	795	7.6076	not calc	not calc

454 Junior (1)	c236-11	MIRA	188	5122871	195066	42287	46	1179	553	1732	6.9523	not calc	not calc
454 Junior (1)	c236-11	Newbler	362	5274907	214937	50603	4	370	517	887	4.4932	not calc	not calc
454 Junior (1+2)	c236-11	MIRA	98	5364151	340118	123481	22	347	389	736	2.8244	not calc	not calc
454 Junior (1+2)	c236-11	Newbler	216	5326447	385415	119080	1	196	363	559	2.969	not calc	not calc
454 Junior (2)	c236-11	MIRA	150	5124508	193239	51086	26	820	465	1285	7.1697	not calc	not calc
454 Junior (2)	c236-11	Newbler	415	5276110	230383	39823	5	462	652	1114	4.328	not calc	not calc
Ion Torrent (1)	c236-11	MIRA	366	5352778	144998	48188	54	394	1771	2165	4.1383	not calc	not calc
Ion Torrent (1)	c236-11	Newbler	553	5256982	234450	58782	3	419	1850	2269	4.8337	not calc	not calc
Ion Torrent (1+2)	c236-11	MIRA	385	5379335	225465	60073	57	336	1508	1844	3.985	not calc	not calc
Ion Torrent (1+2)	c236-11	Newbler	577	5254254	155086	55431	1	363	1508	1871	5.0554	not calc	not calc
Ion Torrent (2)	c236-11	MIRA	376	5342081	196317	41831	69	416	2519	2935	3.926	not calc	not calc
Ion Torrent (2)	c236-11	Newbler	589	5254958	224797	51211	1	429	1866	2295	4.8281	not calc	not calc
MiSeq (contigs)	c236-11	CLC	311	5292732	227129	82564	9	270	303	573	3.9489	not calc	not calc
MiSeq (contigs)	c236-11	MIRA	214	5353451	341174	81730	32	203	287	490	3.4491	not calc	not calc
MiSeq (contigs)	c236-11	Velvet	612	5333187	170725	46119	7	302	337	639	3.9768	not calc	not calc
MiSeq (scaffolds)	c236-11	CLC	200	5298061	288834	100763	9	273	267	540	3.5143	not calc	not calc
MiSeq (scaffolds)	c236-11	Velvet	505	5339506	289526	111638	15	355	390	745	3.9977	not calc	not calc

Supplementary Table 5

Category	Platform	Reference	Gene	Length	Mis-matches	Indels	identity	agree with 280?
st678.fa	280 Reference	Reference	ADK6	536	0	0	100.00%	agreement
st678.fa	454 Junior (1)	Newbler	ADK6	536	0	0	100.00%	agreement
st678.fa	454 Junior (1)	MIRA	ADK6	376	0	0	70.15%	disagreement
st678.fa	454 Junior (1+2)	Newbler	ADK6	536	0	0	100.00%	agreement
st678.fa	454 Junior (1+2)	MIRA	ADK6	536	0	0	100.00%	agreement
st678.fa	454 Junior (2)	Newbler	ADK6	377	0	0	70.34%	disagreement
st678.fa	454 Junior (2)	MIRA	ADK6	536	0	0	100.00%	agreement
st678.fa	C236-11	Reference	ADK6	536	0	0	100.00%	agreement
st678.fa	Ion Torrent (1)	Newbler	ADK6	536	0	0	100.00%	agreement
st678.fa	Ion Torrent (1)	MIRA	ADK6	536	0	0	100.00%	agreement
st678.fa	Ion Torrent (1+2)	Newbler	ADK6	536	0	0	100.00%	agreement
st678.fa	Ion Torrent (1+2)	MIRA	ADK6	536	0	0	100.00%	agreement
st678.fa	Ion Torrent (2)	Newbler	ADK6	536	0	0	100.00%	agreement
st678.fa	Ion Torrent (2)	MIRA	ADK6	536	0	0	100.00%	agreement
st678.fa	MiSeq (contigs)	MIRA	ADK6	536	0	0	100.00%	agreement
st678.fa	MiSeq (contigs)	Velvet	ADK6	536	0	0	100.00%	agreement
st678.fa	MiSeq (contigs)	CLC	ADK6	536	0	0	100.00%	agreement
st678.fa	MiSeq (scaffolds)	Velvet	ADK6	536	0	0	100.00%	agreement
st678.fa	MiSeq (scaffolds)	CLC	ADK6	536	0	0	100.00%	agreement
st678.fa	280 Reference	Reference	FUMC6	469	0	0	100.00%	agreement
st678.fa	454 Junior (1)	Newbler	FUMC6	469	0	0	100.00%	agreement
st678.fa	454 Junior (1)	MIRA	FUMC6	469	0	0	100.00%	agreement
st678.fa	454 Junior (1+2)	Newbler	FUMC6	469	0	0	100.00%	agreement
st678.fa	454 Junior (1+2)	MIRA	FUMC6	469	0	0	100.00%	agreement
st678.fa	454 Junior (2)	Newbler	FUMC6	469	0	0	100.00%	agreement
st678.fa	454 Junior (2)	MIRA	FUMC6	469	0	0	100.00%	agreement
st678.fa	C236-11	Reference	FUMC6	469	0	0	100.00%	agreement
st678.fa	Ion Torrent (1)	Newbler	FUMC6	469	0	0	100.00%	agreement
st678.fa	Ion Torrent (1)	MIRA	FUMC6	469	0	0	100.00%	agreement
st678.fa	Ion Torrent (1+2)	Newbler	FUMC6	469	0	0	100.00%	agreement
st678.fa	Ion Torrent (1+2)	MIRA	FUMC6	469	0	0	100.00%	agreement
st678.fa	Ion Torrent (2)	Newbler	FUMC6	469	0	0	100.00%	agreement
st678.fa	Ion Torrent (2)	MIRA	FUMC6	469	0	0	100.00%	agreement
st678.fa	MiSeq (contigs)	MIRA	FUMC6	469	0	0	100.00%	agreement
st678.fa	MiSeq (contigs)	Velvet	FUMC6	469	0	0	100.00%	agreement
st678.fa	MiSeq (contigs)	CLC	FUMC6	469	0	0	100.00%	agreement
st678.fa	MiSeq (scaffolds)	Velvet	FUMC6	469	0	0	100.00%	agreement
st678.fa	MiSeq (scaffolds)	CLC	FUMC6	469	0	0	100.00%	agreement
st678.fa	280 Reference	Reference	GYRB5	460	0	0	100.00%	agreement
st678.fa	454 Junior (1)	Newbler	GYRB5	460	0	0	100.00%	agreement
st678.fa	454 Junior (1)	MIRA	GYRB5	460	0	0	100.00%	agreement
st678.fa	454 Junior (1+2)	Newbler	GYRB5	460	0	0	100.00%	agreement
st678.fa	454 Junior (1+2)	MIRA	GYRB5	461	1	1	100.00%	disagreement
st678.fa	454 Junior (2)	Newbler	GYRB5	460	0	0	100.00%	agreement
st678.fa	454 Junior (2)	MIRA	GYRB5	460	0	0	100.00%	agreement
st678.fa	C236-11	Reference	GYRB5	460	0	0	100.00%	agreement
st678.fa	Ion Torrent (1)	Newbler	GYRB5	460	0	0	100.00%	agreement
st678.fa	Ion Torrent (1)	MIRA	GYRB5	460	0	0	100.00%	agreement
st678.fa	Ion Torrent (1+2)	Newbler	GYRB5	460	0	0	100.00%	agreement
st678.fa	Ion Torrent (1+2)	MIRA	GYRB5	460	0	0	100.00%	agreement
st678.fa	Ion Torrent (2)	Newbler	GYRB5	460	0	0	100.00%	agreement
st678.fa	Ion Torrent (2)	MIRA	GYRB5	460	0	0	100.00%	agreement
st678.fa	MiSeq (contigs)	MIRA	GYRB5	460	0	0	100.00%	agreement
st678.fa	MiSeq (contigs)	Velvet	GYRB5	460	0	0	100.00%	agreement
st678.fa	MiSeq (contigs)	CLC	GYRB5	460	0	0	100.00%	agreement
st678.fa	MiSeq (scaffolds)	Velvet	GYRB5	460	0	0	100.00%	agreement
st678.fa	MiSeq (scaffolds)	CLC	GYRB5	460	0	0	100.00%	agreement
st678.fa	280 Reference	Reference	ICD136	518	0	0	100.00%	agreement
st678.fa	454 Junior (1)	Newbler	ICD136	518	0	0	100.00%	agreement

