bioRxiv preprint doi: https://doi.org/10.1101/019356; this version posted May 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Real-time strain typing and analysis of antibiotic resistance potential using Nanopore MinION sequencing

Minh Duc Cao[1*], Devika Ganesamoorthy[1*], Alysha G. Elliott[1], Huihui Zhang[1], Matthew A. Cooper[1^] and Lachlan Coin[1^].

[1] Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia.

* These authors contributed equally to the work.

^ Corresponding authors: Matthew Cooper m.cooper@imb.uq.edu.au and Lachlan Coin l.coin@imb.uq.edu.au

Abstract

Clinical pathogen sequencing has significant potential to drive informed treatment of patients with unknown bacterial infection. However, the lack of rapid sequencing technologies with concomitant analysis has impeded clinical adoption in infection diagnosis. Here we demonstrate that commercially-available Nanopore sequencing devices can identify bacterial species and strain information with less than one hour of sequencing time, initial drug-resistance profiles within 2 hours, and a complete resistance profile within 12 hours. We anticipate these devices and associated analysis methods may become useful clinical tools to guide appropriate therapy in time-critical clinical presentations such as bacteraemia and sepsis.

bioRxiv preprint doi: https://doi.org/10.1101/019356; this version posted May 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

INTRODUCTION

High throughput sequencing (HTS) has recently become the frontier technology in genomics and has transformed genetic research[1,2]. The pace of change in this field has been rapid and there have been several new sequencing instruments introduced to the market in recent years. One recent addition to the growing spectrum of HTS field is a portable MinION sequencing device from Oxford Nanopore Technologies.

DNA sequencing with biological nanopore was proposed in the 1990s[3]. However, only recently a prototype version of a nanopore sequencing device, the MinION, was released by Oxford Nanopore Technologies. The MinION sequencer measures the change in electrical current as single-stranded DNA passes through the nanopore and the difference in electrical current determines the nucleotide sequence of each DNA strand[4,5]. Simultaneous base calling of multiple strands of DNA passing through multiple nanopores generates sequence fragments, which permits real-time analysis of the sequence data.

In recent years HTS has become an integrative tool for infectious disease analysis[6,7]. There have been several reports emphasizing the use of HTS methods to characterize clinical isolates, to study the spread of drug resistant microorganisms and to investigate outbreak of infections[8-10]. However, there have been several hurdles to widespread adoption of HTS as the 'method-of-choice' in the clinic for determining the infectious agent and guiding patient treatment: a) lack of portability; b) high cost of the sequencing devices and c) difficulty in obtaining actionable data within a few hours. The miniature, portable and low-cost MinION sequencer overcomes two of these hurdles, but it is currently unclear the extent to which clinically actionable data can be obtained within a few hours using this device.

In this article we report real-time analysis and characterization of three *Klebsiella pneumoniae* strains via Nanopore sequencing (Figure 1). We demonstrate that we can determine the species and strain type of the sequenced sample within an hour of sequencing. Furthermore we show that we can identify ~50% of the drug resistance genes present in a sample within 2 hours of sequencing, and the full drug resistance profile within 12 hours. We also show that Nanopore sequence data can be used for accurate Multi-Locus Sequence Typing (MLST) despite the relatively high base-calling error rates previously reported[11,12]. Our findings support the potential use of Nanopore sequencing for real time analysis of clinical samples for species detection and analysis of antibiotic resistance.

bioRxiv preprint doi: https://doi.org/10.1101/019356; this version posted May 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

