ADAPTIVE NANOPORE SEQUENCING ON MINIATURE FLOW CELL DETECTS EXTENSIVE ANTIMICROBIAL RESISTENCE

PREPRINT

Adrian Viehweger [1*], Mike Marquet [2], Martin Hölzer [3], Nadine Dietze [1], Mathias Pletz [2], Christian Brandt [2]

[1] Institute of Medical Microbiology and Virology, University Hospital Leipzig, Leipzig, Germany [2] Institute for Infectious Diseases and Infection Control, Jena University Hospital, Jena, Germany [3] Methodology and Research Infrastructure, MF1 Bioinformatics, Robert Koch Institute, Berlin, Germany

[*] Corresponding author: adrian.viehweger@medizin.uni-leipzig.de

Keywords Genomic surveillance · Nanopore sequencing · Target enrichment · Antimicrobial resistance · Carbapenemases · Plasmids

Abstract

Rapid screening of hospital admissions to detect asymptomatic carriers of resistant bacteria can prevent pathogen outbreaks. However, the resulting isolates rarely have their genome sequenced due to cost constraints and long turnaround times to get and process the data, limiting their usefulness to the practitioner. Here we use real-time, on-device target enrichment ("adaptive") sequencing on a new type of low-cost nanopore flow cell as a highly multiplexed assay covering 1,147 antimicrobial resistance genes. Using this method, we detected four types of carbapenemase in a single isolate of *Raoultella ornithinolytica* (*NDM*, *KPC*, *VIM*, *OXA*). Further investigation revealed extensive horizontal gene transfer within the underlying microbial consortium, increasing the risk of resistance spreading. Real-time sequencing could thus quickly inform how to monitor this case and its surroundings.

Introduction

Screening patients for multiresistant bacteria on hospital admission can detect asymptomatic colonization early[1] and reduce subsequent complications.[2] However, corresponding isolates rarely have their genome sequenced, which would enable genomic surveillance, and, as a result, source control and reduced spread.[3] Such resistant strains can colonize patients for years, increasing the value of this information.[4] Long-term carriage is surprising in the absence of a selective stimuli such as treatment with antimicrobials. Recently, the underlying microbial consortia in which these strains are embedded have been implicated in resistance maintenance through ongoing horizontal gene transfer of mobile elements.[5,6] This finding suggests that in special cases, genomic surveillance should be expanded to include metagenomic data.[7]

Here we report on a patient with multiple carbapenem-resistant strains detected in a rectal swab. One of the isolates simultaneously carried four carbapenemases, an unusually high number. To support a timely response, we integrated the results from multiple modalities of real-time nanopore sequencing. First, we reconstructed the genomes of individual isolates and then complemented them with metagenomic data from the swab. In a proof-of-concept, we then applied real-time on-device target enrichment of 1,147 resistance genes on a miniature flow cell[8] to create an ultra-high multiplex assay.

Results

During resistance screening of rectal swabs, we found three bacterial species growing on carbapenem agar (*Raoultella ornithinolytica*, *Citrobacter freundii*, and *Citrobacter amalonaticus*). The patient's history revealed no apparent source, although past occupations included work in waste management and training in agriculture, both of which have increased exposure to antibiotic resistance genes.[9] Surprisingly, we detected multiple carbapenemases in *R. ornithinolytica* using PCR (*NDM*, *KPC*, *VIM*; *OXA* was only later identified using sequencing, see below). To identify all resistance genes in the isolates and any putative horizontal transfer between them, we performed real-time nanopore sequencing, both of the isolates individually and of the entire rectal swab, generating in total 3.9 M reads and 23.3 Gb on a standard ("MinION") flow cell.

