TETyper: a bioinformatic pipeline for classifying variation and genetic contexts of transposable elements from short-read whole-genome sequencing data

Anna E Sheppard1,2,*, Nicole Stoesser1, Ian German-Mesner3, Kasi Vegesana3, A Sarah Walker1, Derrick W Crook1, Amy J Mathers4,5

1 Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom
2 National Institute for Health Research Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, University of Oxford, Oxford, United Kingdom
3 Health Information & Technology, University of Virginia Health System, Charlottesville, Virginia, USA
4 Division of Infectious Disease and International Health, Department of Medicine, University of Virginia Health System, Charlottesville, Virginia, USA
5 Clinical Microbiology Laboratory, Department of Pathology, University of Virginia Health System, Charlottesville, Virginia, USA

* Corresponding author: anna.sheppard@ndm.ox.ac.uk

ABSTRACT

Much of the worldwide dissemination of antibiotic resistance has been driven by resistance gene associations with mobile genetic elements (MGEs), such as plasmids and transposons. Although increasing, our understanding of resistance spread remains relatively limited, as methods for tracking mobile resistance genes through multiple species, strains and plasmids are lacking. We have developed a bioinformatic pipeline for tracking variation within, and mobility of, specific transposable elements (TEs), such as transposons carrying antibiotic resistance genes. TETyper takes short-read whole-genome sequencing data as input and identifies single-nucleotide mutations and deletions within the TE of interest, to enable tracking of specific sequence variants, as well as the surrounding genetic context(s), to enable identification of transposition events. To investigate global dissemination of Klebsiella pneumoniae carbapenemase (KPC) and its associated transposon Tn4401, we applied TETyper to a collection of >3000 publicly available Illumina datasets containing blaKPC. This revealed surprising diversity, with >200 distinct flanking genetic contexts for Tn4401, indicating high levels of transposition. Integration of sample metadata revealed insights into associations...
between geographic locations, host species, Tn4401 sequence variants and flanking genetic contexts. To demonstrate the ability of TETyper to cope with high copy number TEs and to track specific short-term evolutionary changes, we also applied it to the insertion sequence IS26 within a defined *K. pneumoniae* outbreak. TETyper is implemented in python and is freely available at https://github.com/aesheppard/TETyper.

---

**IMPACT STATEMENT**

Whole-genome sequencing (WGS) of bacterial pathogens has revolutionised the analysis of global and within-outbreak transmission pathways. However, the study of antibiotic resistance dissemination is more challenging, as resistance genes are often associated with mobile genetic elements (MGEs) that enable gene exchange between different host bacteria. Therefore, standard WGS approaches that focus on host strain relationships may not be informative for understanding resistance gene dissemination. We have developed a bioinformatic tool for analysing WGS data from the perspective of a specific MGE-resistance gene association. The outputs produced identify variation within the MGE, as well as signatures of MGE mobility. This information can then be used to track the movement of the resistance gene, thus overcoming previous limitations by defining relationships from a resistance gene perspective, rather than a host-strain perspective. In an epidemiological context, this can provide insight into specific transmission pathways, thus informing infection control within outbreak scenarios, as well as increasing our understanding of global pathways of resistance dissemination.

---

**INTRODUCTION**

Increasing antibiotic resistance in a range of bacterial pathogens is a major global health threat, but our understanding of resistance gene dissemination remains incomplete. Many resistance genes are carried on mobile genetic elements (MGEs), enabling bacteria to evolve in response to antimicrobial pressures via gene exchange. MGEs of relevance include plasmids, which are extrachromosomal, usually circular DNA structures that can be transferred between different host bacteria, and transposable elements (TEs), which are short stretches of DNA, often carried on plasmids, that can autonomously mobilise to new genomic locations via transposition (1-3). TEs comprise transposons, which carry additional cargo genes, such as antibiotic resistance genes, and insertion sequences (ISs), which comprise only elements necessary for transposition; however, composite structures involving ISs can also be involved in resistance gene mobilisation (4, 5).
Whole-genome sequencing (WGS) has revolutionised the analysis of pathogen transmission by enabling high-resolution insight into chromosomal relatedness (6-9). However, resistance gene dissemination via MGEs is more complicated because horizontal transfer disrupts pairing between resistance genes and host strains. To assess relatedness from the perspective of a mobile resistance gene, it is necessary to examine the gene’s genetic context. However, the most widely used WGS technologies (e.g. Illumina) produce short sequencing reads; these result in fragmented assemblies, with resistance genes often present on very short contigs due to associations with repetitive elements such as TEs. This makes tracking the associated plasmids largely impractical, as assembling complete plasmid sequences from short reads is problematic (10), and reference-based approaches can be unreliable due to transposition or homologous recombination disrupting pairing between host plasmids and resistance genes (11). Alternatively, if a resistance gene has a stable association with a specific transposon, then tracking the transposon may be a better proxy for understanding resistance gene dissemination.