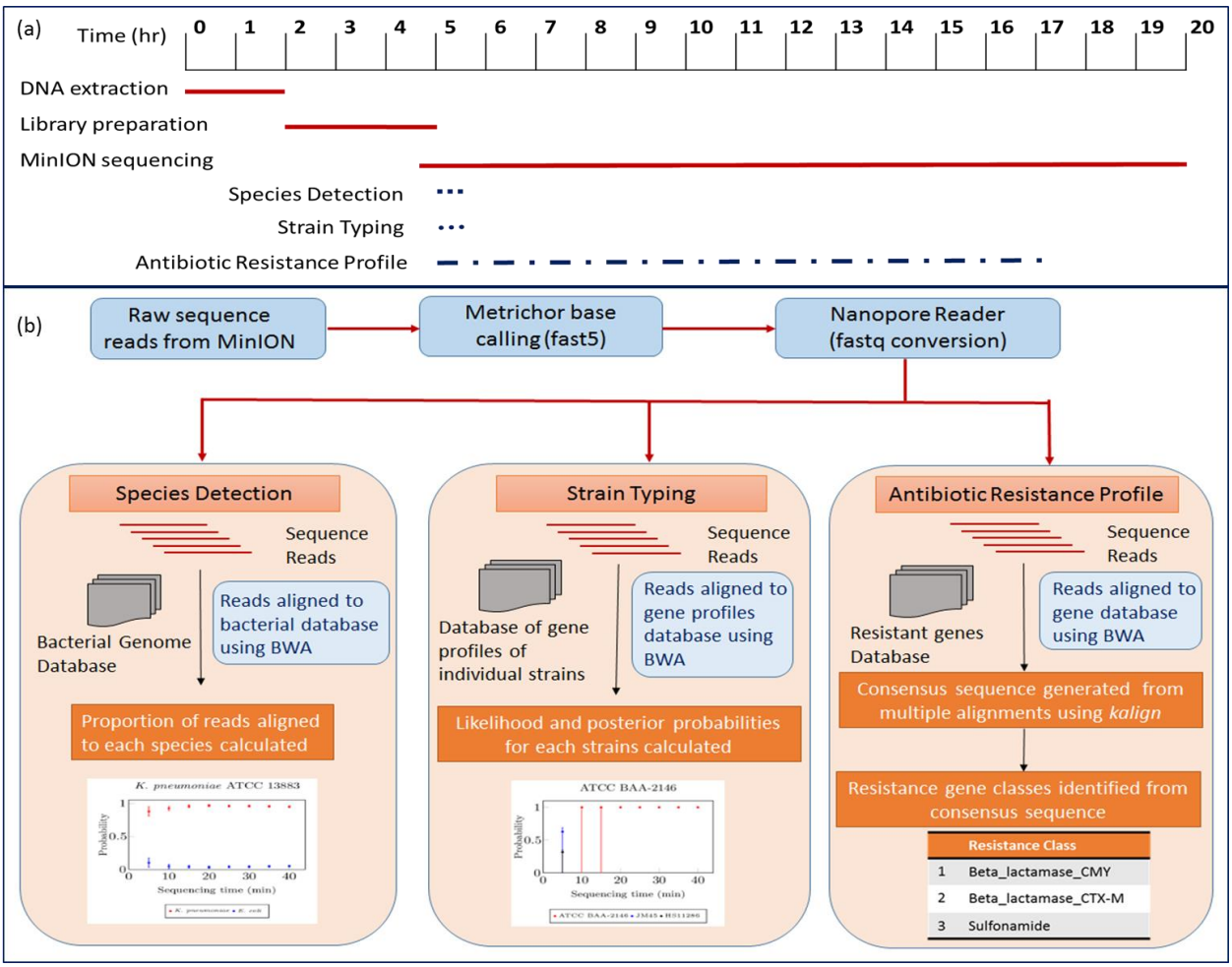

Figure 1: Real-time analysis of nanopore MinION sequence data. (a) Timeline for laboratory workflow and comprehensive bacterial genome analysis using Nanopore MinION sequencing. Species and strain information from a DNA sample can be determined within 3 to 4 hours and antibiotic resistance profile can be determined within 15 hours. (b) Real-time analysis pipeline of Nanopore MinION sequence data. All parts of the pipeline are processed simultaneously to allow for real-time analysis. Raw sequence reads from Nanopore are base called using Metrichor program. Base called reads which are in fast5 fromat are converted into fastq format using our Nanopore reader program. Converted fastq reads are streamed into *species detection, strain typing* and *antibiotic resistance* profile modules.

RESULTS

We sequenced the genomes of three *Klebsiella pneumoniae* strains, ATCC BAA-2146, ATCC 700603 and ATCC 13883 using the Nanopore MinION device. Nanopore sequence reads are classified into three types: template, complement and higher quality 2D reads (i.e. two direction reads, which contains the computationally merged template and complement read). Samples sequenced with the R7 flow cell yielded 12% of 2D reads but it was doubled with the improved R7.3 flow cell (Table 1). Although our sequence

bioRxiv preprint doi: https://doi.org/10.1101/019356; this version posted May 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

yields were lower, the read length and accuracy of the sequence data were similar to other user reports[11-14]. We observed that the majority of data (greater than 75% for R7 and greater than 66% for R7.3) were generated in the first 16 hours of sequencing time (Figure 2).

Table 1: Sequence read statistics

Sample	Nanopore Flow cells	Hours of sequencing	Total reads	Template reads	Complement reads	2D reads	Sequence Yield (Mb)
ATCC BAA-2146	R7	60	37543	26439	6604	4500	184.80
ATCC 700603	R7	60	7208	5139	1233	836	39.37
ATCC 13883	R7.3	36	15564	7212	4555	3797	86.83

