Role of the mobilome in the global dissemination of the carbapenem resistance gene *bla*$_{\text{NDM}}$

Mislav Acman[1*], Ruobing Wang[2], Lucy van Dorp[1], Liam P. Shaw[3], Qi Wang[2], Nina Luhmann[4], Yuyao Yin[2], Shijun Sun[2], Hongbin Chen[2], Hui Wang[2], Francois Balloux[1]

1 UCL Genetics Institute, University College London, Gower Street, London, WC1E 6BT, UK

2 Department of Clinical Laboratory, Peking University People's Hospital, Beijing, 100044, China

3 Department of Zoology, University of Oxford, Oxford OX1 3SZ, UK

4 Warwick Medical School, University of Warwick, Coventry CV4 7AL, UK

* Corresponding Author

E-mail: mislav.acman.17@ucl.ac.uk

Abstract (240 words)

The mobile resistance gene *bla*NDM encodes the NDM enzyme which hydrolyses carbapenems, a class of antibiotics used to treat some of the most severe bacterial infections. *bla*NDM is globally distributed across a variety of Gram-negative bacteria on multiple plasmids, typically located within a highly recombining and transposon-rich genomic region. This complexity means the dynamics underlying the dissemination of *bla*NDM remain poorly resolved. In this work, we compile a dataset of over 6000 bacterial genomes harbouring the *bla*NDM gene including 104 newly generated PacBio hybrid assemblies from clinical and livestock associated isolates across China. We develop a novel computational approach to track structural variants surrounding *bla*NDM in bacterial genomes. This allows us to identify the prevalent genomic contexts of *bla*NDM and reconstruct the key mobile genetic elements and events in its global spread. We estimate that *bla*NDM emerged on a Tn*125* transposon before 1985 but only reached a global prevalence around a decade after its first recorded observation in 2005. We find that the Tn*125* transposon played an important role in early plasmid-mediated jumps of *bla*NDM but was overtaken by other elements in recent years including IS26-flanked pseudo-composite transposons and Tn*3000*. Lastly, we observe a notable correlation between plasmid backbones bearing *bla*NDM and the sampling location of isolates. This observation suggests that the dissemination of resistance genes is mainly driven by successive between-plasmid transposon jumps, with plasmid exchange much more restricted due to the adaptation of plasmids to specific bacterial hosts.

Introduction

Antimicrobial resistance (AMR) poses a major challenge to human and veterinary health. AMR can be conferred by vertically inherited point mutations or via the acquisition of horizontally transmitted 'accessory' genes, often located in transposons and plasmids. The *bla*NDM gene represents a typical example of a mobile AMR gene[1]. *bla*NDM encodes a metallo-β-lactamase capable of hydrolysing most β-lactam antibiotics. These antibiotics are used as a first-line treatment for severe infections and to treat multidrug-resistant Gram-negative bacterial infections. As such, the global prevalence of bacteria carrying *bla*NDM represents a major public health concern.

*bla*NDM was first described in 2008 from a *Klebsiella pneumoniae* isolated from a urinary tract infection in a Swedish patient returning from New Delhi, India[2]. While *bla*NDM now has a worldwide distribution, most of the earliest cases have been linked to the Indian subcontinent, leading to this region being suggested as the likely location for the initial mobilisation event[1,3–6]. NDM-positive *Acinetobacter baumannii* isolates have been retrospectively identified from an Indian hospital in 2005[7], which remain the earliest observations to date. Nevertheless, an NDM-positive *A. pittii* isolate was also isolated in 2006 from a Turkish patient with no history of travel outside Turkey[8].

Although no complete genome sequences are publicly available from these earliest observations, the first NDM-positive isolates from 2005 were shown to carry *bla*NDM on multiple non-conjugative, but potentially mobilizable plasmid backbones[7]. In addition, *bla*NDM in these early isolates was positioned within a complete Tn*125* transposon with IS*26* insertion sequences (ISs) as well as ISCR*27* (IS-containing common region 27), suggesting the possibility of complex patterns of mobility since the gene's initial integration. Subsequent NDM-positive isolates across multiple species consistently harbour either a complete or fragmented IS*Aba125* (an IS constituting Tn*125*), immediately upstream of *bla*NDM, which provides a promoter region[1,5,9,10]. The presence of IS*Aba125* in some form in all NDM-positive isolates to date and the early observations in *A. baumannii* have led to Tn*125* being proposed as the ancestral transposon responsible for the mobilization of *bla*NDM, and *A. baumannii* as the ancestral host[10,11].