All isolate genomes could be reconstructed with high accuracy (< 5 % redundancy, > 95 % completeness). Traditionally, nanopore sequencing has been associated with spurious insertions and deletions (indels).[10] However, we found that under the current iteration of the technology, the quality of genomes rivals that of genomes reconstructed from accurate short reads, as inferred from the "Watson distribution" of protein alignments (see methods and Figure S1). The *R. ornithinolytica* isolate carried nine plasmids and four carbapenemases: *NDM-1*, *KPC-2*, *VIM-1*, and *OXA-1* (Figure 1A). All carbapenemases were encoded on one plasmid each, except *VIM*, which was located on the bacterial chromosome.

The two *Citrobacter* isolates only carried *VIM-1*. An alignment of the genomic region 10 Kb upstream and downstream of *VIM* across the isolates revealed a transposase-mediated resistance transfer, for which we propose the following gene flow: The genomes of *C. freundii* and *C. amalonaticus* both carry *VIM-1* on an *IncHI2* plasmid (> 95 % sequence identity). In *C. freundii*, this transposon then likely copied itself into an *IncN* plasmid with the help of an *ISKpn19* transposase (Figure 1B). The same transposase is found flanking the *VIM* transposon in the *R. ornithinolytica* chromosome, which makes the *IncN* plasmid of *C. freundii* its likely source. A similar transfer pattern was observed for *OXA-1* (data not shown).

Isolate sequencing captured 79.7 % of resistance genes detected in the underlying microbial consortium through metagenomic sequencing (Figure 1C). Of the remainder, few genes were clinically relevant, such as several efflux pumps. Other resistance genes were associated with Gram-positive bacteria, which we did not screen for with culture (Figure 1C). Surprisingly, metagenomics did not detect five resistance gene types (6.8 %), including *KPC*, two out of three *OXA* copies, and two out of four *VIM* copies. This omission likely occurs because the metagenome was dominated by *Proteus vulgaris* (44.6 % of reads), leaving fewer reads (depth) for the carbapenemase-carrying strains (*C. freundii* 19.7 %, *R. ornithinolytica* 1.8 %, *C. amalonaticus* 0.01 %). Selective culture enriched these low-abundant species.