One example of such an association is the *Klebsiella pneumoniae* carbapenemase (*bla*KPC) and its associated replicative transposon Tn4401. *bla*KPC was first identified in the USA in 1996, and has since spread globally, being responsible for a large proportion of carbapenem-resistant Enterobacteriaceae infections worldwide (12-15). Initially, it was largely associated with *K. pneumoniae* multi-locus sequence type (ST) 258 and the IncFII plasmid pKpQIL, but it has since spread to various other plasmids, other *K. pneumoniae* strains, other species of Enterobacteriaceae, and occasionally non-Enterobacteriaceae (11, 16-18). Given the importance of Tn4401 transposition in facilitating this spread, the ability to track transposition events and sequence variation within Tn4401 may be helpful for better understanding *bla*KPC dissemination.

A similar tracking approach may also be useful for investigating the evolution of other TEs of interest. For example, replicative intermolecular transposition of a variety of TEs (including Tn4401) involves target site duplication (TSD), which results in a short (~2-14 bp) direct repeat flanking the newly transposed copy (3). Intramolecular transposition disrupts these flanking repeats, and investigation of TSD sequences can be used to gain insight into historical transposition events, as has been demonstrated for the widely dispersed IS26 (19).

In order to facilitate TE tracking, we developed TETyper, a bioinformatic pipeline for classifying TE variation from short-read WGS data. TETyper takes raw sequencing reads as input, and identifies: 1. Structural variation within a specific TE of interest, 2. Single nucleotide variants (SNVs) within the TE, and 3. flanking genetic context(s) of the TE. Variation within the TE captures signatures of micro-evolution, while the flanking sequences capture signatures of transposition, as every transposition event introduces a new genetic sequence context for the TE. This information can then be utilised for investigating transposition pathways, as well as gaining epidemiological insight in the context of resistance gene dissemination, both within outbreaks and globally. We demonstrate the utility of TETyper by applying it to a collection of >3000 publicly available Illumina datasets containing *bla*KPC and to IS26 within a clonal *K. pneumoniae* outbreak.
THEORY AND IMPLEMENTATION

Description of the TETyper pipeline:

An overview of processing steps is shown in Fig. 1. Firstly, reads are mapped against a reference representing the TE of interest using bwa mem (20). For the remaining steps, only mapped reads are retained.

To classify structural variation within the TE, we focus on deletions relative to the reference, since insertions and other rearrangements are difficult to classify reliably using short-read data. This is achieved by assembling the reads that map to the TE using spades (21), followed by BLASTn (22) to identify missing regions.

To identify SNVs, variant calling is performed under a diploid model using samtools mpileup (23), with variants excluded if they fall within a deleted region as identified above, or if they are not supported by at least one read in each direction. Heterozygous variants are assumed to represent within-sample mixtures (i.e. multiple, slightly different, copies of the TE), and are reported along with the number of reads supporting each nucleotide.

To identify flanking genetic context(s) of the TE, the user specifies the desired length of flanking sequence to classify, which should be short enough that sequencing errors are rare. For the applications below, we use a length equal to TSD length (5 bp for Tn4401 and 8 bp for IS26). Reads mapping to the start/end of the TE sequence are examined, and the sequence of each read immediately adjacent to the start/end of the TE reference, of the length specified above, is extracted. Several filters are used to remove low-quality sequences, and all distinct sequences that pass quality filters are output, along with the number of reads supporting each. In this way, if there is a single copy of the TE present in the sample, there should be a single flanking sequence identified at each end of the TE. On the other hand, every time a TE undergoes transposition, it inserts into a new genetic context. Therefore, if there are X copies of a TE in a sample, then there will be up to X distinct flanking sequences at each end of the TE.

Specific parameters used for running TETyper are described in Supplementary Methods. The TETyper output was validated using a subset of isolates from the datasets described below for which complete, closed references were previously generated using long-read sequence data (11, 24-26) (Supplementary Results).
Application to Tn4401:

To demonstrate the utility of TETyper for epidemiological investigations, we applied it to 3054 bla\textsubscript{KPC} positive Illumina datasets retrieved from a December 2016 snapshot of the European Nucleotide Archive (27) (Supplementary Methods, Supplementary Table 1).