The NDM enzyme itself is of possible chimeric origin[10,12], with the first six amino acids in NDM matching to those in *aphA6*, a gene providing aminoglycoside resistance and also flanked by IS*Aba125*. It is hypothesised that ISCR*27*, which uses a rolling-circle (RC) transposition mechanism[13,14], initially mobilized a progenitor of *bla*NDM in *Xanthomonas sp.* and placed it downstream of IS*Aba125*[10,12,15,16]. At least 29 distinct sequence variants of the NDM enzyme have been described to date[1,17]. The most prevalent of these variants is the first to have been characterised, denoted NDM-1[18]. Different NDM variants are mostly distinguished by a single amino-acid substitution, apart from NDM-18 which carries a tandem repeat of five amino acids. None of the observed substitutions occur in the active site and their functional impact remains under debate[1].

*bla*NDM is found in at least 11 bacterial families and NDM-positive isolates have heterogeneous clonal backgrounds, supporting multiple independent acquisitions of *bla*NDM[1]. Although *bla*NDM has been observed on bacterial chromosomes[19,20] it is most commonly found on plasmids, comprising multiple different backbones or types. Thus far, *bla*NDM has been associated to at least 20 different plasmid types, predominantly IncFIB, IncFII, IncA/C (IncC), IncX3, IncH, and IncL/M, and also in untyped plasmids[1,4,21–24]. Furthermore, even within the same plasmid type, *bla*$_{\mathrm{NDM}}$ can be found in a variety of genomic contexts, often interspersed by multiple ISs and composite transposons[1,12]. The immediate environment of *bla*$_{\mathrm{NDM}}$ has been reported to vary even in isolates from the same patient[23]. Many mobile elements are thought to play important roles in dissemination, including IS*Aba125*, IS*3000*, IS*26*, IS*5*, ISCR1, Tn*3*, Tn*125*, and Tn*3000*[1,23,25–28]. It is therefore clear that the spread of *bla*$_{\mathrm{NDM}}$ was, and is, a multi-layer process involving multiple mobile genetic elements – 'the mobilome'. *bla*$_{\mathrm{NDM}}$ mobility involves diverse processes, including genetic recombination[29,30], transposition, conjugation and transformation of plasmids[31], transduction[32], and transfer through outer-membrane vesicles (OMVs)[33,34].

Previous surveys of *bla*$_{\mathrm{NDM}}$-positive genomes have led to a better understanding of its evolution[1]. However, a major difficulty, as for other AMR genes, is relating the diverse genomic contexts to temporal evolution. Here, we outline an alignment-based method to identify flanking structural variants and use it to build a history of the insertion and mobilization events. We compile a global dataset of more than 6000 NDM-positive isolates. In line with previous studies, we identify Tn*125*, IS*26* and Tn*3000* as the main contributors to *bla*$_{\mathrm{NDM}}$ mobility but go further and estimate the timing of the initial emergence of *bla*$_{\mathrm{NDM}}$ to pre-1990, around two decades prior to its first detection and rapid dissemination. Our findings suggest that this global spread was driven primarily by transposons, with plasmids playing more of a role in local transmission.

Results

A global dataset of *bla*$_{\text{NDM}}$ carriers

We compiled a dataset of 6155 bacterial genomes (7148 contigs) carrying at least one copy of *bla*$_{\text{NDM}}$ (Figure 1). These include: published assemblies from NCBI RefSeq[35] (*n*=2632), NCBI GenBank[36] (*n*=1158) and Enterobase[37] (*n*=1379); bacterial genomes assembled using short read *de novo* assembly from NCBI's Sequence Read Archive (SRA) (*n*=882); and newly generated bacterial genomes isolated from 79 hospitalized patients across China and 25 livestock farms assembled using hybrid PacBio-Illumina *de novo* assembly (*n*=104) (Supplementary Table 1, Supplementary Figure 1). While public genomes have inherent sampling biases, using them is the most comprehensive approach available[1]. Data was included from 251 different Bioprojects, with more than half the samples linked to two large-scale database refinement efforts[38,39].

The dataset included *bla*$_{\text{NDM}}$-positive isolates from 88 states (Figure 1A) mostly collected in Asia, particularly mainland China (*n*=1270), European countries (941), USA (461), Thailand (419) and India (361). At least 27 bacterial genera were represented, with a large fraction of *Klebsiella* and *Escherichia* isolates (2664 and 2154 genomes respectively; Figure 1B; Supplementary Data 1). Collection dates were recorded for 4816 samples (78.25%). Of these, the majority were collected between 2014-2019 (71.05%, Figure 1C). The dataset also includes 55 genomes collected in 2010 or earlier. These include the *K. pneumoniae* isolate from 2008 Sweden in which *bla*$_{\text{NDM}}$ was first characterized[2]; one *Enterobacter hormaechei* isolate from 2008 India[40]; one *S. enterica* isolate from 2008 London, UK[41]; one *A. baumannii* isolate from an individual of Balkan origin collected in Germany in 2007[42,43]; and nine assembled *E. coli* genomes from urine samples collected in Greece in 2007 (Supplementary Data 1).