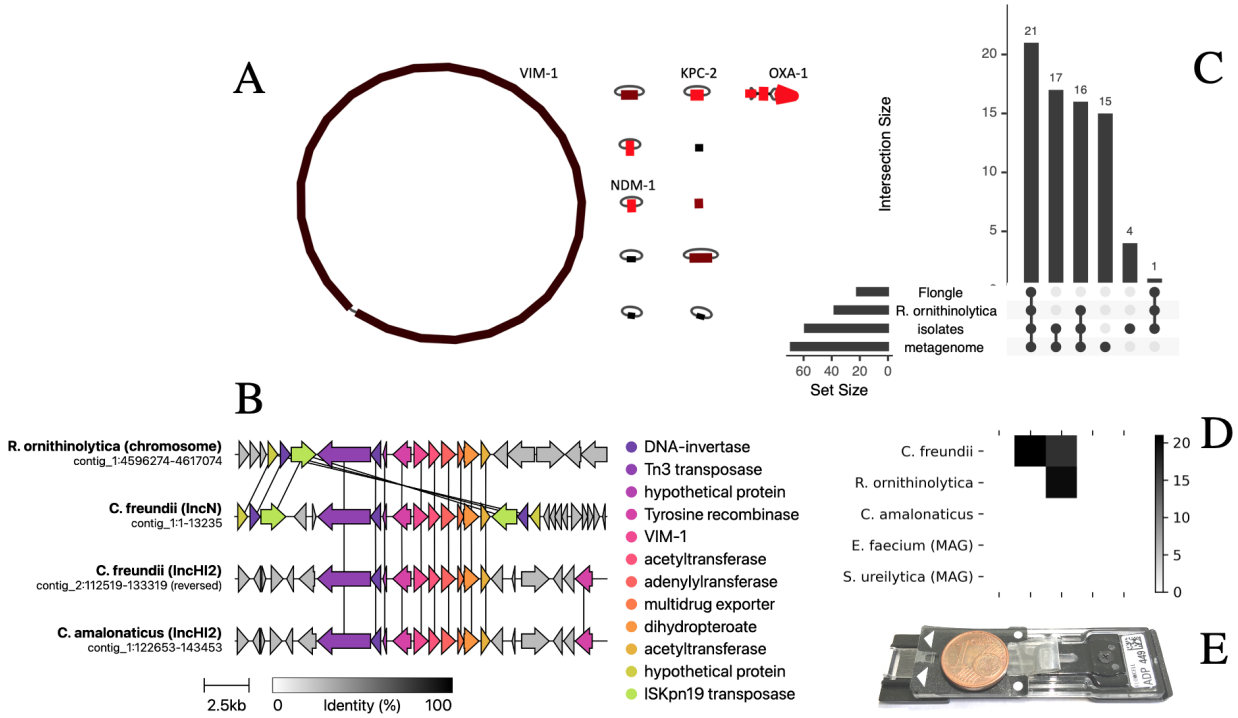

Figure 1: Real-time sequencing reveals extensive resistance load and horizontal gene transfer. **(A)** Genome reconstruction of a strain of *R. ornithinolytica* carrying nine plasmids and four carbapenemase genes. Color-coded coverage from 90x (black, e.g., chromosome) to 250x (red, e.g., plasmid carrying *OXA-1*). **(B)** Gene transfer of *VIM-1* across three strains and four loci. The carbapenemase is flanked by multiple transposases (see annotation), which likely mediate its mobilization. Vertical lines indicate 100 % sequence identity between corresponding genes. **(C)** Comparison of shared resistance genes between the enrichment sequencing run ("Flongle"), the *R. ornithinolytica* isolate, all four "isolates" combined and the "metagenome" assembly. Of all resistance genes identified in the metagenome, 79.7 % were found in the isolates. Surprisingly, several resistance genes were not identified in the metagenome, among them several carbapenemase copies. In the *R. ornithinolytica* isolate genome, about two-thirds of resistance genes were also found using on-device target enrichment. All plasmid-encoded genes among them were detected, including all carbapenemases. **(D)** Pairwise shared sequences between isolates and metagenome-assembled genomes. Putative transfers were defined as loci with a minimum length of one kilobase and 99.9 % sequence identity between each pair of loci. Extensive sequence transfer is observed between the three isolate genomes (and their corresponding bins from the metagenomic assembly). **(E)** Miniature, low-cost flow cell used for on-device target enrichment ("Flongle", Oxford Nanopore Technologies), with a one-cent coin placed on top as scale.

We also observed substantial horizontal gene transfer between our isolate members of the *Enterobacteriaceae* (Figure 1D). For example, *C. freundii* and *R. ornithinolytica* share 15 loci. A region was labeled as a putative transfer if its length exceeded one kilobase with 99.9 % sequence identity between any two genomes. No additional transfer was found in two uncultured, metagenome-assembled genomes (MAGs), namely *Enterococcus faecium* and *Serratia ureilytica*. None of the remaining metagenomic contigs showed putative transfers. Again, metagenomics did not add important information beyond the culture isolates.

The sensitivity of metagenomic sequencing can be increased with depth, but the associated cost limits the applicability in the routine laboratory. Therefore, going in the opposite direction, a recently introduced miniature nanopore flow cell ("Flongle") aims to reduce per-run costs through reduced sequencing yield. Because the yield is reduced, however, targeted sequencing of relevant genes or loci is desirable. Such target enrichment can be performed "on-device", i.e.,

during the sequencing run in real-time and without any changes in the sample preparation, using a method also known as "adaptive sequencing".[11–13] Here, reads are rejected from the pore when the read fragment that already passed through it does not match any sequence in a target database. The nanopore is then free to sequence another molecule.