st678.fa	454 Junior (1)	MIRA	ICD136	518	0	0	100.00%	agreement
st678.fa	454 Junior (1+2)	Newbler	ICD136	518	0	0	100.00%	agreement
st678.fa	454 Junior (1+2)	MIRA	ICD136	518	0	0	100.00%	agreement
st678.fa	454 Junior (2)	Newbler	ICD136	518	0	0	100.00%	agreement
st678.fa	454 Junior (2)	MIRA	ICD136	31	3	0	5.41%	disagreement
st678.fa	C236-11	Reference	ICD136	518	0	0	100.00%	agreement
st678.fa	Ion Torrent (1)	Newbler	ICD136	518	1	1	99.81%	disagreement
st678.fa	Ion Torrent (1)	MIRA	ICD136	518	0	0	100.00%	agreement
st678.fa	Ion Torrent (1+2)	Newbler	ICD136	518	1	1	99.81%	disagreement
st678.fa	Ion Torrent (1+2)	MIRA	ICD136	518	0	0	100.00%	agreement
st678.fa	Ion Torrent (2)	Newbler	ICD136	518	1	1	99.81%	disagreement
st678.fa	Ion Torrent (2)	MIRA	ICD136	518	0	0	100.00%	agreement
st678.fa	MiSeq (contigs)	MIRA	ICD136	518	0	0	100.00%	agreement
st678.fa	MiSeq (contigs)	Velvet	ICD136	518	0	0	100.00%	agreement
st678.fa	MiSeq (contigs)	CLC	ICD136	518	0	0	100.00%	agreement
st678.fa	MiSeq (scaffolds)	Velvet	ICD136	518	0	0	100.00%	agreement
st678.fa	MiSeq (scaffolds)	CLC	ICD136	518	0	0	100.00%	agreement
st678.fa	280 Reference	Reference	MDH9	452	0	0	100.00%	agreement
st678.fa	454 Junior (1)	Newbler	MDH9	452	0	0	100.00%	agreement
st678.fa	454 Junior (1)	MIRA	MDH9	452	0	0	100.00%	agreement
st678.fa	454 Junior (1+2)	Newbler	MDH9	452	0	0	100.00%	agreement
st678.fa	454 Junior (1+2)	MIRA	MDH9	452	0	0	100.00%	agreement
st678.fa	454 Junior (2)	Newbler	MDH9	452	0	0	100.00%	agreement
st678.fa	454 Junior (2)	MIRA	MDH9	452	0	0	100.00%	agreement
st678.fa	C236-11	Reference	MDH9	452	0	0	100.00%	agreement
st678.fa	Ion Torrent (1)	Newbler	MDH9	452	0	0	100.00%	agreement
st678.fa	Ion Torrent (1)	MIRA	MDH9	452	0	0	100.00%	agreement
st678.fa	Ion Torrent (1+2)	Newbler	MDH9	452	0	0	100.00%	agreement
st678.fa	Ion Torrent (1+2)	MIRA	MDH9	452	0	0	100.00%	agreement
st678.fa	Ion Torrent (2)	Newbler	MDH9	452	0	0	100.00%	agreement
st678.fa	Ion Torrent (2)	MIRA	MDH9	452	0	0	100.00%	agreement
st678.fa	MiSeq (contigs)	MIRA	MDH9	452	0	0	100.00%	agreement
st678.fa	MiSeq (contigs)	Velvet	MDH9	452	0	0	100.00%	agreement
st678.fa	MiSeq (contigs)	CLC	MDH9	452	0	0	100.00%	agreement
st678.fa	MiSeq (scaffolds)	Velvet	MDH9	452	0	0	100.00%	agreement
st678.fa	MiSeq (scaffolds)	CLC	MDH9	452	0	0	100.00%	agreement
st678.fa	280 Reference	Reference	PURA7	478	0	0	100.00%	agreement
st678.fa	454 Junior (1)	Newbler	PURA7	478	0	0	100.00%	agreement
st678.fa	454 Junior (1)	MIRA	PURA7	478	0	0	100.00%	agreement
st678.fa	454 Junior (1+2)	Newbler	PURA7	478	0	0	100.00%	agreement
st678.fa	454 Junior (1+2)	MIRA	PURA7	478	0	0	100.00%	agreement
st678.fa	454 Junior (2)	Newbler	PURA7	478	0	0	100.00%	agreement
st678.fa	454 Junior (2)	MIRA	PURA7	478	0	0	100.00%	agreement
st678.fa	C236-11	Reference	PURA7	478	0	0	100.00%	agreement
st678.fa	Ion Torrent (1)	Newbler	PURA7	478	0	0	100.00%	agreement
st678.fa	Ion Torrent (1)	MIRA	PURA7	478	0	0	100.00%	agreement
st678.fa	Ion Torrent (1+2)	Newbler	PURA7	478	0	0	100.00%	agreement
st678.fa	Ion Torrent (1+2)	MIRA	PURA7	478	0	0	100.00%	agreement
st678.fa	Ion Torrent (2)	Newbler	PURA7	478	0	0	100.00%	agreement
st678.fa	Ion Torrent (2)	MIRA	PURA7	478	0	0	100.00%	agreement
st678.fa	MiSeq (contigs)	MIRA	PURA7	478	0	0	100.00%	agreement
st678.fa	MiSeq (contigs)	Velvet	PURA7	478	0	0	100.00%	agreement
st678.fa	MiSeq (contigs)	CLC	PURA7	478	0	0	100.00%	agreement
st678.fa	MiSeq (scaffolds)	Velvet	PURA7	478	0	0	100.00%	agreement
st678.fa	MiSeq (scaffolds)	CLC	PURA7	478	0	0	100.00%	agreement
st678.fa	280 Reference	Reference	RECA7	510	0	0	100.00%	agreement
st678.fa	454 Junior (1)	Newbler	RECA7	510	0	0	100.00%	agreement
st678.fa	454 Junior (1)	MIRA	RECA7	510	0	0	100.00%	agreement
st678.fa	454 Junior (1+2)	Newbler	RECA7	510	0	0	100.00%	agreement
st678.fa	454 Junior (1+2)	MIRA	RECA7	510	0	0	100.00%	agreement
st678.fa	454 Junior (2)	Newbler	RECA7	510	0	0	100.00%	agreement
st678.fa	454 Junior (2)	MIRA	RECA7	510	0	0	100.00%	agreement
st678.fa	C236-11	Reference	RECA7	510	0	0	100.00%	agreement

st678.fa	Ion Torrent (1)	Newbler	RECA7	510	0	0	100.00%	agreement
st678.fa	Ion Torrent (1)	MIRA	RECA7	510	0	0	100.00%	agreement
st678.fa	Ion Torrent (1+2)	Newbler	RECA7	510	1	1	99.80%	disagreement
st678.fa	Ion Torrent (1+2)	MIRA	RECA7	510	0	0	100.00%	agreement
st678.fa	Ion Torrent (2)	Newbler	RECA7	510	0	0	100.00%	agreement
st678.fa	Ion Torrent (2)	MIRA	RECA7	510	1	1	99.80%	disagreement
st678.fa	MiSeq (contigs)	MIRA	RECA7	510	0	0	100.00%	agreement
st678.fa	MiSeq (contigs)	Velvet	RECA7	510	0	0	100.00%	agreement
st678.fa	MiSeq (contigs)	CLC	RECA7	510	0	0	100.00%	agreement
st678.fa	MiSeq (scaffolds)	Velvet	RECA7	510	0	0	100.00%	agreement
st678.fa	MiSeq (scaffolds)	CLC	RECA7	510	0	0	100.00%	agreement
spates.fa	280 Reference	Reference	gi|218697865|ref|YP_	1372	2	0	99.85%	agreement
spates.fa	454 Junior (1)	Newbler	gi|218697865|ref|YP_	1372	2	0	99.85%	agreement
spates.fa	454 Junior (1)	MIRA	gi|218697865|ref|YP_	1372	2	0	99.85%	agreement
spates.fa	454 Junior (1+2)	Newbler	gi|218697865|ref|YP_	1372	2	0	99.85%	agreement
spates.fa	454 Junior (1+2)	MIRA	gi|218697865|ref|YP_	953	2	0	69.31%	disagreement
spates.fa	454 Junior (2)	Newbler	gi|218697865|ref|YP_	1372	2	0	99.85%	agreement
spates.fa	454 Junior (2)	MIRA	gi|218697865|ref|YP_	987	3	0	71.72%	disagreement
spates.fa	C236-11	Reference	gi|218697865|ref|YP_	1372	2	0	99.85%	agreement
spates.fa	Ion Torrent (1)	Newbler	gi|218697865|ref|YP_	1285	2	0	93.51%	disagreement
spates.fa	Ion Torrent (1)	MIRA	gi|218697865|ref|YP_	1363	5	0	98.98%	disagreement
spates.fa	Ion Torrent (1+2)	Newbler	gi|218697865|ref|YP_	1372	2	0	99.85%	agreement
spates.fa	Ion Torrent (1+2)	MIRA	gi|218697865|ref|YP_	572	0	0	41.69%	disagreement
spates.fa	Ion Torrent (2)	Newbler	gi|218697865|ref|YP_	1372	2	0	99.85%	agreement
spates.fa	Ion Torrent (2)	MIRA	gi|218697865|ref|YP_	827	370	18	33.31%	disagreement
spates.fa	MiSeq (contigs)	MIRA	gi|218697865|ref|YP_	1387	618	37	56.05%	disagreement
spates.fa	MiSeq (contigs)	Velvet	gi|218697865|ref|YP_	1372	2	0	99.85%	agreement
spates.fa	MiSeq (contigs)	CLC	gi|218697865|ref|YP_	1372	2	0	99.85%	agreement
spates.fa	MiSeq (scaffolds)	Velvet	gi|218697865|ref|YP_	1372	2	0	99.85%	agreement
spates.fa	MiSeq (scaffolds)	CLC	gi|218697865|ref|YP_	1372	2	0	99.85%	agreement
spates.fa	280 Reference	Reference	gi|30064291|ref|NP_8	1285	0	0	100.00%	agreement
spates.fa	454 Junior (1)	Newbler	gi|30064291|ref|NP_8	679	0	0	52.84%	disagreement
spates.fa	454 Junior (1)	MIRA	gi|30064291|ref|NP_8	846	0	0	65.84%	disagreement
spates.fa	454 Junior (1+2)	Newbler	gi|30064291|ref|NP_8	1285	0	0	100.00%	agreement
spates.fa	454 Junior (1+2)	MIRA	gi|30064291|ref|NP_8	846	0	0	65.84%	disagreement
spates.fa	454 Junior (2)	Newbler	gi|30064291|ref|NP_8	1285	0	0	100.00%	agreement
spates.fa	454 Junior (2)	MIRA	gi|30064291|ref|NP_8	679	0	0	52.84%	disagreement
spates.fa	C236-11	Reference	gi|30064291|ref|NP_8	1285	0	0	100.00%	agreement
spates.fa	Ion Torrent (1)	Newbler	gi|30064291|ref|NP_8	1285	0	0	100.00%	agreement
spates.fa	Ion Torrent (1)	MIRA	gi|30064291|ref|NP_8	1285	0	0	100.00%	agreement
spates.fa	Ion Torrent (1+2)	Newbler	gi|30064291|ref|NP_8	1285	0	0	100.00%	agreement
spates.fa	Ion Torrent (1+2)	MIRA	gi|30064291|ref|NP_8	1285	0	0	100.00%	agreement
spates.fa	Ion Torrent (2)	Newbler	gi|30064291|ref|NP_8	1285	0	0	100.00%	agreement
spates.fa	Ion Torrent (2)	MIRA	gi|30064291|ref|NP_8	1285	0	0	100.00%	agreement
spates.fa	MiSeq (contigs)	MIRA	gi|30064291|ref|NP_8	1285	0	0	100.00%	agreement
spates.fa	MiSeq (contigs)	Velvet	gi|30064291|ref|NP_8	1285	0	0	100.00%	agreement
spates.fa	MiSeq (contigs)	CLC	gi|30064291|ref|NP_8	1229	0	0	95.64%	disagreement
spates.fa	MiSeq (scaffolds)	Velvet	gi|30064291|ref|NP_8	1285	0	0	100.00%	agreement
spates.fa	MiSeq (scaffolds)	CLC	gi|30064291|ref|NP_8	1287	2	2	100.00%	disagreement
spates.fa	280 Reference	Reference	gi|331681632|ref|ZP_	1372	50	1	96.43%	agreement
spates.fa	454 Junior (1)	Newbler	gi|331681632|ref|ZP_	1372	50	1	96.43%	agreement
spates.fa	454 Junior (1)	MIRA	gi|331681632|ref|ZP_	1372	50	1	96.43%	agreement
spates.fa	454 Junior (1+2)	Newbler	gi|331681632|ref|ZP_	1372	50	1	96.43%	agreement
spates.fa	454 Junior (1+2)	MIRA	gi|331681632|ref|ZP_	953	41	1	66.52%	disagreement
spates.fa	454 Junior (2)	Newbler	gi|331681632|ref|ZP_	1372	50	1	96.43%	agreement
spates.fa	454 Junior (2)	MIRA	gi|331681632|ref|ZP_	987	42	1	68.93%	disagreement
spates.fa	C236-11	Reference	gi|331681632|ref|ZP_	1372	50	1	96.43%	agreement
spates.fa	Ion Torrent (1)	Newbler	gi|331681632|ref|ZP_	1285	50	1	90.08%	disagreement
spates.fa	Ion Torrent (1)	MIRA	gi|331681632|ref|ZP_	1363	53	1	95.55%	disagreement
spates.fa	Ion Torrent (1+2)	Newbler	gi|331681632|ref|ZP_	1372	50	1	96.43%	agreement
spates.fa	Ion Torrent (1+2)	MIRA	gi|331681632|ref|ZP_	572	13	0	40.77%	disagreement
spates.fa	Ion Torrent (2)	Newbler	gi|331681632|ref|ZP_	1372	50	1	96.43%	agreement
spates.fa	Ion Torrent (2)	MIRA	gi|331681632|ref|ZP_	827	366	18	33.63%	disagreement