The dataset contained 17 known variants of NDM. NDM-1 was the most abundant (*n*=4127; Supplementary Figure 2A) with NDM-5 (*n*=2394) increasing in prevalence after 2012 (Supplementary Figure 2B and C). Variants showed different associations with plasmid types (Supplementary Figure 2D) and genera (Supplementary Figure 2E) but were fairly evenly distributed across the world except for *bla*$_{\text{NDM-4}}$-carrying isolates largely collected in Southeast Asia and *bla*$_{\text{NDM-9}}$ predominantly found in East Asia (Supplementary Figure 2F).

Plasmid backbones carrying *bla*$_{\mathrm{NDM}}$

We identified 33 different replicon types on 1222 contigs using PlasmidFinder[44] (Figure 1D). The most prevalent replicon type was IncX3 (444 contigs), and abundant types exhibited geographic structure (Supplementary Figure 3). To further identify uncharacterised plasmid types, we mapped 3599 contigs to a set of complete plasmid reference sequences after discarding short contigs (see Methods). This revealed 181 clusters of similar putative plasmid sequences (Supplementary Figure 4; Supplementary Data 2). Most clusters (*n*=105) grouped contigs of the same replicon type and contained a small number of contigs (only 27 clusters included >10 contigs), in line with a diverse and dynamic population of plasmid backbones for *bla*$_{\mathrm{NDM}}$.

The majority (n=2427; 68.4%) of *bla*$_{\mathrm{NDM}}$-carrying contigs were associated with small putative plasmids (<10 Kb; Supplementary Figure 4). While this could suggest small plasmids play a key role as *bla*$_{\mathrm{NDM}}$ carriers, this pattern could also result from consistently fragmented *de novo* assemblies due to duplicated ISs and transposons. Consistent with this latter hypothesis, 610 contigs mapped to pKP-YQ12450 which is likely a 7.8 Kb fragment of a larger plasmid[22]. Conversely, Roach et al. provide evidence that other small *bla*$_{\mathrm{NDM}}$-carrying plasmids (Peruvian pKP-NDM-1_isoforms 1-5) are inherited by descent and are result of transposon-mediated plasmid fusion[45].

Resolving structural variants in the *bla*$_{\mathrm{NDM}}$ flanking regions

To go beyond a static reference-based view of variation around *bla*$_{\mathrm{NDM}}$ and gain a detailed overview of the possible events in its evolution, we developed an alignment-based approach to progressively resolve genomic variation moving upstream or downstream from the gene (see Methods, Figure 2). In brief, a pairwise discontiguous Mega BLAST search (v2.10.1+)[46,47] is applied to all *bla*$_{\mathrm{NDM}}$-carrying contigs to identify all possible homologous regions between each contig pair. Only BLAST hits covering the complete *bla*$_{\mathrm{NDM}}$ gene are retained (Figure 2A). Next, starting from *bla*$_{\mathrm{NDM}}$, a gradually increasing 'splitting threshold' is used to monitor structural variants as they appeared upstream or downstream of the gene. At each step, a network of contigs (nodes) that share a BLAST hit with a minimum length as given by the 'splitting threshold' is assessed (Figure 2B). As we move upstream or downstream and further away from the gene, the network starts to split into smaller clusters, each carrying contigs that share an uninterrupted stretch of homologous DNA, which can be represented as a tree (Figure 2C). This approach treats the upstream and downstream flanking regions separately rather than simultaneously and is agnostic to whether splitting into 'sequence clusters' is caused by structural variants of the same genomic background or different genomic backgrounds.

Upstream of *bla*$_{\mathrm{NDM}}$, >98% of sufficiently long contigs included a ~75 bp fraction of IS*Aba125*, supporting Tn*125* as an ancestral transposon of the *bla*$_{\mathrm{NDM}}$ gene in agreement with previous work[1,5,9,10] (Supplementary Figures 5 and 6). However, the homology of the region upstream of *bla*$_{\mathrm{NDM}}$ falls quickly: within a few hundred base pairs of the *bla*$_{\mathrm{NDM}}$ start codon the region splits into multiple structural variants, none of which dominate the considered pool of contigs (Supplementary Figures 5 and 6). We identified 141 different structural variants within 1200 bp upstream of *bla*$_{\mathrm{NDM}}$. This upstream region contained a high number of ISs (e.g. IS*Aba125* [*n*=243], IS*5* [*n*=426], IS*3000* [*n*=60], IS*Kpn14* [*n*=55], and IS*Ec33* [*n*=147]). This transposition hotspot probably contributes to fragmented assemblies: 2269 contigs were excluded from further analysis for being too short (Supplementary Figure 3).