Structural variation in Tn4401:

There were eight “common” (found in ≥10 samples) structural variants of Tn4401 (Fig. 2a). The ancestral Tn4401b structure (28) was found in 850/3054 (28%) samples. Four variants represented different deletions immediately upstream of bla\textsubscript{KPC}, namely Tn4401d (29, 30), Tn4401a (28), Tn4401h (31) and Tn4401e (29, 30) in 937 (31%), 868 (28%), 40 (1.3%) and 19 (0.6%) samples respectively. The other three variants all represented truncations of Tn4401.

Deletions in the promoter region upstream of bla\textsubscript{KPC} have been shown to result in increased expression (30, 31), which is expected to be advantageous under antibiotic selection pressure. As several of these were observed, with no other common internal deletions across the 10 kb Tn4401 sequence, this suggests that much of the structural variation observed may be due to selection rather than random genetic drift. Truncation of Tn4401 presumably prevents further transposition; one possible reason for the abundance of truncation variants is that they bring other TEs into the vicinity of bla\textsubscript{KPC} (32), thus providing alternative routes for gene mobilisation.

Specific structural variants were generally found in multiple host species, indicating wide horizontal dissemination via inter-species transfer (Fig. 3a). Tn4401b was the most widely disseminated, being found in 10 different genera, while Tn4401a was relatively restricted to \textit{K. pneumoniae} (98%). Several different structural variants were present in USA samples, while other countries generally had a single predominant variant (Fig. 3b), supporting the origination and diversification of bla\textsubscript{KPC} and Tn4401 in Enterobacteriaceae in the USA. However, the dataset was heavily biased towards USA isolates, and for 852/3054 (28%) samples the country of origin was unknown, highlighting limitations in metadata availability.

Single nucleotide variation in Tn4401:

Most SNV variation involved sites within the bla\textsubscript{KPC} gene (Fig. 2b), again implicating selection for KPC function in explaining observed variation. The two most common variants carried bla\textsubscript{KPC-2} and bla\textsubscript{KPC-3}, in 1040/3054 (34%) and 1719 (56%) samples respectively, with each found in several different structural backgrounds (Fig. 3c).

Interestingly, 12 samples showed polymorphism at the site that differentiates bla\textsubscript{KPC-2} and bla\textsubscript{KPC-3} (Fig. 2b), signifying a mixture of both alleles and indicating that these samples most likely contain two copies of Tn4401, one with bla\textsubscript{KPC-2} and one with bla\textsubscript{KPC-3}. Minor allele percentages ranged from 10-49%. These occurred in several Tn4401 structures, including Tn4401a, Tn4401b and Tn4401d, and several host species, including \textit{Escherichia coli}, \textit{K. pneumoniae}, \textit{K. oxytoca} and \textit{Enterobacter cloacae}.

This indicates that the presence of multiple bla\textsubscript{KPC} variants may be a general phenomenon,
suggesting repeated multiple acquisition of $\text{bla}_{KPC}$ and/or repeated mutation converting $\text{bla}_{KPC2}$ to $\text{bla}_{KPC3}$ (or vice versa).

Flanking genetic contexts of Tn4401:

The most common 5 bp sequences flanking Tn4401 were GTTCT/TCTCT, ATTGA/ATTGA and GTTCT/GTTCT, present in 836/3054 (27%), 718 (24%) and 389 (13%) samples respectively (Fig. 3d-f). ATTGA/ATTGA corresponds to the epidemic IncFII pKpQIL plasmid; these samples were almost exclusively Tn4401a-containing $K.\ pneumoniae$, but from a variety of geographic locations. GTTCT/GTTCT corresponds to Tn1/2/3-like elements (including Tn1331; see below), which have been described containing Tn4401 in many different plasmid backbones (11); these samples represented a wider variety of Tn4401 structures and host species. GTTCT/TTTCT is consistent with IncFIA pBK30661/pBK30683-like plasmids, where Tn4401 is adjacent to a partial Tn1331 element on the left side only, presumably as a result of deletion on the right side following initial integration (33).