spates.fa	MiSeq (contigs)	MIRA	gi|331681632|ref|ZP_	1384	612	32	56.31%	disagreement
spates.fa	MiSeq (contigs)	Velvet	gi|331681632|ref|ZP_	1372	50	1	96.43%	agreement
spates.fa	MiSeq (contigs)	CLC	gi|331681632|ref|ZP_	1372	50	1	96.43%	agreement
spates.fa	MiSeq (scaffolds)	Velvet	gi|331681632|ref|ZP_	1372	50	1	96.43%	agreement
spates.fa	MiSeq (scaffolds)	CLC	gi|331681632|ref|ZP_	1372	50	1	96.43%	agreement
spates.fa	280 Reference	Reference	gi|58045130|gb|AAW	1262	7	0	99.45%	agreement
spates.fa	454 Junior (1)	Newbler	gi|58045130|gb|AAW	798	5	0	62.84%	disagreement
spates.fa	454 Junior (1)	MIRA	gi|58045130|gb|AAW	863	13	4	67.35%	disagreement
spates.fa	454 Junior (1+2)	Newbler	gi|58045130|gb|AAW	1213	6	0	95.64%	disagreement
spates.fa	454 Junior (1+2)	MIRA	gi|58045130|gb|AAW	1100	17	0	85.82%	disagreement
spates.fa	454 Junior (2)	Newbler	gi|58045130|gb|AAW	923	5	0	72.74%	disagreement
spates.fa	454 Junior (2)	MIRA	gi|58045130|gb|AAW	897	445	33	35.82%	disagreement
spates.fa	C236-11	Reference	gi|58045130|gb|AAW	1163	6	0	91.68%	disagreement
spates.fa	Ion Torrent (1)	Newbler	gi|58045130|gb|AAW	1196	551	35	51.11%	disagreement
spates.fa	Ion Torrent (1)	MIRA	gi|58045130|gb|AAW	932	5	0	73.45%	disagreement
spates.fa	Ion Torrent (1+2)	Newbler	gi|58045130|gb|AAW	811	6	0	63.79%	disagreement
spates.fa	Ion Torrent (1+2)	MIRA	gi|58045130|gb|AAW	646	6	1	50.71%	disagreement
spates.fa	Ion Torrent (2)	Newbler	gi|58045130|gb|AAW	791	6	0	62.20%	disagreement
spates.fa	Ion Torrent (2)	MIRA	gi|58045130|gb|AAW	811	5	0	63.87%	disagreement
spates.fa	MiSeq (contigs)	MIRA	gi|58045130|gb|AAW	1263	10	1	99.29%	disagreement
spates.fa	MiSeq (contigs)	Velvet	gi|58045130|gb|AAW	860	5	0	67.75%	disagreement
spates.fa	MiSeq (contigs)	CLC	gi|58045130|gb|AAW	929	5	0	73.22%	disagreement
spates.fa	MiSeq (scaffolds)	Velvet	gi|58045130|gb|AAW	860	5	0	67.75%	disagreement
spates.fa	MiSeq (scaffolds)	CLC	gi|58045130|gb|AAW	929	5	0	73.22%	disagreement
antigens.fa	280 Reference	Reference	gi|14517807|gb|AAK6	370	1	0	99.73%	agreement
antigens.fa	454 Junior (1)	Newbler	gi|14517807|gb|AAK6	76	56	11	5.41%	disagreement
antigens.fa	454 Junior (1)	MIRA	gi|14517807|gb|AAK6	76	56	11	5.41%	disagreement
antigens.fa	454 Junior (1+2)	Newbler	gi|14517807|gb|AAK6	76	56	11	5.41%	disagreement
antigens.fa	454 Junior (1+2)	MIRA	gi|14517807|gb|AAK6	76	56	11	5.41%	disagreement
antigens.fa	454 Junior (2)	Newbler	gi|14517807|gb|AAK6	76	56	11	5.41%	disagreement
antigens.fa	454 Junior (2)	MIRA	gi|14517807|gb|AAK6	76	56	11	5.41%	disagreement
antigens.fa	C236-11	Reference	gi|14517807|gb|AAK6	370	1	0	99.73%	agreement
antigens.fa	Ion Torrent (1)	Newbler	gi|14517807|gb|AAK6	370	1	0	99.73%	agreement
antigens.fa	Ion Torrent (1)	MIRA	gi|14517807|gb|AAK6	323	2	0	86.76%	disagreement
antigens.fa	Ion Torrent (1+2)	Newbler	gi|14517807|gb|AAK6	370	1	0	99.73%	agreement
antigens.fa	Ion Torrent (1+2)	MIRA	gi|14517807|gb|AAK6	323	1	0	87.03%	disagreement
antigens.fa	Ion Torrent (2)	Newbler	gi|14517807|gb|AAK6	323	1	0	87.03%	disagreement
antigens.fa	Ion Torrent (2)	MIRA	gi|14517807|gb|AAK6	315	1	0	84.86%	disagreement
antigens.fa	MiSeq (contigs)	MIRA	gi|14517807|gb|AAK6	370	1	0	99.73%	agreement
antigens.fa	MiSeq (contigs)	Velvet	gi|14517807|gb|AAK6	370	1	0	99.73%	agreement
antigens.fa	MiSeq (contigs)	CLC	gi|14517807|gb|AAK6	370	1	0	99.73%	agreement
antigens.fa	MiSeq (scaffolds)	Velvet	gi|14517807|gb|AAK6	370	1	0	99.73%	agreement
antigens.fa	MiSeq (scaffolds)	CLC	gi|14517807|gb|AAK6	370	1	0	99.73%	agreement
antigens.fa	280 Reference	Reference	gi|14517809|gb|AAK6	431	1	0	99.77%	agreement
antigens.fa	454 Junior (1)	Newbler	gi|14517809|gb|AAK6	342	4	0	78.42%	disagreement
antigens.fa	454 Junior (1)	MIRA	gi|14517809|gb|AAK6	57	41	1	3.71%	disagreement
antigens.fa	454 Junior (1+2)	Newbler	gi|14517809|gb|AAK6	327	1	0	75.64%	disagreement
antigens.fa	454 Junior (1+2)	MIRA	gi|14517809|gb|AAK6	279	7	0	63.11%	disagreement
antigens.fa	454 Junior (2)	Newbler	gi|14517809|gb|AAK6	269	71	2	45.94%	disagreement
antigens.fa	454 Junior (2)	MIRA	gi|14517809|gb|AAK6	21	2	0	4.41%	disagreement
antigens.fa	C236-11	Reference	gi|14517809|gb|AAK6	431	1	0	99.77%	agreement
antigens.fa	Ion Torrent (1)	Newbler	gi|14517809|gb|AAK6	431	1	0	99.77%	agreement
antigens.fa	Ion Torrent (1)	MIRA	gi|14517809|gb|AAK6	309	0	0	71.69%	disagreement
antigens.fa	Ion Torrent (1+2)	Newbler	gi|14517809|gb|AAK6	431	1	0	99.77%	agreement
antigens.fa	Ion Torrent (1+2)	MIRA	gi|14517809|gb|AAK6	431	1	0	99.77%	agreement
antigens.fa	Ion Torrent (2)	Newbler	gi|14517809|gb|AAK6	310	0	0	71.93%	disagreement
antigens.fa	Ion Torrent (2)	MIRA	gi|14517809|gb|AAK6	306	0	0	71.00%	disagreement
antigens.fa	MiSeq (contigs)	MIRA	gi|14517809|gb|AAK6	431	1	0	99.77%	agreement
antigens.fa	MiSeq (contigs)	Velvet	gi|14517809|gb|AAK6	431	1	0	99.77%	agreement
antigens.fa	MiSeq (contigs)	CLC	gi|14517809|gb|AAK6	431	1	0	99.77%	agreement
antigens.fa	MiSeq (scaffolds)	Velvet	gi|14517809|gb|AAK6	431	1	0	99.77%	agreement
antigens.fa	MiSeq (scaffolds)	CLC	gi|14517809|gb|AAK6	431	1	0	99.77%	agreement
stx2.fa	280 Reference	Reference	stx2A	319	0	0	100.00%	agreement