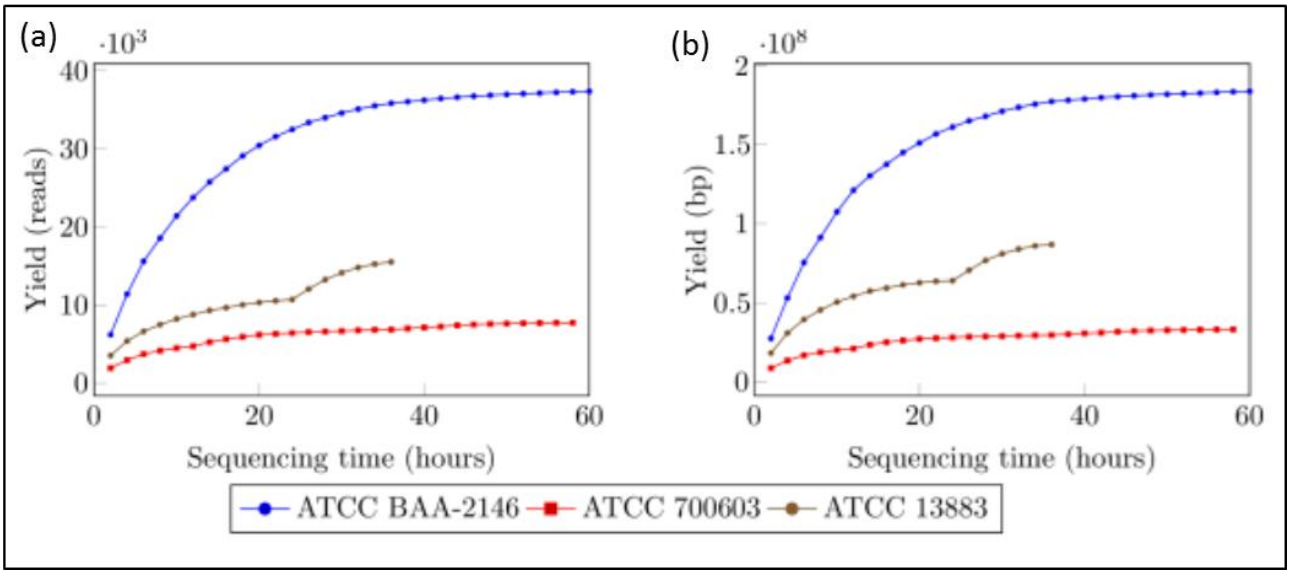

Figure 2: Sequencing yield over time for all 3 *Klebsiella pneumoniae* samples (a) Number of reads over time and (b) Number of base pairs sequenced over time.

bioRxiv preprint doi: https://doi.org/10.1101/019356; this version posted May 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Species Detection:

To illustrate the potential of rapid 'real-time' species detection with the MinION sequencer, we built a 'streaming' species detection computational pipeline. As each new read is sequenced, our pipeline updates an estimate of the proportion of DNA present in the sample which belongs to each of 2,785 bacterial genomes currently available in GenBank (http://www.ncbi.nlm.nih.gov/genbank/), as well as an estimate of the uncertainty in this proportion (See Methods).

In all three sequenced samples, we successfully detected *K.pneumoniae* as the major species present in the isolate with 99% confidence. This was achieved with as little as 500 sequence reads requiring less than 20 minutes of sequencing time (Figure 3). We assessed our species detection method on an *Escherichia coli* (strain K12 MG1655) Nanopore MinION sequencing data set published by Quick et al[14]. Our method successfully identified the species in their sequence data set with approximately the same amount of data required for our samples.

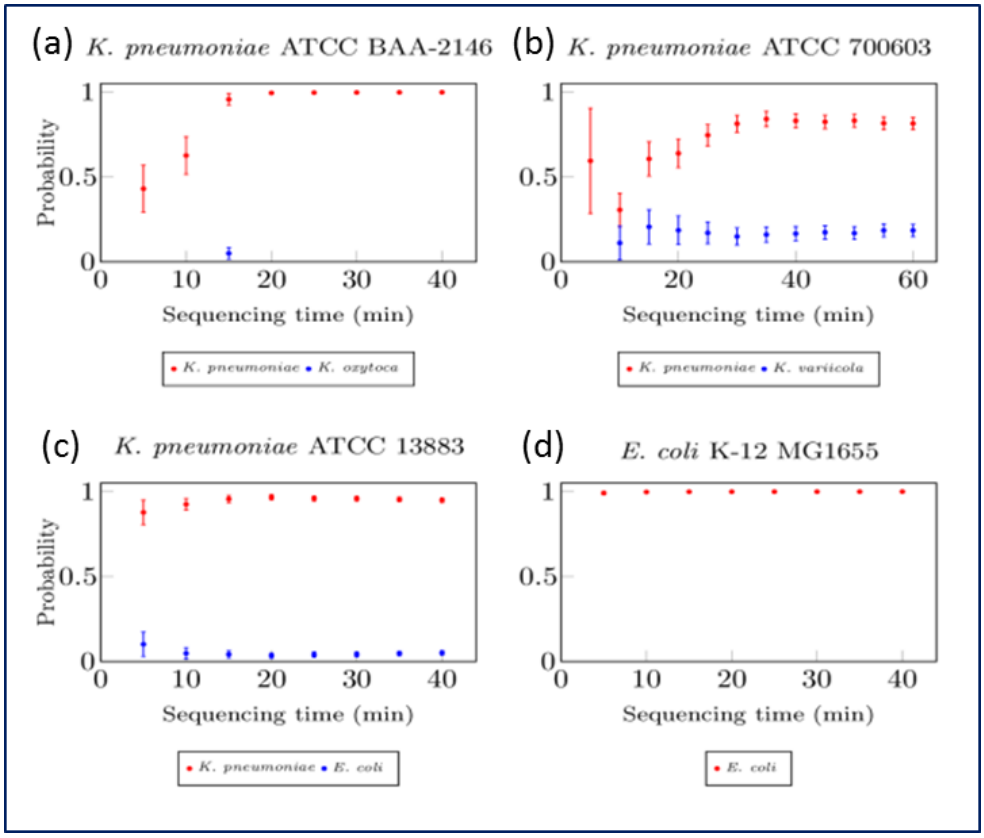

Figure 3: Real-time identification of bacterial species from Nanopore MinION sequence reads on four different bacterial genomes. (a) *K.pneumoniae* ATCC BAA-2146, (b) *K.pneumoniae* ATCC 700603, (c) *K.pneumoniae* ATCC 13883 and (d) *Escherichia coli* K12 MG1655.

bioRxiv preprint doi: https://doi.org/10.1101/019356; this version posted May 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Interestingly, our pipeline identified approximately 20% of the ATCC 700603 sample (*K.pneumoniae*) as *Klebsiella variicola*. We downloaded the assembly of ATCC 700603 strain (Accession ID NZ_AOGO00000000.1) and performed phylogenetic analysis with 5 *K.variicola* assemblies and 9 *K.pneumoniae* complete genomes available on GenBank (Supplementary Figure 1). We found that the *K.pneumoniae* ATCC 700603 strain was in fact an ancestor of all other *K.pneumoniae* and *K.variicola* strains with available sequencing data in GenBank. This explains the shared identity detected in Nanopore sequencing data. Finally, when we included the ATCC 700603 assembly in the bacterial database, the species detection pipeline identified *K.pneumoniae* as the only species in the ATCC 700603 sample (Supplementary Figure 2).