The downstream flanking region exhibits more gradual structural diversification than the upstream region, with one dominant putative ancestral background (Figure 3). As illustrated by the stem of the tree of structural variants, many of the 7014 contigs analysed contained complete sequences of the same set of genes: *ble* (6863 contigs), *trpF* (6038), *dsbD* (5551), *cutA* (2731), *groS* (2175), *groL* (1631). When restricted to *bla*$_{\mathrm{NDM}}$-positive contigs of sufficient length to possibly harbour the full repertoire of these genes (*n*=3786), almost half carry all of them (*n*=1,631; 43.1%). In addition, we find dominant structural variants associated with various source databases and sequence lengths hence diminishing the impact of the sampling bias (Supplementary Figure 7)

Early events in the spread of *bla*$_{\mathrm{NDM}}$

While we did not observe any strong overall signal in the distribution of associated plasmid backbones, bacterial genera, or sampling locations, closer examination of mobilome features common to sufficiently large numbers of isolates indicated early events in the spread of *bla*$_{\mathrm{NDM}}$. The putative ancestral Tn*125* background, with an uninterrupted downstream IS*Aba125* element, was seen in contigs mainly from *Acinetobacter* and *Klebsiella* (Figure 3 top). Conversely, the diversity of bacterial genera carrying IS*Aba125* upstream is more substantial (Supplementary Figure 5 top). Only 203 contigs carried a complete IS*Aba125* downstream of *bla*$_{\mathrm{NDM}}$, of which 147 carried an IS*Aba125* sequence in proximity (<9 Kb) to the *bla*$_{\mathrm{NDM}}$ start codon. These account for a minority (7%; 147/2097) of isolates when sufficiently long contigs are considered. This supports the initial dissemination of *bla*$_{\mathrm{NDM}}$ by Tn*125* to other plasmid backbones predominately being mediated by *Acinetobacter* and *Klebsiella*, after which the transposon was disrupted by other rearrangements.

IS*3000*, both upstream and downstream, was almost exclusively associated with samples from *Klebsiella* (Figure 3 and Supplementary Figure 5). Thus, as suggested by Campos et al.[26], Tn*3000* – a composite transposon made of two copies of IS*3000* – likely re-mobilized *bla*$_{\mathrm{NDM}}$ following the 'fossilization' of Tn*125*; our findings suggest this secondary mobilization primarily happened in *Klebsiella* species. Tn*5403* was found extensively associated with IncN2 plasmids (Figure 3) which could have placed *bla*$_{\mathrm{NDM}}$ in this background via cointegrate intermediate as previously suggested by Poirel et. al.[9] Some elements of the mobilome were geographically linked e.g., IS5 which was predominantly found upstream of *bla*$_{\mathrm{NDM}}$ on IncX3 plasmids in *Escherichia* from East Asia (Supplementary Figure 5). IS*5* is known to enhance transcription of nearby promoters in *E. coli*[48] and its abundance and positioning just upstream of *bla*$_{\mathrm{NDM}}$ suggests a similar role in this case.

One of the most commonly identified transposable elements in the downstream flanking region (~30% prevalence) was ISCR1 (IS91 family transposase) (Figure 3) always accompanied by *sul1* and occasionally in configuration with *ant1* or *pspF*, *ampR*, and *dap* genes. In some cases, a small and possibly fragmented putative IS, which we refer to as 'IS-?', is found further downstream. IS-? bears little similarity to known ISs and it is unclear what role it plays in the mobility of *bla*$_{\mathrm{NDM}}$. ISCR1 is found at various positions downstream of *bla*$_{\mathrm{NDM}}$ and often in *Escherichia* and *Klebsiella* species. We note that, in most cases, the orientation of ISCR1 should prevent this element from mobilizing *bla*$_{\mathrm{NDM}}$ (Figure 3)[14]. Nevertheless, the prevalence of this element could be due to the several AMR genes it can mobilize, such as *sul1* or *ampR*. ISCR1s are mainly found in complex class 1 integrons[14], however, not many annotated integrase genes are located within the vicinity of *bla*$_{\mathrm{NDM}}$. In fact, only 15 contigs were found to have an integrase <50 Kb away from *bla*$_{\mathrm{NDM}}$ and none showed any consistency in integrase placement with respect to *bla*$_{\mathrm{NDM}}$. This suggests integrases play a minor role in the dissemination of *bla*$_{\mathrm{NDM}}$.