stx2.fa	454 Junior (1)	Newbler	stx2A	319	0	0	100.00%	agreement
stx2.fa	454 Junior (1)	MIRA	stx2A	206	1	0	64.26%	disagreement
stx2.fa	454 Junior (1+2)	Newbler	stx2A	319	0	0	100.00%	agreement
stx2.fa	454 Junior (1+2)	MIRA	stx2A	319	0	0	100.00%	agreement
stx2.fa	454 Junior (2)	Newbler	stx2A	319	0	0	100.00%	agreement
stx2.fa	454 Junior (2)	MIRA	stx2A	292	0	0	91.54%	disagreement
stx2.fa	C236-11	Reference	stx2A	319	0	0	100.00%	agreement
stx2.fa	Ion Torrent (1)	Newbler	stx2A	319	0	0	100.00%	agreement
stx2.fa	Ion Torrent (1)	MIRA	stx2A	319	0	0	100.00%	agreement
stx2.fa	Ion Torrent (1+2)	Newbler	stx2A	198	6	3	60.19%	disagreement
stx2.fa	Ion Torrent (1+2)	MIRA	stx2A	319	0	0	100.00%	agreement
stx2.fa	Ion Torrent (2)	Newbler	stx2A	319	0	0	100.00%	agreement
stx2.fa	Ion Torrent (2)	MIRA	stx2A	198	6	3	60.19%	disagreement
stx2.fa	MiSeq (contigs)	MIRA	stx2A	319	0	0	100.00%	agreement
stx2.fa	MiSeq (contigs)	Velvet	stx2A	319	0	0	100.00%	agreement
stx2.fa	MiSeq (contigs)	CLC	stx2A	319	0	0	100.00%	agreement
stx2.fa	MiSeq (scaffolds)	Velvet	stx2A	319	0	0	100.00%	agreement
stx2.fa	MiSeq (scaffolds)	CLC	stx2A	319	0	0	100.00%	agreement
stx2.fa	280 Reference	Reference	stx2B	89	0	0	100.00%	agreement
stx2.fa	454 Junior (1)	Newbler	stx2B	89	0	0	100.00%	agreement
stx2.fa	454 Junior (1)	MIRA	stx2B	89	0	0	100.00%	agreement
stx2.fa	454 Junior (1+2)	Newbler	stx2B	89	0	0	100.00%	agreement
stx2.fa	454 Junior (1+2)	MIRA	stx2B	89	0	0	100.00%	agreement
stx2.fa	454 Junior (2)	Newbler	stx2B	89	0	0	100.00%	agreement
stx2.fa	454 Junior (2)	MIRA	stx2B	77	0	0	86.52%	disagreement
stx2.fa	C236-11	Reference	stx2B	89	0	0	100.00%	agreement
stx2.fa	Ion Torrent (1)	Newbler	stx2B	89	0	0	100.00%	agreement
stx2.fa	Ion Torrent (1)	MIRA	stx2B	89	0	0	100.00%	agreement
stx2.fa	Ion Torrent (1+2)	Newbler	stx2B	89	0	0	100.00%	agreement
stx2.fa	Ion Torrent (1+2)	MIRA	stx2B	89	0	0	100.00%	agreement
stx2.fa	Ion Torrent (2)	Newbler	stx2B	89	0	0	100.00%	agreement
stx2.fa	Ion Torrent (2)	MIRA	stx2B	89	0	0	100.00%	agreement
stx2.fa	MiSeq (contigs)	MIRA	stx2B	89	0	0	100.00%	agreement
stx2.fa	MiSeq (contigs)	Velvet	stx2B	89	0	0	100.00%	agreement
stx2.fa	MiSeq (contigs)	CLC	stx2B	89	0	0	100.00%	agreement
stx2.fa	MiSeq (scaffolds)	Velvet	stx2B	89	0	0	100.00%	agreement
stx2.fa	MiSeq (scaffolds)	CLC	stx2B	89	0	0	100.00%	agreement
tellerium.fa	280 Reference	Reference	TerA	385	0	0	100.00%	agreement
tellerium.fa	454 Junior (1)	Newbler	TerA	385	0	0	100.00%	agreement
tellerium.fa	454 Junior (1)	MIRA	TerA	385	0	0	100.00%	agreement
tellerium.fa	454 Junior (1+2)	Newbler	TerA	385	0	0	100.00%	agreement
tellerium.fa	454 Junior (1+2)	MIRA	TerA	385	0	0	100.00%	agreement
tellerium.fa	454 Junior (2)	Newbler	TerA	385	0	0	100.00%	agreement
tellerium.fa	454 Junior (2)	MIRA	TerA	385	0	0	100.00%	agreement
tellerium.fa	C236-11	Reference	TerA	385	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1)	Newbler	TerA	287	0	0	74.55%	disagreement
tellerium.fa	Ion Torrent (1)	MIRA	TerA	379	2	0	97.92%	disagreement
tellerium.fa	Ion Torrent (1+2)	Newbler	TerA	287	0	0	74.55%	disagreement
tellerium.fa	Ion Torrent (1+2)	MIRA	TerA	380	2	0	98.18%	disagreement
tellerium.fa	Ion Torrent (2)	Newbler	TerA	385	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (2)	MIRA	TerA	380	2	0	98.18%	disagreement
tellerium.fa	MiSeq (contigs)	MIRA	TerA	385	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	Velvet	TerA	385	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	CLC	TerA	385	0	0	100.00%	agreement
tellerium.fa	MiSeq (scaffolds)	Velvet	TerA	385	0	0	100.00%	agreement
tellerium.fa	MiSeq (scaffolds)	CLC	TerA	385	0	0	100.00%	agreement
tellerium.fa	280 Reference	Reference	TerB	151	0	0	100.00%	agreement
tellerium.fa	454 Junior (1)	Newbler	TerB	151	0	0	100.00%	agreement
tellerium.fa	454 Junior (1)	MIRA	TerB	151	0	0	100.00%	agreement
tellerium.fa	454 Junior (1+2)	Newbler	TerB	151	0	0	100.00%	agreement
tellerium.fa	454 Junior (1+2)	MIRA	TerB	151	0	0	100.00%	agreement
tellerium.fa	454 Junior (2)	Newbler	TerB	151	0	0	100.00%	agreement
tellerium.fa	454 Junior (2)	MIRA	TerB	151	0	0	100.00%	agreement