Multi-locus Sequence Typing:

K.pneumoniae are conventionally strain typed using an MLST system (http://bigsdb.web.pasteur.fr/klebsiella/klebsiella.html), which requires accurate genotyping to distinguish the alleles of seven house-keeping genes[15]. Previous reports indicating a high base-calling error rate[11,12] suggested that MLST typing may be challenging with MinION sequence data.

We developed a pipeline to carry out MLST typing using MinION sequence data which first corrects errors in the raw sequence reads and subsequently combines information across multiple SNPs in a likelihood-based framework (See Methods). Table 2 presents the top five highest score types (in log-likelihood) for three *K.pneumoniae* strains using Nanopore sequencing. In all three strains, the correct types were the highest score out of 1678 types available in the MLST database. However, we noticed that the typing system also outputted several other types with the same likelihood (i.e., types 751 and 864 for strain ATCC BAA-2146 and type 851 for strain ATCC 700603). We examined the profiles of these types, and found that for strain ATCC BAA-2146, types 751 and 864 differ to the correct type 11 by only one SNP from the total of 3012 bases in seven genes (see Supplementary Note 3). For strain ATCC 700603, type 851 differs to type 489 by two alleles (in genes phoE and tonB), but there was only one read mapped to each of these genes. These results suggest a more accurate strain-typing methodology would consider all of the sequenced reads, rather than just those covering 7 house-keeping genes, so we further devised a method for strain-typing which was based on presence or absence of genes.

bioRxiv preprint doi: https://doi.org/10.1101/019356; this version posted May 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 2: Multi-locus Strain-typing results for three *K.pneumoniae* strains. The top five probable MLST types are shown for each sample.

	ATCC BAA-2146 (MLST Type 11)		**ATCC 700603 (MLST Type 489)**		**ATCC 13883 (MLST Type 3)**	
Rank	**Type**	**Score**	**Type**	**Score**	**Type**	**Score**
1	**11**	1985.47	**489**	418.45	**3**	1451.65
2	*751*	*1985.47*	*851*	*418.45*	136	1450.21
3	*864*	*1985.47*	257	413.57	38	1444.81
4	1080	1984.46	356	413.57	1106	1444.19
5	1680	1982.62	414	413.57	931	1441.44

Strain Detection by presence or absence of genes

We propose a novel 'real-time' strain typing method to identify the bacterial strain from the Nanopore sequence reads based on the presence or absence patterns of genes. We obtained the genome assemblies of 235 *K.pneumoniae* strains from GenBank (http://www.ncbi.nlm.nih.gov/genbank/) and Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) and annotated them with Prokka[16] to create a gene database and a gene profile for each *K.pneumoniae* strain. Our pipeline identifies which genes in the database are present in each read (fully or partially) as it is generated by the nanopore sequencer, and use this information to update the posterior probability of each of the 235 strains , as well as the 99% confidence intervals in this estimate (see Methods).

We applied this pipeline to the 3 sequenced *K.pneumoniae* strains (Figure 4). We successfully identified the corresponding strains from the sequence data with 99% confidence within 25 minutes of sequencing time and with as low as 500 sequencing reads.

bioRxiv preprint doi: https://doi.org/10.1101/019356; this version posted May 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

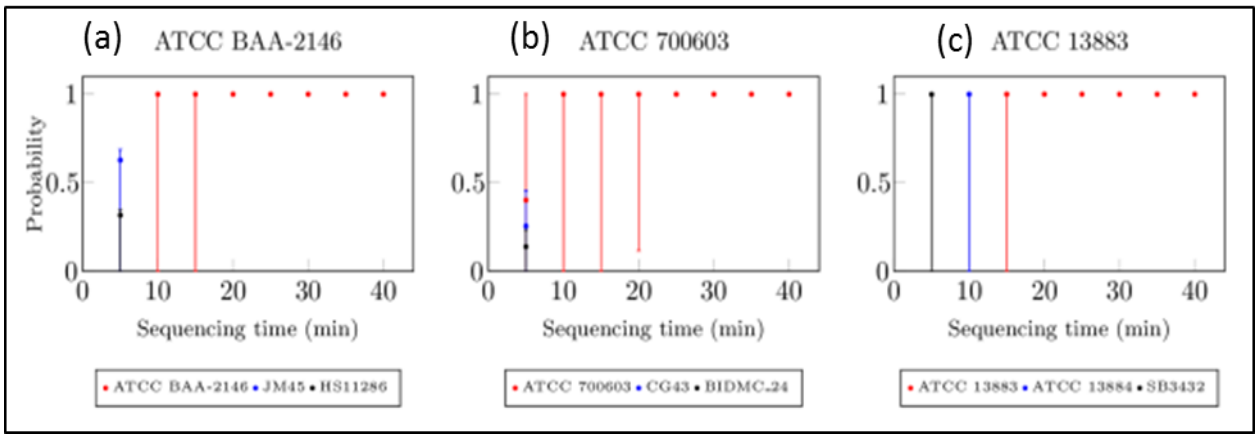

Figure 4: Real-time identification of strain type from Nanopore MinION sequence reads on three different *K.pneumoniae* strains (a) ATCC BAA-2146 (b) ATCC 700603 and (c) ATCC 13883 strains successfully identified using the pipeline.