Another notable ISCR element is ISCR27 which is consistently found immediately downstream of the *groL* gene at high prevalence (33.1% of sufficiently long contigs; Figure 3). Contrary to its ISCR1 relative, ISCR27 is correctly oriented to mobilize *bla*$_{\text{NDM}}$ as is presumed to have happened during the initial mobilization of the progenitor of *bla*$_{\text{NDM}}$[10]. However, we find no evidence of subsequent ISCR27 mobility. The origin of rolling-circle replication of ISCR27 (*oriIS*; GCGGTTGAACTTCCTATACC) is located 236 bp downstream of the ISCR27 transposase stop codon. The region downstream of this stop codon in all structural variants bearing a complete ISCR27 is highly conserved for at least 750 bp (Figure 3).

Subsequent rearrangements dominated by IS*26*

Three sharp drops in the number of considered contigs at particular distances downstream of *bla*$_{\text{NDM}}$ (see Figure 3, e.g., region 3000-3300bp) prompted us to investigate these distinct cut-offs. We mapped 781 raw Illumina paired-end sequencing reads from our dataset back to their matching *bla*$_{\text{NDM}}$ contigs. The read overhangs (≥50bp) that mapped to the downstream end of the contigs were screened against the ISFinder database[49]. The ≥50bp overhangs associated with 3000-3300 long flanks downstream of *bla*$_{\text{NDM}}$ corresponding to the largest observed drop almost exclusively match the left inverted repeat (IRL) of the IS*26* sequence (Supplementary Figure 8). Another hotspot, associated to IS*26* was found around 7,500bp, while at around 7,800bp a number of overhanging reads mapped to IS*Aba125*. These positions roughly match the third drop in the number of contigs observed 7500-8000bp downstream of *bla*$_{\text{NDM}}$. No ISs were found to match the second drop in number of contigs (5000-5250bp).

IS*26*, although often found in two adjacent copies forming a seemingly composite transposon, is a so-called pseudo-composite (or pseudo-compound) transposon[50]. In contrast to composite transposons, a fraction of DNA flanked by the two IS*26* is mobilized either via cointegrate formation or in the form of a circular translocatable unit (TU), which consists of a single IS*26* element and a mobilized fraction of DNA, and inserts preferentially next to another IS*26*[50,51]. Taken together, the presented results, including Supplementary Figure 8, suggest three possible explanations for the presence of short *bla*$_{\text{NDM}}$ carrying contigs in the dataset: (i) the presence of IS*26* TUs in the host cell; (ii) other circular DNA formations mediated by plasmid recombination, transposons[9,45] or ISCR elements[13,52]; (iii) missassembly of contigs due to presence of multiple copies of the same ISs[53].

To further investigate the mobility of *bla*$_{\text{NDM}}$, we characterised the most common (pseudo-)composite transposons theoretically capable of mobilizing *bla*$_{\text{NDM}}$ (Supplementary Figure 9). These were defined as stretches of DNA flanked by two matching complete or partial ISs <30 Kb apart and enclosing *bla*$_{\text{NDM}}$. In total, we identified 640 composite transposons in 468 contigs which comprised 31 different types with the most frequent being: IS*26* (231 instances), IS*3000* (forming Tn*3000*; 168), IS*Aba125* (forming Tn*125*; 138 instances), and IS*15* (28) (Supplementary Figure 9B). Interestingly, we observe 80 cases where >2 of the same IS flank *bla*$_{\text{NDM}}$. These are mostly IS*26* (59) which could indicate the presence of cointegrate formation[50] and showcases increased activity of this particular insertion element. Only 431 of the 640 putative composite transposons identified contained both complete flanking ISs, while others had at least one IS partially truncated. In addition, 1681 ISCR27, and 150 ISCR1 were found in similar proximity and appropriate orientation to mobilize *bla*$_{\text{NDM}}$ (Supplementary Figure 9B). However, as mentioned earlier, their role in transposition of *bla*$_{\text{NDM}}$ appears minor.