We then used on-device target enrichment ("adaptive sequencing") in a multiplex resistance assay with 1,147 representative target genes (see methods). We generated about 5.4 Mb on the miniature flow cell within four hours from the carbapenemase-rich *R. ornithinolytica* isolate (Figure 1E). 97.2 % of reads were rejected; of those, 0.2 % (n=43) were false negative. Correspondingly, 2.8 % of reads were accepted, of which 20.4 % (n=104) were true positive, i.e., could be found in the target database. A positive database hit was defined as a read with at least 100 bp mapped to a target with a minimum of 50 % matching positions (Figure S2). 57.9 % of the resistance genes found in the high-quality genome reconstruction were found using adaptive sampling, too, including all four carbapenemases (Figure 1C). The probability of detection was determined by genomic location: All un-detected genes were located on the chromosome, and all plasmid-encoded resistance genes were detected (odds ratio 26.7, $p < 0.001$), likely because plasmids are present in higher copy numbers relative to the chromosome (Figure 1A). Since many resistance determinants are located on plasmids, we argue that enrichment sequencing is a promising approach for antimicrobial gene detection in routine settings. Compared to sequencing without enrichment, adaptive sequencing could increase the number of detected resistance gene copies by a factor of 2.2, thus more than doubling sensitivity at the observed sequencing yield.

Discussion

We detected a highly resistant consortium during hospital admission screening, including a strain that carried four carbapenemases. Real-time nanopore sequencing comprehensively characterized three resistant culture isolates within 48 hours, documenting many resistance genes as well as extensive gene transfer between isolates. This short turn-around time helped shape the public health response. For example, transposon-encoded VIM and OXA carbapenemases meant that associated wards could be monitored for the occurrence of these genes in other members of the *Enterobacteriaceae*.

Metagenomic sequencing of the corresponding rectal swap added little information and did not detect several important resistance genes. It might be that deeper sequencing would increase sensitivity, but because the carbapenemase-carrying strains were low abundant, in practice, this procedure would not be cost-competitive in a routine setting. Cultural screening as a first step reliably identified the strains that carried clinically relevant resistance genes.

We then performed on-device target enrichment of the most resistant culture isolate and were able to identify all plasmid-encoded resistance genes and nearly two-thirds of all resistance genes known to be present. The real-time search encompassed 1,147 representative genes in an ultra-high multiplex assay. As a proof of concept, we argue that this low-cost approach could be a valuable complement to routine microbiology and takes us closer to an effective point-of-care resistance screening, especially given the continued rapid improvements in the underlying technology.[14]

Methods

Culture and DNA extraction

All samples were streaked on carbapenemase chromogenic agar plates (CHROMagar, Paris, France). Carbapenemase carriage was confirmed using PCR and phenotypically using microdilution MIC testing. DNA was extracted from culture isolates and rectal swabs using the ZymoBIOMICS DNA Miniprep extraction kit according to the manufacturer's instructions. The cell disruption was conducted three times for five minutes with the Speedmill Plus (Analytik Jena, Germany).

Library preparation

Two sequencing runs were performed: One on a standard flow cell ("MinION"), multiplexing three culture isolates and a metagenomic sample, and the second on a miniature flow cell ("Flongle"). DNA quantification steps were performed using the dsDNA HS assay for Qubit (Invitrogen, US). DNA was size-selected by cleaning up with 0.45x volume of Ampure XP buffer (Beckman Coulter, Brea, CA, USA) and eluted in 60 l EB buffer (Qiagen, Hilden, Germany). The libraries were prepared from 1.5 g input DNA. For multiple samples we used the SQK-LSK109 kit (Oxford Nanopore Technologies, Oxford, UK) and the Native Barcoding Expansion-Kit (EXP-NBD104), according to the manufacturer's protocol. For the Flongle run we used the SQK-RBK004 kit from the same manufacturer.