For 386/3054 (13%) samples, there was no flanking sequence identified on one or both sides of Tn4401, consistent with Tn4401 truncation. These truncated Tn4401 structures are presumably not transpositionally active, so TETyper would be unable to capture ongoing mobilisation events in these cases. For the vast majority, however, Tn4401 appeared to be intact; 2379 (78%) samples had a single flanking sequence identified on each side of Tn4401, indicating a single intact copy, and 289 (9%) had multiple flanking sequences on one or both sides, suggesting multiple copies.

Of those with a single copy, 1418/2379 (60%) had the same 5 bp sequence at both ends of Tn4401, consistent with TSD following standard intermolecular transposition. Surprisingly, 961/2379 (40%) showed different 5 bp sequences, indicating disruption of TSD signatures and suggesting that intramolecular transposition of Tn4401 may be relatively common.

Altogether, there were 193 and 214 distinct 5 bp sequences flanking the left and right sides of Tn4401 respectively, and a total of 273 distinct flanking sequence profiles, suggesting relatively frequent transposition. This diversity indicates that the classification of flanking genetic contexts in this way may be useful for epidemiological tracking by providing higher genetic resolution than strain typing alone.

Application to IS26:

To demonstrate the utility of TETyper for analysing specific TE mobility events within well-defined sampling frames, we applied it to IS26 for 34 closely related $K.\ pneumoniae$ ST15 isolates from an NDM-1 outbreak in Nepal (26). These isolates varied in the number and sequence of genetic contexts of IS26, with evidence for 4-14 copies per isolate (Fig. 4). In some cases, the TETyper output provided higher genetic resolution than a standard phylogenetic approach, with IS26 flanking sequence profiles differing between pairs of isolates with 0 chromosomal SNVs (Fig. 4; PMK21b vs PMK18/21a/21d and PMK24 vs PMK22/25).
CONCLUSION

We have developed a novel bioinformatic pipeline, TETyper, for classifying sequence variation and flanking genetic contexts of TEs from short-read WGS data. We demonstrate the utility of TETyper by applying it to Tn4401 for a large, global $bla_{KPC}$ collection, as well as IS26 within a small, defined outbreak. This revealed surprising diversity in both cases, and provided insights into patterns of transposition and mutational change within Tn4401. In an epidemiological context, the within-TE variation and transposition signatures identified by TETyper could be used to facilitate higher resolution resistance gene tracking related to gene mobility than is currently possible using other WGS-based methods.

AUTHOR STATEMENTS

Funding information

The research was funded by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at University of Oxford in partnership with Public Health England (PHE), in collaboration with University of Virginia, and by a contract from the US Centers for Disease Control and Prevention (CDC) Broad Agency Announcement (BAA 2016-N-17812). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health, Public Health England or the CDC. NS is supported by a Public Health England/University of Oxford Clinical Lectureship.

Acknowledgements

We thank Zamin Iqbal and Phelim Bradley for assistance with BIGSI. We also thank the HPRU Steering Group.

Ethical statement

No experiments involving humans or animals were performed for this study.

Conflicts of interest

The authors declare that there are no conflicts of interest.
ABBREVIATIONS

MGE – mobile genetic element
TE – transposable element
IS – insertion sequence
WGS – whole-genome sequencing
KPC – Klebsiella pneumoniae carbapenemase
TSD – target site duplication
SNV – single nucleotide variant

REFERENCES

1. Carattoli A. Plasmids and the spread of resistance. Int J Med Microbiol. 2013;303(6-7):298-304.
2. Siguier P, Gourbeyre E, Varani A, Ton-Hoang B, Chandler M. Everyman’s Guide to Bacterial Insertion Sequences. Microbiol Spectr. 2015;3(2):MDNA3-0030-2014.
3. Partridge SR. Resistance mechanisms in Enterobacteriaceae. Pathology. 2015;47(3):276-84.
4. Harmer CJ, Moran RA, Hall RM. Movement of IS26-associated antibiotic resistance genes occurs via a translocatable unit that includes a single IS26 and preferentially inserts adjacent to another IS26. MBio. 2014;5(5):e01801-14.
5. Poirel L, Lartigue MF, Decousser JW, Nordmann P. ISEcp1B-mediated transposition of blaCTX-M in Escherichia coli. Antimicrob Agents Chemother. 2005;49(1):447-50.
6. Walker TM, Lalor MK, Broda A, Saldana Ortega L, Morgan M, Parker L, et al. Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an observational study. Lancet Respir Med. 2014;2(4):285-92.
7. Eyre DW, Cule ML, Wilson DJ, Griffiths D, Vaughan A, O’Connor L, et al. Diverse sources of C. difficile infection identified on whole-genome sequencing. N Engl J Med. 2013;369(13):1195-205.
8. Senn L, Clerc O, Zanetti G, Basset P, Prod’hom G, Gordon NC, et al. The Stealthy Superbug: the Role of Asymptomatic Enteric Carriage in Maintaining a Long-Term Hospital Outbreak of ST228 Methicillin-Resistant Staphylococcus aureus. MBio. 2016;7(1):e02039-15.
9. Kwong JC, Lane CR, Romanes F, Gonçalves da Silva A, Easton M, Cronin K, et al. Translating genomics into practice for real-time surveillance and response to carbapenemase-producing Enterobacteriaceae: evidence from a complex multi-institutional KPC outbreak. PeerJ. 2018;6:e4210.