In the majority of cases, composite transposons Tn*125* and Tn*3000* were found to have a consistent length ranging from 7-10Kb (Supplementary Figure 9A). Similarly, ISCR1 and ISCR27 are found at fixed positions downstream of *bla*NDM. However, the lengths of transposons formed by IS*15*, a known variant of IS*26*[54], and especially IS*26* were found to be more variable. Pairs of IS*26* are found to be 2.5-30Kb apart again consistent with increased activity and multiple independent insertions. We note that IS*15* and IS*26* occur at increased presence in samples collected in East and Southeast Asia (Supplementary Figure 9C). These occur roughly equally in *Escherichia* and *Klebsiella* genera (Supplementary Figure 9D) and are associated to multiple plasmid backbones, but predominantly on IncF plasmids (Supplementary Figure 9E). Tn*125* and Tn3000 have a notable predominance in the Indian subcontinent (Supplementary Figure 9C) and largely in the *Acinetobacter* and *Klebsiella* genera respectively (Supplementary Figure 9D).

Molecular dating of key events

We estimated the relative timing of the formation of the Tn*125* and Tn*3000* transposons (see Methods). After selecting only contigs with conserved transposon configurations we aligned each transposon region and identified the likely root (i.e., ancestral) sequence by assessing temporal patterns (Supplementary Figures 10 and 11; see Methods). Overall, we observed fewer SNPs, mostly located within the transposase gene, in the alignment of Tn*3000* compared to Tn*125*, but observed a significant temporal signal for both (Supplementary Figures 12-13). We also assessed temporal signal for three other prevalent insertion events (Figure 3), namely: *bla*NDM with downstream ISCR27, *bla*NDM with correctly oriented downstream *folP*-ISCR1 (+ strand), and *bla*NDM – dsbD with downstream ISCR1 (- strand) ending with an unknown putative IS (labelled IS-?). However, no significant temporal signal was recovered for these events.

This Bayesian analysis indicated that the most recent common ancestor (MRCA) of the Tn*125* transposon carrying the *bla*NDM gene dated to before 1990 (Figure 4A). While the time intervals are uncertain, the results are consistent with a MRCA in the mid-20th century – strikingly half a century prior to the first reported Tn*125*-*bla*NDM-positive isolates[7]. Conversely, the mobilization of *bla*NDM by Tn*3000* is estimated to have happened later at the turn of the millennium (Figure 4B). These findings are consistent with a wider narrative whereby the spread of *bla*NDM was initially driven by Tn*125* mobilization before subsequent transposition by Tn*3000*, IS*26* and others.

Temporal diversity in *bla*NDM isolates suggests role of plasmids

The earliest samples in our dataset are from 2007 to 2010 and comprise 21 *bla*NDM-positive isolates. These already encompass seven bacterial species, collected in eight countries spanning four geographic regions (17 clinical samples and four of unknown origin from South Asia, Middle East, Oceania, and Europe). Such a wide host and geographic distribution, even in the earliest available genomes, illustrates the extraordinarily high mobility of *bla*NDM at this stage and is consistent with our molecular dating estimates.

In order to trace the progress of *bla*NDM's rapid spread after 2005 (coinciding with the first published observations), we measured diversity over time for several metadata categories, including country, genera, plasmid backbone and IS presence (Supplementary Figure 16; see Methods). The change in diversity of the countries associated to *bla*NDM-positive isolates was used to approximate the broad patterns of global dissemination of *bla*NDM. Our results are consistent with the spread stabilising between 2013-2015, with a gradual decline in diversity afterwards (Supplementary Figure 16A). This observation supports a scenario whereby the global dissemination of NDM took place over 8-10 years. Temporal diversity of bacterial genera was largely unchanged, consistent with *bla*$_{\mathrm{NDM}}$ having been highly mobile across genera since at least 2005 (Supplementary Figure 16B).

The estimated change in the diversity of countries associated to *bla*$_{\mathrm{NDM}}$-positive isolates was positively correlated with other metadata categories (Supplementary Figure 17) suggesting it holds information which can be leveraged to reconstruct dissemination trends. The strongest correlation was found between the diversity of countries and plasmid backbones (ρ = 0.864 [0.691-0.964]) supporting a strong dependence between the two (Supplementary Figure 17B). To further investigate this relationship, we assessed the correlation between genetic and geographic distance between pairs of confirmed plasmid contigs (tested for IncF, IncX3, IncC, IncN2 and confirmed plasmid contigs >10kb) as a function of the distance downstream of *bla*$_{\mathrm{NDM}}$ gene (Supplementary Figure 18, see Methods).