tellerium.fa	C236-11	Reference	TerB	151	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1)	Newbler	TerB	151	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1)	MIRA	TerB	151	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1+2)	Newbler	TerB	151	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1+2)	MIRA	TerB	151	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (2)	Newbler	TerB	151	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (2)	MIRA	TerB	151	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	MIRA	TerB	151	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	Velvet	TerB	151	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	CLC	TerB	151	0	0	100.00%	agreement
tellerium.fa	MiSeq (scaffolds)	Velvet	TerB	151	0	0	100.00%	agreement
tellerium.fa	MiSeq (scaffolds)	CLC	TerB	151	0	0	100.00%	agreement
tellerium.fa	280 Reference	Reference	TerC	346	0	0	100.00%	agreement
tellerium.fa	454 Junior (1)	Newbler	TerC	346	0	0	100.00%	agreement
tellerium.fa	454 Junior (1)	MIRA	TerC	346	0	0	100.00%	agreement
tellerium.fa	454 Junior (1+2)	Newbler	TerC	346	0	0	100.00%	agreement
tellerium.fa	454 Junior (1+2)	MIRA	TerC	346	0	0	100.00%	agreement
tellerium.fa	454 Junior (2)	Newbler	TerC	346	0	0	100.00%	agreement
tellerium.fa	454 Junior (2)	MIRA	TerC	346	0	0	100.00%	agreement
tellerium.fa	C236-11	Reference	TerC	346	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1)	Newbler	TerC	346	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1)	MIRA	TerC	346	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1+2)	Newbler	TerC	346	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1+2)	MIRA	TerC	346	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (2)	Newbler	TerC	346	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (2)	MIRA	TerC	346	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	MIRA	TerC	346	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	Velvet	TerC	346	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	CLC	TerC	346	0	0	100.00%	agreement
tellerium.fa	MiSeq (scaffolds)	Velvet	TerC	346	0	0	100.00%	agreement
tellerium.fa	MiSeq (scaffolds)	CLC	TerC	346	0	0	100.00%	agreement
tellerium.fa	280 Reference	Reference	TerD	192	0	0	100.00%	agreement
tellerium.fa	454 Junior (1)	Newbler	TerD	192	0	0	100.00%	agreement
tellerium.fa	454 Junior (1)	MIRA	TerD	192	0	0	100.00%	agreement
tellerium.fa	454 Junior (1+2)	Newbler	TerD	192	0	0	100.00%	agreement
tellerium.fa	454 Junior (1+2)	MIRA	TerD	192	0	0	100.00%	agreement
tellerium.fa	454 Junior (2)	Newbler	TerD	192	0	0	100.00%	agreement
tellerium.fa	454 Junior (2)	MIRA	TerD	192	0	0	100.00%	agreement
tellerium.fa	C236-11	Reference	TerD	192	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1)	Newbler	TerD	154	0	0	80.21%	disagreement
tellerium.fa	Ion Torrent (1)	MIRA	TerD	154	0	0	80.21%	disagreement
tellerium.fa	Ion Torrent (1+2)	Newbler	TerD	192	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1+2)	MIRA	TerD	154	0	0	80.21%	disagreement
tellerium.fa	Ion Torrent (2)	Newbler	TerD	154	0	0	80.21%	disagreement
tellerium.fa	Ion Torrent (2)	MIRA	TerD	154	0	0	80.21%	disagreement
tellerium.fa	MiSeq (contigs)	MIRA	TerD	192	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	Velvet	TerD	192	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	CLC	TerD	192	0	0	100.00%	agreement
tellerium.fa	MiSeq (scaffolds)	Velvet	TerD	192	0	0	100.00%	agreement
tellerium.fa	MiSeq (scaffolds)	CLC	TerD	192	0	0	100.00%	agreement
tellerium.fa	280 Reference	Reference	TerE	191	0	0	100.00%	agreement
tellerium.fa	454 Junior (1)	Newbler	TerE	191	0	0	100.00%	agreement
tellerium.fa	454 Junior (1)	MIRA	TerE	191	0	0	100.00%	agreement
tellerium.fa	454 Junior (1+2)	Newbler	TerE	191	0	0	100.00%	agreement
tellerium.fa	454 Junior (1+2)	MIRA	TerE	191	0	0	100.00%	agreement
tellerium.fa	454 Junior (2)	Newbler	TerE	191	0	0	100.00%	agreement
tellerium.fa	454 Junior (2)	MIRA	TerE	150	0	0	78.53%	disagreement
tellerium.fa	C236-11	Reference	TerE	191	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1)	Newbler	TerE	191	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1)	MIRA	TerE	132	7	0	65.45%	disagreement
tellerium.fa	Ion Torrent (1+2)	Newbler	TerE	132	7	0	65.45%	disagreement
tellerium.fa	Ion Torrent (1+2)	MIRA	TerE	191	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (2)	Newbler	TerE	132	7	0	65.45%	disagreement

tellerium.fa	Ion Torrent (2)	MIRA	TerE	191	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	MIRA	TerE	191	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	Velvet	TerE	191	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	CLC	TerE	191	0	0	100.00%	agreement
tellerium.fa	MiSeq (scaffolds)	Velvet	TerE	191	0	0	100.00%	agreement
tellerium.fa	MiSeq (scaffolds)	CLC	TerE	191	0	0	100.00%	agreement
tellerium.fa	280 Reference	Reference	TerF	413	0	0	100.00%	agreement
tellerium.fa	454 Junior (1)	Newbler	TerF	413	0	0	100.00%	agreement
tellerium.fa	454 Junior (1)	MIRA	TerF	210	122	12	21.31%	disagreement
tellerium.fa	454 Junior (1+2)	Newbler	TerF	413	0	0	100.00%	agreement
tellerium.fa	454 Junior (1+2)	MIRA	TerF	413	0	0	100.00%	agreement
tellerium.fa	454 Junior (2)	Newbler	TerF	413	0	0	100.00%	agreement
tellerium.fa	454 Junior (2)	MIRA	TerF	210	122	12	21.31%	disagreement
tellerium.fa	C236-11	Reference	TerF	413	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1)	Newbler	TerF	291	0	0	70.46%	disagreement
tellerium.fa	Ion Torrent (1)	MIRA	TerF	291	0	0	70.46%	disagreement
tellerium.fa	Ion Torrent (1+2)	Newbler	TerF	291	0	0	70.46%	disagreement
tellerium.fa	Ion Torrent (1+2)	MIRA	TerF	291	0	0	70.46%	disagreement
tellerium.fa	Ion Torrent (2)	Newbler	TerF	291	0	0	70.46%	disagreement
tellerium.fa	Ion Torrent (2)	MIRA	TerF	291	0	0	70.46%	disagreement
tellerium.fa	MiSeq (contigs)	MIRA	TerF	413	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	Velvet	TerF	413	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	CLC	TerF	413	0	0	100.00%	agreement
tellerium.fa	MiSeq (scaffolds)	Velvet	TerF	413	0	0	100.00%	agreement
tellerium.fa	MiSeq (scaffolds)	CLC	TerF	413	0	0	100.00%	agreement
tellerium.fa	280 Reference	Reference	TerW	77	0	0	100.00%	agreement
tellerium.fa	454 Junior (1)	Newbler	TerW	77	0	0	100.00%	agreement
tellerium.fa	454 Junior (1)	MIRA	TerW	42	0	0	54.55%	disagreement
tellerium.fa	454 Junior (1+2)	Newbler	TerW	77	0	0	100.00%	agreement
tellerium.fa	454 Junior (1+2)	MIRA	TerW	77	0	0	100.00%	agreement
tellerium.fa	454 Junior (2)	Newbler	TerW	42	0	0	54.55%	disagreement
tellerium.fa	454 Junior (2)	MIRA	TerW	77	0	0	100.00%	agreement
tellerium.fa	C236-11	Reference	TerW	77	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1)	Newbler	TerW	77	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1)	MIRA	TerW	51	0	0	66.23%	disagreement
tellerium.fa	Ion Torrent (1+2)	Newbler	TerW	77	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1+2)	MIRA	TerW	77	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (2)	Newbler	TerW	77	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (2)	MIRA	TerW	77	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	MIRA	TerW	77	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	Velvet	TerW	77	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	CLC	TerW	77	0	0	100.00%	agreement
tellerium.fa	MiSeq (scaffolds)	Velvet	TerW	77	0	0	100.00%	agreement
tellerium.fa	MiSeq (scaffolds)	CLC	TerW	77	0	0	100.00%	agreement
tellerium.fa	280 Reference	Reference	TerX	212	0	0	100.00%	agreement
tellerium.fa	454 Junior (1)	Newbler	TerX	212	0	0	100.00%	agreement
tellerium.fa	454 Junior (1)	MIRA	TerX	212	0	0	100.00%	agreement
tellerium.fa	454 Junior (1+2)	Newbler	TerX	212	0	0	100.00%	agreement
tellerium.fa	454 Junior (1+2)	MIRA	TerX	212	0	0	100.00%	agreement
tellerium.fa	454 Junior (2)	Newbler	TerX	212	0	0	100.00%	agreement
tellerium.fa	454 Junior (2)	MIRA	TerX	212	0	0	100.00%	agreement
tellerium.fa	C236-11	Reference	TerX	212	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1)	Newbler	TerX	212	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1)	MIRA	TerX	203	4	0	93.87%	disagreement
tellerium.fa	Ion Torrent (1+2)	Newbler	TerX	212	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1+2)	MIRA	TerX	212	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (2)	Newbler	TerX	212	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (2)	MIRA	TerX	203	0	0	95.75%	disagreement
tellerium.fa	MiSeq (contigs)	MIRA	TerX	212	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	Velvet	TerX	212	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	CLC	TerX	212	0	0	100.00%	agreement
tellerium.fa	MiSeq (scaffolds)	Velvet	TerX	212	0	0	100.00%	agreement
tellerium.fa	MiSeq (scaffolds)	CLC	TerX	212	0	0	100.00%	agreement