10. Arredondo-Alonso S, Willems RJ, van Schaik W, Schürch AC. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. Microb Genom. 2017;3(10):e000128.

11. Sheppard AE, Stoesser N, Wilson DJ, Sebra R, Kasarskis A, Anson LW, et al. Nested Russian Doll-Like Genetic Mobility Drives Rapid Dissemination of the Carbapenem Resistance Gene $\text{bla}_{\text{KPC}}$. Antimicrob Agents Chemother. 2016;60(6):3767-78.

12. Yigit H, Queenan AM, Anderson GJ, Domenech-Sanchez A, Biddle JW, Steward CD, et al. Novel carbapenem-hydrolyzing beta-lactamase, KPC-1, from a carbapenem-resistant strain of Klebsiella pneumoniae. Antimicrob Agents Chemother. 2001;45(4):1151-61.

13. Munoz-Price LS, Poirel L, Bonomo RA, Schwaber MJ, Daikos GL, Cormican M, et al. Clinical epidemiology of the global expansion of Klebsiella pneumoniae carbapenemases. Lancet Infect Dis. 2013;13(9):785-96.

14. Lee CR, Lee JH, Park KS, Kim YB, Jeong BC, Lee SH. Global Dissemination of Carbapenem-Producing Klebsiella pneumoniae: Epidemiology, Genetic Context, Treatment Options, and Detection Methods. Front Microbiol. 2016;7:895.

15. Bonomo RA, Burd EM, Conly J, Limbago BM, Poirel L, Segre JA, et al. Carbapenemase-Producing Organisms: A Global Scourge! Clin Infect Dis. 2017.

16. Mathers AJ, Peirano G, Pitout JD. The role of epidemic resistance plasmids and international high-risk clones in the spread of multidrug-resistant Enterobacteriaceae. Clin Microbiol Rev. 2015;28(3):565-91.

17. Zhang Y, Wang Q, Yin Y, Chen H, Jin L, Gu B, et al. Epidemiology of Carbapenem-Resistant Enterobacteriaceae Infections: Report from the China CRE Network. Antimicrob Agents Chemother. 2018;62(2).

18. Correa A, Del Campo R, Perenguez M, Blanco VM, Rodríguez-Baños M, Perez F, et al. Dissemination of high-risk clones of extensively drug-resistant Pseudomonas aeruginosa in colombia. Antimicrob Agents Chemother. 2015;59(4):2421-5.

19. He S, Hickman AB, Varani AM, Siguier P, Chandler M, Dekker JP, et al. Insertion Sequence IS26 Reorganizes Plasmids in Clinically Isolated Multidrug-Resistant Bacteria by Replicative Transposition. MBio. 2015;6(3):e00762.

20. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013:arXiv:1303.3997v2 [q-bio.GN].

21. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455-77.

22. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
23. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987-93.

24. Mathers AJ, Stoesser N, Chai W, Carroll J, Barry K, Cheruvanky A, et al. Chromosomal Integration of the Klebsiella pneumoniae Carbapenemase Gene, bla\textsubscript{KPC}, in Klebsiella Species Is Elusive but Not Rare. Antimicrob Agents Chemother. 2017;61(3).

25. Sheppard AE, Stoesser N, Sebra R, Kasarskis A, Deikus G, Anson L, et al. Complete Genome Sequence of KPC-Producing Klebsiella pneumoniae Strain CAV1193. Genome Announc. 2016;4(1).

26. Stoesser N, Giess A, Batty EM, Sheppard AE, Walker AS, Wilson DJ, 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. Antimicrob Agents Chemother. 2014;58(12):7347-57.