Nanopore sequencing and on-device target enrichment

All DNA was sequenced on the GridION using a FLO-MIN106D (MinION) and FLO-FGL001 (Flongle) flow cell, respectively, (`MinKNOW` software v4.1.2), all from Oxford Nanopore Technologies. For on-device target enrichment, active channel selection was applied. As target database, we created a dereplicated version of the `CARD` database of resistance genes (v3.1.3)[15] using `mmseqs2 easy-cluster` (v13.45111)[16] using a minimum sequence identity of 0.95 and minimum coverage of 0.8 in coverage mode 1. We thereby reduced the database from 2,979 to 1,147 representative genes. We performed this step to reduce the search space that the adaptive sequencing algorithm has to map against. The reduction halves the database size because many resistance genes such as *CTX* have over one hundred documented isoforms, which would lead to uninformative multi-mappings. Reads were basecalled using the `guppy` GPU basecaller (high accuracy model, v4.2.2, Oxford Nanopore Technologies). For isolate genomes, reads were assigned to their respective barcodes only if matching adapters were detected on both ends of the read to avoid cross-contamination.

Data analysis

Isolate data were assembled using `flye` (v2.9)[17] and consensus sequences corrected using `racon` (v1.4.3)[18] and `medaka` (v1.4.3, github.com/nanoporetech/medaka). Read mapping was performed using `minimap2` (v2.22-r1101).[19] Genome quality was confirmed using `checkm` (v1.1.3).[20] To further assess the accuracy of the nanopore-only assemblies and especially the presence of indels in the absence of a ground truth short-read assembly, we looked at what we call "Watson distribution" (github.com/phiweger/ideel), after its inventor Mick Watson. It is a reference-free way to estimate the quality of a consensus genome reconstruction from error-prone long reads. The intuition is as follows: If a genome sequence contains many spurious indels that cause frameshifts in the coding sequence, it will manifest as an increased

number of pseudogenes. Bacterial genomes only carry few pseudogenes and rarely more than 200 because they are quickly lost from the genome.[21] Thus, we expect most proteins in the genome to be matchable about 100 % in length to a reference protein catalog, creating a distribution that peaks at 1. Resistance gene annotation was performed using `abricate` (v1.0.1, github.com/tseemann/abricate) against the `CARD` database (see above). Taxonomic assignments were performed using single-copy marker genes[22] as well as k-mers using sourmash (v4.2).[23] The metagenomic data were analyzed as described elsewhere.[24]

Availability

All sequencing data will be released in the final publication under NCBI project ID `PRJXXXXX`. Assemblies of isolate genomes and MAGs have been deposited with the Open Science Framework (osf.io) under project ID `wt7gc`.

References

[1] Harbarth, S. *et al.* Universal screening for methicillin-resistant staphylococcus aureus at hospital admission and nosocomial infection in surgical patients. *JAMA* **299**, 1149–1157 (2008).

[2] Xu, L., Sun, X. & Ma, X. Systematic review and meta-analysis of mortality of patients infected with carbapenem-resistant klebsiella pneumoniae. *Ann. Clin. Microbiol. Antimicrob.* **16**, 18 (2017).

[3] Armstrong, G. L. *et al.* Pathogen genomics in public health. *N. Engl. J. Med.* **381**, 2569–2580 (2019).

[4] Jousset, A. B. *et al.* A 4.5-year Within-Patient evolution of a Colistin-Resistant klebsiella pneumoniae Carbapenemase-Producing k. pneumoniae sequence type 258. *Clin. Infect. Dis.* **67**, 1388–1394 (2018).

[5] Hall, J. P. J., Wood, A. J., Harrison, E. & Brockhurst, M. A. Source-sink plasmid transfer dynamics maintain gene mobility in soil bacterial communities. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 8260–8265 (2016).

[6] León-Sampedro, R. *et al.* Pervasive transmission of a carbapenem resistance plasmid in the gut microbiota of hospitalized patients. *Nat Microbiol* **6**, 606–616 (2021).

[7] Carr, V. R. & Chaguza, C. Metagenomics for surveillance of respiratory pathogens. *Nat. Rev. Microbiol.* **19**, 285 (2021).