tellerium.fa	280 Reference	Reference	TerY1	197	0	0	100.00%	agreement
tellerium.fa	454 Junior (1)	Newbler	TerY1	197	0	0	100.00%	agreement
tellerium.fa	454 Junior (1)	MIRA	TerY1	197	0	0	100.00%	agreement
tellerium.fa	454 Junior (1+2)	Newbler	TerY1	197	0	0	100.00%	agreement
tellerium.fa	454 Junior (1+2)	MIRA	TerY1	197	0	0	100.00%	agreement
tellerium.fa	454 Junior (2)	Newbler	TerY1	197	0	0	100.00%	agreement
tellerium.fa	454 Junior (2)	MIRA	TerY1	197	0	0	100.00%	agreement
tellerium.fa	C236-11	Reference	TerY1	197	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1)	Newbler	TerY1	197	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1)	MIRA	TerY1	197	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1+2)	Newbler	TerY1	197	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1+2)	MIRA	TerY1	197	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (2)	Newbler	TerY1	197	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (2)	MIRA	TerY1	197	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	MIRA	TerY1	197	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	Velvet	TerY1	197	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	CLC	TerY1	197	0	0	100.00%	agreement
tellerium.fa	MiSeq (scaffolds)	Velvet	TerY1	197	0	0	100.00%	agreement
tellerium.fa	MiSeq (scaffolds)	CLC	TerY1	197	0	0	100.00%	agreement
tellerium.fa	280 Reference	Reference	TerY2	212	0	0	100.00%	agreement
tellerium.fa	454 Junior (1)	Newbler	TerY2	212	0	0	100.00%	agreement
tellerium.fa	454 Junior (1)	MIRA	TerY2	212	0	0	100.00%	agreement
tellerium.fa	454 Junior (1+2)	Newbler	TerY2	212	0	0	100.00%	agreement
tellerium.fa	454 Junior (1+2)	MIRA	TerY2	212	0	0	100.00%	agreement
tellerium.fa	454 Junior (2)	Newbler	TerY2	212	0	0	100.00%	agreement
tellerium.fa	454 Junior (2)	MIRA	TerY2	212	0	0	100.00%	agreement
tellerium.fa	C236-11	Reference	TerY2	212	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1)	Newbler	TerY2	212	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1)	MIRA	TerY2	212	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1+2)	Newbler	TerY2	212	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1+2)	MIRA	TerY2	212	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (2)	Newbler	TerY2	212	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (2)	MIRA	TerY2	212	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	MIRA	TerY2	212	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	Velvet	TerY2	212	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	CLC	TerY2	212	0	0	100.00%	agreement
tellerium.fa	MiSeq (scaffolds)	Velvet	TerY2	212	0	0	100.00%	agreement
tellerium.fa	MiSeq (scaffolds)	CLC	TerY2	212	0	0	100.00%	agreement
tellerium.fa	280 Reference	Reference	TerY3	346	0	0	100.00%	agreement
tellerium.fa	454 Junior (1)	Newbler	TerY3	346	0	0	100.00%	agreement
tellerium.fa	454 Junior (1)	MIRA	TerY3	343	4	0	97.98%	disagreement
tellerium.fa	454 Junior (1+2)	Newbler	TerY3	346	0	0	100.00%	agreement
tellerium.fa	454 Junior (1+2)	MIRA	TerY3	346	0	0	100.00%	agreement
tellerium.fa	454 Junior (2)	Newbler	TerY3	326	1	0	93.93%	disagreement
tellerium.fa	454 Junior (2)	MIRA	TerY3	346	0	0	100.00%	agreement
tellerium.fa	C236-11	Reference	TerY3	346	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1)	Newbler	TerY3	346	1	0	99.71%	disagreement
tellerium.fa	Ion Torrent (1)	MIRA	TerY3	346	1	0	99.71%	disagreement
tellerium.fa	Ion Torrent (1+2)	Newbler	TerY3	346	1	0	99.71%	disagreement
tellerium.fa	Ion Torrent (1+2)	MIRA	TerY3	346	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (2)	Newbler	TerY3	343	1	0	98.84%	disagreement
tellerium.fa	Ion Torrent (2)	MIRA	TerY3	346	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	MIRA	TerY3	346	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	Velvet	TerY3	346	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	CLC	TerY3	346	0	0	100.00%	agreement
tellerium.fa	MiSeq (scaffolds)	Velvet	TerY3	346	0	0	100.00%	agreement
tellerium.fa	MiSeq (scaffolds)	CLC	TerY3	346	0	0	100.00%	agreement
tellerium.fa	280 Reference	Reference	TerZ	193	0	0	100.00%	agreement
tellerium.fa	454 Junior (1)	Newbler	TerZ	193	0	0	100.00%	agreement
tellerium.fa	454 Junior (1)	MIRA	TerZ	193	0	0	100.00%	agreement
tellerium.fa	454 Junior (1+2)	Newbler	TerZ	193	0	0	100.00%	agreement
tellerium.fa	454 Junior (1+2)	MIRA	TerZ	193	0	0	100.00%	agreement
tellerium.fa	454 Junior (2)	Newbler	TerZ	193	0	0	100.00%	agreement

tellerium.fa	454 Junior (2)	MIRA	TerZ	193	0	0	100.00%	agreement
tellerium.fa	C236-11	Reference	TerZ	193	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1)	Newbler	TerZ	193	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1)	MIRA	TerZ	193	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1+2)	Newbler	TerZ	193	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (1+2)	MIRA	TerZ	193	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (2)	Newbler	TerZ	193	0	0	100.00%	agreement
tellerium.fa	Ion Torrent (2)	MIRA	TerZ	193	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	MIRA	TerZ	193	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	Velvet	TerZ	193	0	0	100.00%	agreement
tellerium.fa	MiSeq (contigs)	CLC	TerZ	193	0	0	100.00%	agreement
tellerium.fa	MiSeq (scaffolds)	Velvet	TerZ	193	0	0	100.00%	agreement
tellerium.fa	MiSeq (scaffolds)	CLC	TerZ	193	0	0	100.00%	agreement
antibiotics.fa	280 Reference	Reference	TetA	424	0	0	100.00%	agreement
antibiotics.fa	454 Junior (1)	Newbler	TetA	424	0	0	100.00%	agreement
antibiotics.fa	454 Junior (1)	MIRA	TetA	424	0	0	100.00%	agreement
antibiotics.fa	454 Junior (1+2)	Newbler	TetA	424	0	0	100.00%	agreement
antibiotics.fa	454 Junior (1+2)	MIRA	TetA	424	0	0	100.00%	agreement
antibiotics.fa	454 Junior (2)	Newbler	TetA	424	0	0	100.00%	agreement
antibiotics.fa	454 Junior (2)	MIRA	TetA	424	0	0	100.00%	agreement
antibiotics.fa	C236-11	Reference	TetA	424	0	0	100.00%	agreement
antibiotics.fa	Ion Torrent (1)	Newbler	TetA	424	0	0	100.00%	agreement
antibiotics.fa	Ion Torrent (1)	MIRA	TetA	424	0	0	100.00%	agreement
antibiotics.fa	Ion Torrent (1+2)	Newbler	TetA	424	0	0	100.00%	agreement
antibiotics.fa	Ion Torrent (1+2)	MIRA	TetA	424	0	0	100.00%	agreement
antibiotics.fa	Ion Torrent (2)	Newbler	TetA	259	3	0	60.38%	disagreement
antibiotics.fa	Ion Torrent (2)	MIRA	TetA	261	0	0	61.56%	disagreement
antibiotics.fa	MiSeq (contigs)	MIRA	TetA	424	0	0	100.00%	agreement
antibiotics.fa	MiSeq (contigs)	Velvet	TetA	424	0	0	100.00%	agreement
antibiotics.fa	MiSeq (contigs)	CLC	TetA	424	0	0	100.00%	agreement
antibiotics.fa	MiSeq (scaffolds)	Velvet	TetA	424	0	0	100.00%	agreement
antibiotics.fa	MiSeq (scaffolds)	CLC	TetA	424	0	0	100.00%	agreement
antibiotics.fa	280 Reference	Reference	blaCTX-M-15	280	0	0	100.00%	agreement
antibiotics.fa	454 Junior (1)	Newbler	blaCTX-M-15	262	161	4	36.07%	disagreement
antibiotics.fa	454 Junior (1)	MIRA	blaCTX-M-15	262	161	4	36.07%	disagreement
antibiotics.fa	454 Junior (1+2)	Newbler	blaCTX-M-15	280	0	0	100.00%	agreement
antibiotics.fa	454 Junior (1+2)	MIRA	blaCTX-M-15	280	0	0	100.00%	agreement
antibiotics.fa	454 Junior (2)	Newbler	blaCTX-M-15	280	0	0	100.00%	agreement
antibiotics.fa	454 Junior (2)	MIRA	blaCTX-M-15	87	61	17	9.29%	disagreement
antibiotics.fa	C236-11	Reference	blaCTX-M-15	280	0	0	100.00%	agreement
antibiotics.fa	Ion Torrent (1)	Newbler	blaCTX-M-15	280	0	0	100.00%	agreement
antibiotics.fa	Ion Torrent (1)	MIRA	blaCTX-M-15	280	0	0	100.00%	agreement
antibiotics.fa	Ion Torrent (1+2)	Newbler	blaCTX-M-15	280	0	0	100.00%	agreement
antibiotics.fa	Ion Torrent (1+2)	MIRA	blaCTX-M-15	280	0	0	100.00%	agreement
antibiotics.fa	Ion Torrent (2)	Newbler	blaCTX-M-15	280	0	0	100.00%	agreement
antibiotics.fa	Ion Torrent (2)	MIRA	blaCTX-M-15	280	0	0	100.00%	agreement
antibiotics.fa	MiSeq (contigs)	MIRA	blaCTX-M-15	280	0	0	100.00%	agreement
antibiotics.fa	MiSeq (contigs)	Velvet	blaCTX-M-15	280	0	0	100.00%	agreement
antibiotics.fa	MiSeq (contigs)	CLC	blaCTX-M-15	280	0	0	100.00%	agreement
antibiotics.fa	MiSeq (scaffolds)	Velvet	blaCTX-M-15	280	0	0	100.00%	agreement
antibiotics.fa	MiSeq (scaffolds)	CLC	blaCTX-M-15	280	0	0	100.00%	agreement
antibiotics.fa	280 Reference	Reference	blaT	52	1	0	85.00%	agreement
antibiotics.fa	454 Junior (1)	Newbler	blaT	52	1	0	85.00%	agreement
antibiotics.fa	454 Junior (1)	MIRA	blaT	52	1	0	85.00%	agreement
antibiotics.fa	454 Junior (1+2)	Newbler	blaT	52	1	0	85.00%	agreement
antibiotics.fa	454 Junior (1+2)	MIRA	blaT	52	1	0	85.00%	agreement
antibiotics.fa	454 Junior (2)	Newbler	blaT	52	1	0	85.00%	agreement
antibiotics.fa	454 Junior (2)	MIRA	blaT	39	24	0	25.00%	disagreement
antibiotics.fa	C236-11	Reference	blaT	52	1	0	85.00%	agreement
antibiotics.fa	Ion Torrent (1)	Newbler	blaT	52	1	0	85.00%	agreement
antibiotics.fa	Ion Torrent (1)	MIRA	blaT	52	1	0	85.00%	agreement
antibiotics.fa	Ion Torrent (1+2)	Newbler	blaT	52	1	0	85.00%	agreement
antibiotics.fa	Ion Torrent (1+2)	MIRA	blaT	52	1	0	85.00%	agreement