Antibiotic resistance detection:

The antibiotic resistance profile of the three *K.pneumoniae* strains were characterised with Nanopore MinION sequencing data. We obtained antibiotic drug resistance genes from "The Comprehensive Antibiotic Resistance Database" (http://arpcard.mcmaster.ca/)[17], and grouped them into 38 different antibiotic classes based on the gene annotation. We applied a pipeline to detect antibiotic resistance genes from the sequencing data and to report the classes of these genes in real-time (See Methods). The resulting antibiotic gene profiles were then validated with the gene profiles we obtained from the reference genome sequences of the respective strains.

bioRxiv preprint doi: https://doi.org/10.1101/019356; this version posted May 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 3: Timeline of detecting antibiotic resistance gene classes from Nanopore MinION sequencing data on three *K.pneumoniae* strains ATCC BAA-2146, ATCC 700603 and ATCC 13883.

Time	Classes detected	TP/FP	Sensitivity	Specificity	Accuracy	Data
***K.pneumoniae* ATCC BAA-2146 (17 classes^)**						
1 hour	Aminoglycoside Beta_lactamase_CMY Beta_lactamase_CTX-M Beta_lactam_OXA Beta_lactam_SHV Beta_lactamase_others Daptomycin-Rifampin Sulfonamide	TP TP TP TP TP TP TP TP	47.06%	100.00%	76.32%	2817 reads 11.66Mb
2 hours	Fluoroquinolone Trimethoprim	TP	58.82%	100.00%	81.58%	6215 reads 27.31Mb
3 hours	Aminocoumarin Beta_lactamase_TEM Bleomycin	TP TP TP	76.47%	100%	89.47%	9072 reads 41.44Mb
7 hours	Chloramphenicol-Phenicol Elfamycin Tetracycline	FP TP FP	88.23%	95.23%	92.11%	17150reads 83.66Mb
12 hours	NDM-1 Fosfomycin	TP TP	94.12%	90.48%	92.11%	23761 reads 121.15Mb
60 hours	No more classes detected					37543 reads 184.80Mb
***K. pneumoniae* ATCC 700603 (9 classes^)**						
1 hour	Aminocoumarin Beta_lactam_SHV Fosfomycin	TP TP TP	33.33%	100.00%	84.21%	1133 reads 4.86Mb
2 hours	Daptomycin-Rifampin	TP	44.44%	100.00%	86.84%	1938reads 8.59Mb
4 hours	Beta_lactamase_others	TP	55.55%	100.00%	89.47%	2960reads 13.34Mb
7 hours	Beta_lactam_OXA	TP	66.67%	100.00%	92.11%	3978 reads 17.80Mb
60 hours	No more classes detected					7748 reads 33.07Mb
***K. pneumoniae* ATCC 13883 (8 classes^)**						
1 hour	Beta_lactam_SHV Beta_lactam_others Daptomycin-Rifampin Fosfomycin Peptide_antibiotic Sulfonamide	TP TP TP TP FP TP	62.50%	96.67%	89.47%	2158 reads 9.63Mb
2 hours	Aminocoumarin Elfamycin Fluoroquinolone	TP TP TP	100.00%	96.67%	97.37%	3543 reads 18.12Mb
12 hours	Aminoglycoside	FP	100.00%	93.33%	94.73%	8793 reads 54.03 Mb
36 hours	No more classes detected					15564 reads 86.83Mb

^ Number of antibiotic classes detected in the reference TP – true positive, FP – false positive

bioRxiv preprint doi: https://doi.org/10.1101/019356; this version posted May 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 3 shows the classes of antibiotic genes detected from Nanopore MinION sequencing of three *K.pneumoniae* strains over time. For the NDM-1 producing *K.pneumoniae* strain ATCC BAA-2146 (17 classes of antibiotic resistance classes based on the reference genome), the pipeline detected 18 classes after 12 hours of sequencing with 94.12% sensitivity and 90.48% specificity. Almost 50% of the classes (8 out of 17) were detected after just 1 hour of sequencing. We observed similar performance for *K.pneumoniae* type strain ATCC 13883 where 6 out of 8 classes were detected after 1 hour with one false positive (specificity 96.67%). After 12 hours of sequencing, all classes were detected including 2 false positives (sensitivity 100% and specificity 93.33%). For the multi-drug resistant *K.pneumoniae* strain ATCC 700603, the pipeline only detected 6 out of 9 classes, however without any false positives (sensitivity 66.67% and specificity 100%). The reduced sensitivity for this sample is most likely due to the low sequence yield (33Mb of data in total).

To evaluate the overall performance of the antibiotic gene profiles identification, we varied the stringent threshold to obtain different levels of sensitivity. Figure 5 shows the *Receiver Operating Characteristic* curves of antibiotic resistance gene class detection after 12 hours of sequencing of three *K.pneumoniae* strains. In all three strains, the pipeline was able to detect 100% antibiotic resistance gene classes with only up to 23.80% false positive rates. On the other hand, it detected over three quarters of the antibiotic resistance gene classes without any false positive. Notably, it attained a perfect identification result (AUC=1.0) for sample ATCC 13383, which was sequenced with the better chemistry R7.3.