27. Bradley P, den Bakker H, Rocha E, McVean G, Iqbal Z. Real-time search of all bacterial and viral genomic data. bioRxiv. 2017.

28. Naas T, Cuzon G, Villegas MV, Lartigue MF, Quinn JP, Nordmann P. Genetic structures at the origin of acquisition of the beta-lactamase bla\textsubscript{KPC} gene. Antimicrob Agents Chemother. 2008;52(4):1257-63.

29. Kitchel B, Rasheed JK, Endimiani A, Hujer AM, Anderson KF, Bonomo RA, et al. Genetic factors associated with elevated carbapenem resistance in KPC-producing Klebsiella pneumoniae. Antimicrob Agents Chemother. 2010;54(10):4201-7.

30. Naas T, Cuzon G, Truong HV, Nordmann P. Role of IS\textsubscript{Kpn7} and deletions in bla\textsubscript{KPC} gene expression. Antimicrob Agents Chemother. 2012;56(9):4753-9.

31. Cheruvanky A, Stoesser N, Sheppard AE, Crook DW, Hoffman PS, Weddle E, et al. Enhanced Klebsiella pneumoniae Carbapenemase Expression from a Novel Tn\textsubscript{4401} Deletion. Antimicrob Agents Chemother. 2017;61(6).

32. Stoesser N, Sheppard AE, Peirano G, Anson LW, Pankhurst L, Sebra R, et al. Genomic epidemiology of global Klebsiella pneumoniae carbapenemase (KPC)-producing Escherichia coli. Sci Rep. 2017;7(1):5917.

33. Chen L, Chavda KD, Melano RG, Hong T, Rojtman AD, Jacobs MR, et al. Molecular survey of the dissemination of two bla\textsubscript{KPC}-harboring IncFIA plasmids in New Jersey and New York hospitals. Antimicrob Agents Chemother. 2014;58(4):2289-94.

34. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268-74.
FIGURES AND TABLES

Fig. 1: Overview of processing steps in the TETyper pipeline.

Fig. 2: Structure of Tn4401 showing common structural variants (a) and SNVs (b). Only variants found in at least 10 samples are shown. For simplicity, SNV variants are all illustrated within a Tn4401b structural background.

Fig. 3: Associations between Tn4401 structural variants, SNVs, flanking genetic contexts, host species, and countries of origin amongst a collection of 3054 blaKPC samples from the European Nucleotide Archive. (a) Distributions of host species for different structural variants of Tn4401. (b) Distributions of Tn4401 structural variants for different countries of origin. (c) Distributions of Tn4401 structural variants for different Tn4401 SNVs. (d-f) Distributions of Tn4401 structural variants (d), host species (e) and countries of origin (f) for different Tn4401 flanking genetic contexts.

Fig. 4: Variation in IS26 amongst 34 ST15 K. pneumoniae isolates from an NDM-1 outbreak. TETyper output is annotated alongside a maximum likelihood phylogeny that was generated using IQ-TREE version 1.3.13 (34), after mapping to the MGH78578 reference as previously described (26). Branch lengths are shown as SNVs per genome.

Supplementary Table 1: TETyper output and sample metadata for blaKPC datasets.

Supplementary Table 2: Validation of TETyper output for 24 blaKPC isolates with complete long-read assemblies available.

Supplementary Table 3: IS26 elements present in PMK1 long-read assembly.
Illumina reads

Map to TE reference

Mapped reads

De novo assembly

BLASTn against TE reference

TE structural variation (deletions)

Variant calling

Homozygous SNVs

Heterozygous SNVs

Extract reads overlapping TE ends

Filter out low-quality sequences

TE flanking genetic contexts
**a**

Samples (%)

|   | Samples (%) |
|---|-------------|
| Tn4401b | 850 (28%) |
| Tn4401d | 937 (31%) |
| Tn4401a | 868 (28%) |
| 7128-9197 | 146 (5%) |
| Tn4401h | 40 (1%) |
| Tn4401e | 19 (0.6%) |
| 1-9767 | 17 (0.6%) |
| 4728-10006 | 13 (0.4%) |

**b**

|   | Samples (%) |
|---|-------------|
| 1 | 1040 (34%) |
| C8015T | 1719 (56%) |
| C7509G T7919G | 52 (2%) |
| C126T | 23 (0.8%) |
| A6687G | 16 (0.5%) |
| C4202T | 15 (0.5%) |
| C8015Y | 12 (0.4%) |