antibiotics.fa	Ion Torrent (2)	Newbler	blaT	52	1	0	85.00%	agreement
antibiotics.fa	Ion Torrent (2)	MIRA	blaT	52	1	0	85.00%	agreement
antibiotics.fa	MiSeq (contigs)	MIRA	blaT	52	1	0	85.00%	agreement
antibiotics.fa	MiSeq (contigs)	Velvet	blaT	52	1	0	85.00%	agreement
antibiotics.fa	MiSeq (contigs)	CLC	blaT	52	1	0	85.00%	agreement
antibiotics.fa	MiSeq (scaffolds)	Velvet	blaT	52	1	0	85.00%	agreement
antibiotics.fa	MiSeq (scaffolds)	CLC	blaT	52	1	0	85.00%	agreement
antibiotics.fa	280 Reference	Reference	blaTEM	286	0	0	100.00%	agreement
antibiotics.fa	454 Junior (1)	Newbler	blaTEM	286	0	0	100.00%	agreement
antibiotics.fa	454 Junior (1)	MIRA	blaTEM	286	0	0	100.00%	agreement
antibiotics.fa	454 Junior (1+2)	Newbler	blaTEM	286	0	0	100.00%	agreement
antibiotics.fa	454 Junior (1+2)	MIRA	blaTEM	286	0	0	100.00%	agreement
antibiotics.fa	454 Junior (2)	Newbler	blaTEM	286	0	0	100.00%	agreement
antibiotics.fa	454 Junior (2)	MIRA	blaTEM	42	28	2	4.90%	disagreement
antibiotics.fa	C236-11	Reference	blaTEM	286	0	0	100.00%	agreement
antibiotics.fa	Ion Torrent (1)	Newbler	blaTEM	286	0	0	100.00%	agreement
antibiotics.fa	Ion Torrent (1)	MIRA	blaTEM	178	0	0	62.24%	disagreement
antibiotics.fa	Ion Torrent (1+2)	Newbler	blaTEM	286	0	0	100.00%	agreement
antibiotics.fa	Ion Torrent (1+2)	MIRA	blaTEM	178	0	0	62.24%	disagreement
antibiotics.fa	Ion Torrent (2)	Newbler	blaTEM	286	0	0	100.00%	agreement
antibiotics.fa	Ion Torrent (2)	MIRA	blaTEM	178	0	0	62.24%	disagreement
antibiotics.fa	MiSeq (contigs)	MIRA	blaTEM	286	0	0	100.00%	agreement
antibiotics.fa	MiSeq (contigs)	Velvet	blaTEM	286	0	0	100.00%	agreement
antibiotics.fa	MiSeq (contigs)	CLC	blaTEM	286	0	0	100.00%	agreement
antibiotics.fa	MiSeq (scaffolds)	Velvet	blaTEM	286	0	0	100.00%	agreement
antibiotics.fa	MiSeq (scaffolds)	CLC	blaTEM	286	0	0	100.00%	agreement
antibiotics.fa	280 Reference	Reference	irod7_orf00018 amino	209	0	0	100.00%	agreement
antibiotics.fa	454 Junior (1)	Newbler	irod7_orf00018 amino	209	0	0	100.00%	agreement
antibiotics.fa	454 Junior (1)	MIRA	irod7_orf00018 amino	209	0	0	100.00%	agreement
antibiotics.fa	454 Junior (1+2)	Newbler	irod7_orf00018 amino	209	0	0	100.00%	agreement
antibiotics.fa	454 Junior (1+2)	MIRA	irod7_orf00018 amino	209	0	0	100.00%	agreement
antibiotics.fa	454 Junior (2)	Newbler	irod7_orf00018 amino	209	0	0	100.00%	agreement
antibiotics.fa	454 Junior (2)	MIRA	irod7_orf00018 amino	54	38	2	7.66%	disagreement
antibiotics.fa	C236-11	Reference	irod7_orf00018 amino	209	0	0	100.00%	agreement
antibiotics.fa	Ion Torrent (1)	Newbler	irod7_orf00018 amino	209	0	0	100.00%	agreement
antibiotics.fa	Ion Torrent (1)	MIRA	irod7_orf00018 amino	209	0	0	100.00%	agreement
antibiotics.fa	Ion Torrent (1+2)	Newbler	irod7_orf00018 amino	209	0	0	100.00%	agreement
antibiotics.fa	Ion Torrent (1+2)	MIRA	irod7_orf00018 amino	209	0	0	100.00%	agreement
antibiotics.fa	Ion Torrent (2)	Newbler	irod7_orf00018 amino	209	0	0	100.00%	agreement
antibiotics.fa	Ion Torrent (2)	MIRA	irod7_orf00018 amino	209	0	0	100.00%	agreement
antibiotics.fa	MiSeq (contigs)	MIRA	irod7_orf00018 amino	209	0	0	100.00%	agreement
antibiotics.fa	MiSeq (contigs)	Velvet	irod7_orf00018 amino	209	0	0	100.00%	agreement
antibiotics.fa	MiSeq (contigs)	CLC	irod7_orf00018 amino	209	0	0	100.00%	agreement
antibiotics.fa	MiSeq (scaffolds)	Velvet	irod7_orf00018 amino	209	0	0	100.00%	agreement
antibiotics.fa	MiSeq (scaffolds)	CLC	irod7_orf00018 amino	209	0	0	100.00%	agreement
antibiotics.fa	280 Reference	Reference	lcl||EcE24377A_B000	200	1	0	99.50%	agreement
antibiotics.fa	454 Junior (1)	Newbler	lcl||EcE24377A_B000	200	1	0	99.50%	agreement
antibiotics.fa	454 Junior (1)	MIRA	lcl||EcE24377A_B000	200	1	0	99.50%	agreement
antibiotics.fa	454 Junior (1+2)	Newbler	lcl||EcE24377A_B000	200	1	0	99.50%	agreement
antibiotics.fa	454 Junior (1+2)	MIRA	lcl||EcE24377A_B000	200	1	0	99.50%	agreement
antibiotics.fa	454 Junior (2)	Newbler	lcl||EcE24377A_B000	200	1	0	99.50%	agreement
antibiotics.fa	454 Junior (2)	MIRA	lcl||EcE24377A_B000	54	38	2	8.00%	disagreement
antibiotics.fa	C236-11	Reference	lcl||EcE24377A_B000	200	1	0	99.50%	agreement
antibiotics.fa	Ion Torrent (1)	Newbler	lcl||EcE24377A_B000	200	1	0	99.50%	agreement
antibiotics.fa	Ion Torrent (1)	MIRA	lcl||EcE24377A_B000	200	1	0	99.50%	agreement
antibiotics.fa	Ion Torrent (1+2)	Newbler	lcl||EcE24377A_B000	200	1	0	99.50%	agreement
antibiotics.fa	Ion Torrent (1+2)	MIRA	lcl||EcE24377A_B000	200	1	0	99.50%	agreement
antibiotics.fa	Ion Torrent (2)	Newbler	lcl||EcE24377A_B000	200	1	0	99.50%	agreement
antibiotics.fa	Ion Torrent (2)	MIRA	lcl||EcE24377A_B000	200	1	0	99.50%	agreement
antibiotics.fa	MiSeq (contigs)	MIRA	lcl||EcE24377A_B000	200	1	0	99.50%	agreement
antibiotics.fa	MiSeq (contigs)	Velvet	lcl||EcE24377A_B000	200	1	0	99.50%	agreement
antibiotics.fa	MiSeq (contigs)	CLC	lcl||EcE24377A_B000	200	1	0	99.50%	agreement
antibiotics.fa	MiSeq (scaffolds)	Velvet	lcl||EcE24377A_B000	200	1	0	99.50%	agreement