No relationship was detected for IncX3 and IncN2 plasmids (Supplementary Figure 18A and B) likely due to the lack of long plasmid sequences and deficient geographic distance between pairs of plasmids as both replicon types are mostly localized to China and India respectively (Supplementary Figure 3). However, in all other cases aside from IncN2 plasmids, a peak in the correlation recovered between genetic and geographical distance was observed immediately downstream of *bla*$_{\mathrm{NDM}}$ possibly signifying more recent and local genome reshuffling events (Supplementary Figure 18). More importantly, in IncF and IncC, and other confirmed plasmid contigs, a notable and gradual increase in the strength of correlation was noted further downstream as more plasmid sequence is included in the analysis (Supplementary Figure 18B, C, and D). These trends suggest that plasmids carrying *bla*$_{\mathrm{NDM}}$ are geographically structured and that dissemination of *bla*$_{\mathrm{NDM}}$ is a fundamentally spatial process. This would be consistent with the existence of plasmid niches: settings to which particular plasmids are more adapted.

Discussion

In this study, we have characterised the extant structural variation around *bla*_{NDM} in a large global dataset in order to reconstruct its evolutionary history and the main actors underlying its spread. Our results, largely summarized in Figure 3, highlight an ancestral background of *bla*NDM as well as several insertion events and a myriad of other genetic reshuffling, together pointing to an early emergence of *bla*NDM followed by a more recent and rapid dissemination globally. Genetic reshuffling and mobilization of *bla*NDM by multiple transposons aided its rapid dissemination via a multitude of plasmid backbones.

We go beyond previous smaller studies by dating the MRCA of the hypothesised ancestral form – the transposon Tn*125*, together with *bla*NDM in its chimeric form[10] – to pre-1990, and possibly well back into the mid-20th century. A likely scenario is an origin in *Acinetobacter* in the Indian subcontinent. We note that Tn*125* is mostly present in *Acinetobacter* and *Klebsiella* species and it is likely this transposon played an important role in early plasmid jumps of *bla*NDM, given it is the dominant transposon in our comprehensive dataset which encompasses the ancestral genetic background of *bla*NDM – *groS*/*groL* genes and ISCR27 sequence. We also estimate the formation of a secondary transposon, involving Tn*3000*, which remobilized the region likely in *Klebsiella* species sometime between the 1980s and early 2000s. However, we suggest Tn*3000* likely played a lesser role in the early spread of *bla*NDM as it does not include the ISCR27 found at least partially preserved in many samples.

In total, 31 different putative transposons were identified within our dataset. Their role, together with integrons and other transposable elements, is likely mostly minor or disruptive, as suggested for ISCR1. However, we do identify IS*26* as of interest, given it frequently forms putative transposons in our dataset, especially in IncF plasmids. IS*26* is known for its increased activity and rearrangement of plasmids in clinical isolates[55] and has been observed to drive within-plasmid heterogeneity even in a single *E. coli* isolate[56]. Thus, IS26 flanked pseudo-composite transposons likely represent the most important contributor to genetic reshuffling of *bla*NDM in recent times.

Our assessment of temporal diversity of countries of origin of *bla*NDM positive isolates supports a globalisation peak in 2013-2015. Such a rapid 8-10 year world-wide spread has been suggested for other important mobile resistance genes such as the *mcr-1* gene, mediating colistin resistance[57]. The extent to which this model of 'rapid global spread' applies to other transposon-borne resistance elements remains to be determined.

We found 33 different plasmid types carrying *bla*NDM and a positive correlation between genetic distance calculated for differing lengths of plasmid backbones and geographic distances of sampling locations. Such an observation is consistent with the existence of a constraint on plasmid spread, i.e. plasmid niches. Such niches may exist as a result of local ecological and evolutionary pressures acting on particular plasmid backbones. Such forces may include country boundaries limiting population movement, region-specific patterns in antibiotic usage, influence of co-resistance, plasmid fitness costs, conjugation rates and copy numbers, the narrow host range of the majority of bacterial plasmids[58], or plasmids being associated with particular locations or environmental niches[59], all may contribute to restricting plasmid geographical range. Thus, an introduction of another plasmid into a foreign plasmid niche may lead to plasmid loss or fast adaptation by, for instance, acquisition of resistance and other accessory elements. This hypothetical scenario also provides an opportunity for resistance to spread by transposition or recombination, by which a new resistance gene could establish itself into another plasmid niche. In the case of *bla*$_{\mathrm{NDM}}$, this would also imply that after the initial introduction of *bla*$_{\mathrm{NDM}}$ to a geographic region, dissemination and persistence of the gene could proceed idiosyncratically – selection for carbapenem resistance being just one of many selective pressures acting on plasmid diversity.