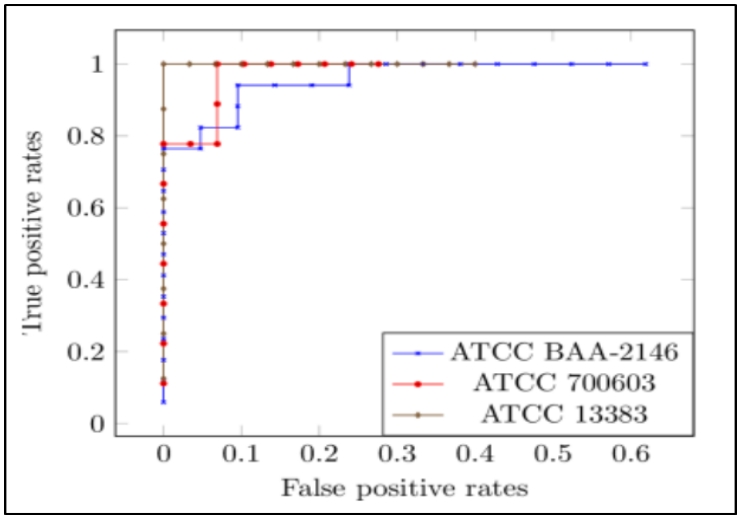

Figure 5: Receiver operating characteristic curves of antibiotic gene profiles identification of three *K.pneumoniae* strains ATCC BAA-2146, ATCC 700603 and ATCC 13883 after 12 hours of sequencing.

bioRxiv preprint doi: https://doi.org/10.1101/019356; this version posted May 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Discussion

One of the major bottlenecks to routine establishment of whole genome pathogen sequencing in the clinic is the lack of a sequencing technology that can return definitive results within a few hours. In this paper we have demonstrated the potential of the Nanopore MinION sequencing device to return clinically actionable results on the pathogen species and strain with less than half an hour of sequencing time, and on the drug resistance profile, useful for therapeutic treatment, within a few hours. This is a major step forward for realising the promise of whole genome sequencing in the clinic. Bradley et al have also recently shown they could accurately identify the drug resistance profile of a Mycobacterium Tuberculosis sample from 8hrs of MinION sequencing using a de-Bruijn graph approach[18] (BioRXiv paper).

One of the major advantages of a whole-genome sequencing approach to drug resistance profiling is that it is not necessary to restrict the analysis to a limited panel of drug-resistance tests but it is possible to discover the complete drug resistance profile in a sample. By having a complete picture of the drug-resistance profile within a few hours, the clinician is able to design an antibiotic treatment regimen which is both much more likely to succeed and less likely to induce further antibiotic resistance.

Whole genome sequencing has further advantages that we have not explored here, but have been addressed by others, including the ability to track the progression of an outbreak. Traditional approaches to this problem require high confidence SNP calls, which are likely to become more viable as the Nanopore sequencing technology improves and as error-correction algorithms become more sophisticated.

Clearly, real-time Nanopore sequencing has a promising future in clinical pathogen sequencing.

bioRxiv preprint doi: https://doi.org/10.1101/019356; this version posted May 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

METHODS

DNA extraction

Bacterial strains *K.pneumoniae* ATCC 13883, ATCC 700603 and ATCC BAA-2146 were obtained from American Type Culture Collection (ATCC, USA). Bacterial cultures were grown overnight from a single colony at 37 °C with shaking (180 rpm). Whole cell DNA was extracted from the cultures using the DNeasy Blood and Tissue Kit (QIAGEN®, Cat # 69504) according to the bacterial DNA extraction protocol with enzymatic lysis pre-treatment.

MinION library preparation - R7

Library preparation was performed using the Genomic DNA Sequencing kit (SQK-MAPP-002) (Nanopore) according to the manufacturer's instruction. Briefly, 1 μg of genomic DNA was sheared to 10kb fragment size using a Covaris g-TUBE. The sheared DNA was end repaired using the NEBNext End Repair Module (New England Biolabs) in a total volume of 100 μL and incubated at 20°C for 30 minutes. The end repaired DNA was purified using 1x volume (100 μL) Agencourt Ampure XP beads (Beckman Coulter) according to the manufacturer's instructions. Purified end repair products were eluted in 42 μL of molecular grade water and dA-tailing was performed using the NEBNext dA-tailing module (New England Biolabs) in a total volume of 50 μL and incubated at 37°C for 30 minutes. Ligation was performed using the reagents supplied by Nanopore and T4 DNA ligase from New England Biolabs. The dA-tailed DNA was mixed with 10 μL of adapter mix, 10 μL of HP adapter, 20 μL of 5x ligation buffer and 10 μL of T4 DNA ligase (20000 units per reaction) and incubated at room temperature for 10 minutes. The adapter-ligated DNA was purified using 0.4x volume (40 μL) Agencourt Ampure XP beads (Beckman Coulter) according to the manufacturer's instructions with slight modifications. Nanopore supplied wash buffer and elution buffer was used and only a single wash was performed. Samples were eluted in 25 μL of elution buffer. The ligated DNA was mixed with 10 μL of tether and incubated at room temperature for 10 minutes. Finally, 15 μL of HP motor was added to the reaction and incubated at room temperature for 16 hours.

MinION library preparation - R7.3

For the R7.3 flow cells an updated Genomic Sequencing kit (SQK-MAPP-003) (Nanopore) was used according to the manufacturer's instruction. Purified end repair products were eluted in 25 μL of molecular grade water and dA-tailing was performed in a total volume of 30 μL. The dA-tailed DNA was mixed with 10 μL of adapter mix,10 μL of HP adapter and 50 μL of Blunt/TA ligase master mix (New England Biolabs) and

bioRxiv preprint doi: https://doi.org/10.1101/019356; this version posted May 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

incubated at room temperature for 10 minutes. The adapter-ligated DNA was purified using 0.4 x volume (40 μL) Agencourt Ampure XP beads (Beckman Coulter) according to the manufacturer's instructions with slight modifications. Nanopore supplied wash buffer and elution buffer was used and only a single wash was performed. Samples were eluted in 25 μL of elution buffer.