antibiotics.fa	MiSeq (scaffolds)	CLC	lcl||EcE24377A_B0003	200	1	0	99.50%	agreement
microcins.fa	280 Reference	Reference	microcin H47 secretior	383	0	0	100.00%	agreement
microcins.fa	454 Junior (1)	Newbler	microcin H47 secretior	206	0	0	53.79%	disagreement
microcins.fa	454 Junior (1)	MIRA	microcin H47 secretior	383	0	0	100.00%	agreement
microcins.fa	454 Junior (1+2)	Newbler	microcin H47 secretior	291	0	0	75.98%	disagreement
microcins.fa	454 Junior (1+2)	MIRA	microcin H47 secretior	383	0	0	100.00%	agreement
microcins.fa	454 Junior (2)	Newbler	microcin H47 secretior	291	0	0	75.98%	disagreement
microcins.fa	454 Junior (2)	MIRA	microcin H47 secretior	277	221	62	14.62%	disagreement
microcins.fa	C236-11	Reference	microcin H47 secretior	383	0	0	100.00%	agreement
microcins.fa	Ion Torrent (1)	Newbler	microcin H47 secretior	256	11	0	63.97%	disagreement
microcins.fa	Ion Torrent (1)	MIRA	microcin H47 secretior	383	0	0	100.00%	agreement
microcins.fa	Ion Torrent (1+2)	Newbler	microcin H47 secretior	383	0	0	100.00%	agreement
microcins.fa	Ion Torrent (1+2)	MIRA	microcin H47 secretior	383	0	0	100.00%	agreement
microcins.fa	Ion Torrent (2)	Newbler	microcin H47 secretior	351	9	0	89.30%	disagreement
microcins.fa	Ion Torrent (2)	MIRA	microcin H47 secretior	291	0	0	75.98%	disagreement
microcins.fa	MiSeq (contigs)	MIRA	microcin H47 secretior	383	0	0	100.00%	agreement
microcins.fa	MiSeq (contigs)	Velvet	microcin H47 secretior	383	0	0	100.00%	agreement
microcins.fa	MiSeq (contigs)	CLC	microcin H47 secretior	383	0	0	100.00%	agreement
microcins.fa	MiSeq (scaffolds)	Velvet	microcin H47 secretior	383	0	0	100.00%	agreement
microcins.fa	MiSeq (scaffolds)	CLC	microcin H47 secretior	383	0	0	100.00%	agreement
adhesins.fa	280 Reference	Reference	AggA	167	0	0	100.00%	agreement
adhesins.fa	454 Junior (1)	Newbler	AggA	105	0	0	62.87%	disagreement
adhesins.fa	454 Junior (1)	MIRA	AggA	35	26	0	5.39%	disagreement
adhesins.fa	454 Junior (1+2)	Newbler	AggA	167	0	0	100.00%	agreement
adhesins.fa	454 Junior (1+2)	MIRA	AggA	167	0	0	100.00%	agreement
adhesins.fa	454 Junior (2)	Newbler	AggA	167	0	0	100.00%	agreement
adhesins.fa	454 Junior (2)	MIRA	AggA	172	27	10	86.83%	disagreement
adhesins.fa	C236-11	Reference	AggA	167	0	0	100.00%	agreement
adhesins.fa	Ion Torrent (1)	Newbler	AggA	167	0	0	100.00%	agreement
adhesins.fa	Ion Torrent (1)	MIRA	AggA	167	0	0	100.00%	agreement
adhesins.fa	Ion Torrent (1+2)	Newbler	AggA	167	0	0	100.00%	agreement
adhesins.fa	Ion Torrent (1+2)	MIRA	AggA	167	0	0	100.00%	agreement
adhesins.fa	Ion Torrent (2)	Newbler	AggA	167	0	0	100.00%	agreement
adhesins.fa	Ion Torrent (2)	MIRA	AggA	167	0	0	100.00%	agreement
adhesins.fa	MiSeq (contigs)	MIRA	AggA	167	0	0	100.00%	agreement
adhesins.fa	MiSeq (contigs)	Velvet	AggA	167	0	0	100.00%	agreement
adhesins.fa	MiSeq (contigs)	CLC	AggA	167	0	0	100.00%	agreement
adhesins.fa	MiSeq (scaffolds)	Velvet	AggA	167	0	0	100.00%	agreement
adhesins.fa	MiSeq (scaffolds)	CLC	AggA	167	0	0	100.00%	agreement
adhesins.fa	280 Reference	Reference	AggB	81	0	0	75.70%	agreement
adhesins.fa	454 Junior (1)	Newbler	AggB	117	11	10	99.07%	disagreement
adhesins.fa	454 Junior (1)	MIRA	AggB	28	18	6	9.35%	disagreement
adhesins.fa	454 Junior (1+2)	Newbler	AggB	81	0	0	75.70%	agreement
adhesins.fa	454 Junior (1+2)	MIRA	AggB	117	11	10	99.07%	disagreement
adhesins.fa	454 Junior (2)	Newbler	AggB	81	0	0	75.70%	agreement
adhesins.fa	454 Junior (2)	MIRA	AggB	117	11	10	99.07%	disagreement
adhesins.fa	C236-11	Reference	AggB	81	0	0	75.70%	agreement
adhesins.fa	Ion Torrent (1)	Newbler	AggB	116	9	9	100.00%	disagreement
adhesins.fa	Ion Torrent (1)	MIRA	AggB	116	9	9	100.00%	disagreement
adhesins.fa	Ion Torrent (1+2)	Newbler	AggB	116	9	9	100.00%	disagreement
adhesins.fa	Ion Torrent (1+2)	MIRA	AggB	81	0	0	75.70%	agreement
adhesins.fa	Ion Torrent (2)	Newbler	AggB	116	9	9	100.00%	disagreement
adhesins.fa	Ion Torrent (2)	MIRA	AggB	116	9	9	100.00%	disagreement
adhesins.fa	MiSeq (contigs)	MIRA	AggB	81	0	0	75.70%	agreement
adhesins.fa	MiSeq (contigs)	Velvet	AggB	81	0	0	75.70%	agreement
adhesins.fa	MiSeq (contigs)	CLC	AggB	81	0	0	75.70%	agreement
adhesins.fa	MiSeq (scaffolds)	Velvet	AggB	81	0	0	75.70%	agreement
adhesins.fa	MiSeq (scaffolds)	CLC	AggB	81	0	0	75.70%	agreement
adhesins.fa	280 Reference	Reference	AggC	860	0	0	100.00%	agreement
adhesins.fa	454 Junior (1)	Newbler	AggC	769	0	0	89.42%	disagreement
adhesins.fa	454 Junior (1)	MIRA	AggC	853	571	48	32.79%	disagreement
adhesins.fa	454 Junior (1+2)	Newbler	AggC	769	0	0	89.42%	disagreement
adhesins.fa	454 Junior (1+2)	MIRA	AggC	471	7	0	53.95%	disagreement

adhesins.fa	454 Junior (2)	Newbler	AggC	772	2	0	89.53%	disagreement
adhesins.fa	454 Junior (2)	MIRA	AggC	738	7	0	85.00%	disagreement
adhesins.fa	C236-11	Reference	AggC	860	0	0	100.00%	agreement
adhesins.fa	Ion Torrent (1)	Newbler	AggC	737	1	0	85.58%	disagreement
adhesins.fa	Ion Torrent (1)	MIRA	AggC	404	0	0	46.98%	disagreement
adhesins.fa	Ion Torrent (1+2)	Newbler	AggC	731	0	0	85.00%	disagreement
adhesins.fa	Ion Torrent (1+2)	MIRA	AggC	769	0	0	89.42%	disagreement
adhesins.fa	Ion Torrent (2)	Newbler	AggC	737	1	0	85.58%	disagreement
adhesins.fa	Ion Torrent (2)	MIRA	AggC	731	0	0	85.00%	disagreement
adhesins.fa	MiSeq (contigs)	MIRA	AggC	860	0	0	100.00%	agreement
adhesins.fa	MiSeq (contigs)	Velvet	AggC	550	0	0	63.95%	disagreement
adhesins.fa	MiSeq (contigs)	CLC	AggC	860	0	0	100.00%	agreement
adhesins.fa	MiSeq (scaffolds)	Velvet	AggC	550	0	0	63.95%	disagreement
adhesins.fa	MiSeq (scaffolds)	CLC	AggC	860	0	0	100.00%	agreement
adhesins.fa	280 Reference	Reference	AggD	254	0	0	39.20%	agreement
adhesins.fa	454 Junior (1)	Newbler	AggD	254	0	0	39.20%	agreement
adhesins.fa	454 Junior (1)	MIRA	AggD	225	17	0	32.10%	disagreement
adhesins.fa	454 Junior (1+2)	Newbler	AggD	254	0	0	39.20%	agreement
adhesins.fa	454 Junior (1+2)	MIRA	AggD	254	0	0	39.20%	agreement
adhesins.fa	454 Junior (2)	Newbler	AggD	241	7	2	36.11%	disagreement
adhesins.fa	454 Junior (2)	MIRA	AggD	254	0	0	39.20%	agreement
adhesins.fa	C236-11	Reference	AggD	254	0	0	39.20%	agreement
adhesins.fa	Ion Torrent (1)	Newbler	AggD	241	7	2	36.11%	disagreement
adhesins.fa	Ion Torrent (1)	MIRA	AggD	141	0	0	21.76%	disagreement
adhesins.fa	Ion Torrent (1+2)	Newbler	AggD	241	7	2	36.11%	disagreement
adhesins.fa	Ion Torrent (1+2)	MIRA	AggD	141	0	0	21.76%	disagreement
adhesins.fa	Ion Torrent (2)	Newbler	AggD	241	7	2	36.11%	disagreement
adhesins.fa	Ion Torrent (2)	MIRA	AggD	101	7	2	14.51%	disagreement
adhesins.fa	MiSeq (contigs)	MIRA	AggD	254	0	0	39.20%	agreement
adhesins.fa	MiSeq (contigs)	Velvet	AggD	231	0	0	35.65%	disagreement
adhesins.fa	MiSeq (contigs)	CLC	AggD	254	0	0	39.20%	agreement
adhesins.fa	MiSeq (scaffolds)	Velvet	AggD	231	0	0	35.65%	disagreement
adhesins.fa	MiSeq (scaffolds)	CLC	AggD	254	0	0	39.20%	agreement

Chapter 6

Statement of contribution to work

6.1 Paper I

NJL helped plan the sequencing runs. NJL designed and performed the bioinformatics analysis (assembly, alignment, SNP calling) and helped draft the manuscript.

6.2 Paper II

NJL performed the bioinformatics analysis (assembly, alignment, SNP calling, comparison of strains). NJL designed validation primers for PCR. NJL wrote the manuscript with MH.

6.3 Paper III

NJL initiated the crowd-sourcing analysis, performed the Ion Torrent and Illumina assemblies and deposited the sequences in Genbank, performed phylogenetic and comparative analysis of the outbreak strain including recapitulating existing crowd-sourcing results, generated the circular comparison figure and wrote the supplementary materials.

6.4 Paper IV

NJL helped conceive the project. NJL designed the bioinformatics analysis and performed sequence read mapping, de novo assembly, assembly comparison, pathogen biology comparison, wrote the analysis scripts and created the Github repository. NJL write the results and methods section of the manuscript and prepared all figures, tables and supplementary materials.