[8] de Siqueira, G. M. V., Pereira-dos Santos, F. M., Silva-Rocha, R. & Guazzaroni, M.-E. Nanopore sequencing provides rapid and reliable insight into microbial profiles of intensive care units. *Frontiers in Public Health* **9**, 1231 (2021).

[9] He, Y. *et al.* Antibiotic resistance genes from livestock waste: occurrence, dissemination, and treatment. *npj Clean Water* **3**, 1–11 (2020).

[10] Watson, M. & Warr, A. Errors in long-read assemblies can critically affect protein prediction. *Nat. Biotechnol.* **37**, 124–126 (2019).

[11] Loose, M., Malla, S. & Stout, M. Real-time selective sequencing using nanopore technology. *Nat. Methods* **13**, 751–754 (2016).

[12] Kovaka, S., Fan, Y., Ni, B., Timp, W. & Schatz, M. C. Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. *Nat. Biotechnol.* **39**, 431–441 (2021).

[13] Payne, A. *et al.* Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nat. Biotechnol.* **39**, 442–450 (2020).

[14] Koren, S., Phillippy, A. M., Simpson, J. T., Loman, N. J. & Loose, M. Reply to 'errors in long-read assemblies can critically affect protein prediction'. *Nat. Biotechnol.* **37**, 127–128 (2019).

[15] Jia, B. *et al.* CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **45**, D566–D573 (2017).

[16] Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

[17] Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).

[18] Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).

[19] Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

[20] Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).

[21] Thomson, N. R. *et al.* Comparative genome analysis of salmonella enteritidis PT4 and salmonella gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. *Genome Res.* **18**, 1624–1637 (2008).

[22] Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* **36**, 1925–1927 (2019).

[23] Pierce, N. T., Irber, L., Reiter, T., Brooks, P. & Brown, C. T. Large-scale sequence comparisons with *sourmash*. *F1000Res.* **8**, 1006 (2019).

[24] Van Damme, R. *et al.* Metagenomics workflow for hybrid assembly, differential coverage binning, metatranscriptomics and pathway analysis (MUFFIN). *PLoS Comput. Biol.* **17**, e1008716 (2021).

Supplement

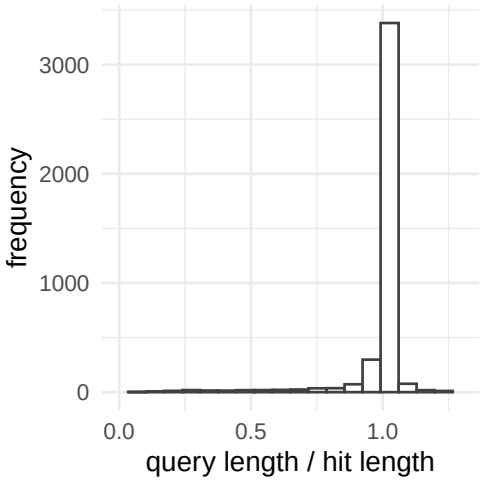

Figure S1: "Watson distribution" of *R. ornithinolytica* genome reconstruction. This heuristic is a reference-free way to estimate the quality of a consensus genome reconstruction from error-prone long reads (see methods). For example, a peak at 1 indicates a high-quality reconstruction without substantial spurious insertions and deletions, a typical error in earlier iterations of the Nanopore technology.

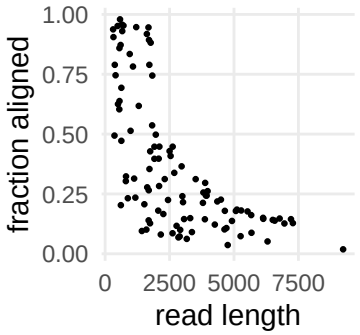

Figure S2: Mapping profile of accepted reads from adaptive sequencing (target enrichment). A positive database hit was defined as a read with at least 100 bp mapped to a target with a minimum of 50 % matching positions.