MinION Sequencing

For each sample a new MinION flow cell (R7 or R7.3) was used for sequencing. The flow cell was inserted into the MinION device and prior to sequencing, the flow cell was primed using 150 μL of EP buffer twice with 10 minute incubation after each addition. The sequencing library mix was prepared by combining 6 μL of library with 140 μL of EP buffer and 4 μL of fuel mix. The library mix was loaded onto the MinION flow cell and the Genomic DNA 48 hour sequencing protocol was initiated on the MinKNOW software. The MinION flow cell was topped up with fresh library mix for every 12 hours as required.

MinION data analysis

The sequence read data were base called with the Metrichor software using the workflow r7 2D version 1.12. We developed Nanopore Reader (*npreader)* program (available at https://github.com/mdcao/npReader) to convert base-called sequence data in fast5 format to fastq format. The *npreader* also extracted the time that each read was sequenced and used this information to sort the read sequences in order they were produced. We streamlined read data in this order to the analysis pipelines presented below and took measures every five minutes of sequencing data.

Species typing

We downloaded the bacterial genome database on GenBank (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/, accessed 19 Nov 2014), which contained high quality genomes of 2785 bacterial strains from 1487 species (See Supplementary Spread-sheets). Our species typing pipeline streamed read data from *npreader* directly to BWA[19] (see Supplementary Note 1) which aligned the reads to the bacterial genome database. Output from *BWA* was streamed directly into our species typing pipeline, which calculated the proportion of reads aligned to each of these species. We used the MultinomialCI package in R[20] to calculate 95% confidence intervals in the value of this proportion.

bioRxiv preprint doi: https://doi.org/10.1101/019356; this version posted May 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

MLST typing

Nanopore MinION sequence reads from *Klebsiella pneumoniae* strains were aligned to the seven house-keeping genes specified by the MLST system[15] using BWA[19]. We then collected reads that were aligned to a gene and performed a multiple alignment on them using *kalign2*[21]. The consensus sequence created from the multiple alignment was then globally aligned to all alleles of the gene using an implementation of the Finite State Machine[22] for global alignment. The score of a MLST type was determined by the sum of the scores of seven alleles making up the type.

Strain typing

We built a database of genomes and genes of 235 *K.pneumoniae* strains, sourcing from GenBank and the Sequence Read Archive. In particular, we downloaded 9 complete *K.pneumoniae* genomes and 165 draft *K.pneumoniae* assemblies from GenBank (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/ and ftp://ftp.ncbi.nlm.nih.gov/genomes/ASSEMBLY_BACTERIA/Klebsiella_pneumoniae/, accessed 19 Nov 2014). In addition, we obtained the whole genome sequencing of 311 *K. pneumoniae* strains from European Nucleotide Archive (Study ERP000165: http://www.ebi.ac.uk/ena/data/view/ERP000165) These whole genome sequencing datasets were trimmed by *trimmomatic*[23] and then assembled by *spades*[24]. We then selected assemblies that had N50 greater than 100kb, resulting in a list of 235 genomes and assemblies (See Supplementary Spreadsheet1.xlsx). We annotated these 235 strains using *Prokka*[16] to obtain 1,210,072 gene sequences, and grouped these sequences into 23,945 unique genes based on 85% sequence identity. This also contributed the gene profile of each strain.

We align each MinION read to the gene database using BWA[19] *(See Supplementary Note 2)*. Whenever a read r_i aligns to a gene in this database, we calculate the likelihood of observing this alignment under a null model in which we assume that we could have observed an alignment with any gene in the collection of strains in proportion to the number of times a homolog of that gene is found, i.e.

P(r_i overlaps g | any strain S) =

(number of times g is in any strain) / (total number of genes in all strains)

We also calculate the likelihood under the alternative model that our sample consists of strain S_k in a 80%/20% mixture with all other strains as

P(r_i overlaps g | S_k) = 0.8 * 1/(number of genes in S_k) + 0.2 * P($_i$ r overlaps g | any strain S)

bioRxiv preprint doi: https://doi.org/10.1101/019356; this version posted May 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

This likelihood is a mixture of the probability of observing the alignment conditional on species S_k and the null model. This ensures that the probability is never zero, and allows for a low level of incorrect gene annotation due to mis-alignment.

Finally we calculate the posterior probability using Bayes theorem

$$P(S_k|r_{1..}r_m) = \Pi_{i=1..m} P(r_i \mid S_k) / \Sigma_k \Pi_{i=1..m} P(r_i \mid S_k)$$

Antibiotic resistance gene classes detection

We manually classified the 3936 genes from "The Comprehensive Antibiotic Resistance Database" (http://arpcard.mcmaster.ca/)[17] into antibiotics classes and selected 38 antibiotics classes. We then created a benchmark for validation of antibiotic resistance profile identification by aligning the reference genomes of the three strains against the gene database using *dnadiff*[25]. Genes that were aligned with above 90% identity to the reference were considered presented in the genomes.

Our analysis pipeline aligned MinION sequencing data into this gene database using BWA[19] in a streamline fashion, and examined genes that had reads aligned to. Due to the high error rates of raw MinION read data, we noticed a high rate of false positive genes. To further reduce false positives, we used *kalign2*[21] to perform a multiple alignment of reads that were aligned to the same gene. The consensus sequence resulting from the multiple alignment was then compared with the gene sequence using the Finite State Machine[22]. The pipeline then reported gene classes based on the genes detected.

Data Availability:

All sequence data are available in European Nucleotide Archive – Study Accession Number ERP010377 (http://www.ebi.ac.uk/ena/data/view/ERP010377). All analysis pipelines are available in http://www.genomicsresearch.org/public/researcher/npAnalysis/.

Author Contributions:

M.D.C., D.G., M.A.C. and L.C. conceived the study, performed the analysis and wrote the first draft of the manuscript. D.G performed the MinION sequencing. M.D.C., H.Z., and L.C performed the bioinformatics

bioRxiv preprint doi: https://doi.org/10.1101/019356; this version posted May 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

analysis. A.G.E performed the bacterial cultures and DNA extractions. All authors contributed to editing the final manuscript.

Competing Financial Interests:

M.A.C is a participant of Oxford Nanopore's MinION Access Programme (MAP) and received the MinION device and flow cells used for this study free of charge.

bioRxiv preprint doi: https://doi.org/10.1101/019356; this version posted May 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

REFERENCES

1 Boyd, S. D. Diagnostic applications of high-throughput DNA sequencing. *Annual review of pathology* **8**, 381-410, doi:10.1146/annurev-pathol-020712-164026 (2013).

2 Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K. & Mardis, E. R. The next-generation sequencing revolution and its impact on genomics. *Cell* **155**, 27-38, doi:10.1016/j.cell.2013.09.006 (2013).

3 Kasianowicz, J. J., Brandin, E., Branton, D. & Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 13770-13773 (1996).

4 Branton, D. *et al.* The potential and challenges of nanopore sequencing. *Nature biotechnology* **26**, 1146-1153, doi:10.1038/nbt.1495 (2008).

5 Stoddart, D., Heron, A. J., Mikhailova, E., Maglia, G. & Bayley, H. Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 7702-7707, doi:10.1073/pnas.0901054106 (2009).

6 Dunne, W. M., Jr., Westblade, L. F. & Ford, B. Next-generation and whole-genome sequencing in the diagnostic clinical microbiology laboratory. *European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology* **31**, 1719-1726, doi:10.1007/s10096-012-1641-7 (2012).

7 Fricke, W. F. & Rasko, D. A. Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. *Nature reviews. Genetics* **15**, 49-55, doi:10.1038/nrg3624 (2014).

8 Hudson, C. M., Bent, Z. W., Meagher, R. J. & Williams, K. P. Resistance determinants and mobile genetic elements of an NDM-1-encoding Klebsiella pneumoniae strain. *PloS one* **9**, e99209, doi:10.1371/journal.pone.0099209 (2014).

9 Petty, N. K. *et al.* Global dissemination of a multidrug resistant Escherichia coli clone. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 5694-5699, doi:10.1073/pnas.1322678111 (2014).

10 Stoesser, N. *et al.* Genome sequencing of an extended series of NDM-producing Klebsiella pneumoniae isolates from neonatal infections in a Nepali hospital characterizes the extent of community- versus hospital-associated transmission in an endemic setting. *Antimicrobial agents and chemotherapy* **58**, 7347-7357, doi:10.1128/AAC.03900-14 (2014).

11 Ashton, P. M. *et al.* MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nature biotechnology* **33**, 296-300, doi:10.1038/nbt.3103 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/019356; this version posted May 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

12 Kilianski, A. *et al.* Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer. *GigaScience* **4**, 12, doi:10.1186/s13742-015-0051-z (2015).

13 Jain, M. *et al.* Improved data analysis for the MinION nanopore sequencer. *Nature methods* **12**, 351-356, doi:10.1038/nmeth.3290 (2015).

14 Quick, J., Quinlan, A. R. & Loman, N. J. A reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer. *GigaScience* **3**, 22, doi:10.1186/2047-217X-3-22 (2014).

15 Diancourt, L., Passet, V., Verhoef, J., Grimont, P. A. & Brisse, S. Multilocus sequence typing of Klebsiella pneumoniae nosocomial isolates. *Journal of clinical microbiology* **43**, 4178-4182, doi:10.1128/JCM.43.8.4178-4182.2005 (2005).

16 Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069, doi:10.1093/bioinformatics/btu153 (2014).

17 McArthur, A. G. *et al.* The comprehensive antibiotic resistance database. *Antimicrobial agents and chemotherapy* **57**, 3348-3357, doi:10.1128/AAC.00419-13 (2013).

18 Bradley, P. *et al. Rapid antibiotic resistance predictions from genome sequence data for S. aureus and M. tuberculosis*. (2015).

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

20 Sison, C. P. & Glaz, J. Simultaneous Confidence Intervals and Sample Size Determination for Multinomial Proportions. *Journal of the American Statistical Association* **90**, 366-369 (1995).

21 Lassmann, T., Frings, O. & Sonnhammer, E. L. Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic acids research* **37**, 858-865, doi:10.1093/nar/gkn1006 (2009).

22 Allison, L., Wallace, C. S. & Yee, C. N. Finite-state models in the alignment of macromolecules. *Journal of molecular evolution* **35**, 77-89 (1992).

23 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).

24 Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology* **19**, 455-477, doi:10.1089/cmb.2012.0021 (2012).

25 Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome biology* **5**, R12, doi:10.1186/gb-2004-5-2-r12 (2004).