The importance of transposon movement has been previously demonstrated by work on plasmid networks[58,60], as well as several papers promoting a Russian-doll model of resistance mobility[57,61]. Considering our results, we suggest a conceptual framework of AMR gene dissemination across genera where plasmid mobility is for the most part restricted. Although plasmids can facilitate rapid spread within species and geographical regions, plasmid transfer is not the main driver of widespread dissemination. Instead, most plasmid horizontal transfers are likely only transient, with plasmids generally failing to establish themselves in the new bacterial host. Though, such aborted plasmid exchanges still provide a crucial opportunity for between-plasmid transposon jumps and genetic recombination to spread AMR genes across bacterial species.

Methods

Compiling the curated dataset of NDM sequences

We compiled an extensive dataset of 6155 bacterial genomes carrying the *bla*$_{\mathrm{NDM}}$ gene from several publicly available databases. 2632, 1158, and 1379 fully assembled genomes were downloaded from NCBI Reference Sequence Database[35,62] (RefSeq; accessed on 15th of April 2021), NCBI's GenBank[36] (accessed on 15th of April 2021), and EnteroBase (accessed on 27th of April 2021)[37] respectively. The EnteroBase repository was screened for *bla*$_{\mathrm{NDM}}$ using BlastFrost (v1.0.0)[63] allowing for one mismatch. In addition, we used the Bitsliced Genomic Signature Index (BIGSI) tool (v0.3)[64] to identify all Sequence Read Archive (SRA) unassembled reads which carry the *bla*$_{\mathrm{NDM}}$ gene. At the time of writing, a publicly available BIGSI demo did not include sequencing datasets from after December 2016. Therefore, we manually indexed and screened an additional 355,375 SRA bacterial sequencing datasets starting from January 2017 to January 2019. We required the presence of 95% of *bla*$_{\mathrm{NDM\text{-}1}}$ *k*-mers to identify NDM-positive samples from raw SRA reads. This led to the inclusion of a further 882 isolates. The dataset also included 104 NDM-positive genomes from 79 hospitalized patients across China and 25 livestock farms selected from two previous studies[65,66]. These were sequenced using paired-end Illumina (Illumina HiSeq 2500) and PacBio (PacBio RS II). The sequencing reads are available on the Short Read Archive (SRA) under accession number PRJNA761884. All reads were *de novo* assembled using Unicycler (v0.4.8)[67] with default parameters while also specifying hybrid mode for those isolates for which we had both Illumina short-read and PacBio long read sequencing data. Spades (v3.11.1)[68] was applied, without additional polishing, for cases where Unicycler assemblies failed to resolve.

Assembled genomes were retained when they derived from a single BioSample identifier. Contigs carrying the *bla*$_{\mathrm{NDM}}$ gene were identified using BLAST (v2.10.1+)[46]. Forty-eight contigs were found to carry more than one copy of *bla*$_{\mathrm{NDM}}$ and were not included in our analyses and eighty-eight contigs were excluded due to having partial (<90%) *bla*$_{\mathrm{NDM}}$ hits. Fourteen assemblies had a single *bla*$_{\mathrm{NDM}}$ gene split into two contigs; these 28 contigs were also excluded. Several contigs were also removed due to poor assembly quality. The filtering resulted in a dataset of 7148 contigs (6,155 samples) which were used in all subsequent analyses. Of these, 958 assembled genomes were found to contain *bla*$_{\mathrm{NDM}}$ on multiple (mostly two) contigs, each harbouring a single and complete copy of *bla*$_{\mathrm{NDM}}$. Even though the information about sequencing platform or assembly methods of most samples from RefSeq, GenBank and Enterobase databases could not be determined, the distribution of *bla*$_{\mathrm{NDM}}$-positive contig lengths (Supplementary Figure 1) reveals they are likely to be based on short reads with the minority of contigs, mostly from RefSeq, reaching the quality of a hybrid *de novo* assembly. The full table of contigs and metadata considered is available as Supplementary Data 1.

Annotating the dataset

Full metadata for each genome was collected from its respective database and the R package 'taxize'[69] used to retrieve taxonomic information for each sample. In the case of samples for which exact sampling coordinates were not provided, Geocoding API from Google cloud computing services was used to retrieve coordinates based on location names.

Coding sequences (CDS) of all NDM-positive contigs were annotated using the annotation tool Prokka (v1.14.6)[70] and Roary (v3.13.0)[71] run with minimum blastp percentage identity of 90% (-i 0.9) and disabled paralog splitting (-s). To identify plasmid replicon types[72], contigs were screened against the PlasmidFinder replicon database (version 2020-02-25)[44] using BLAST (v2.10.1+)[46] where only BLAST hits with a minimum coverage of 80% and percentage identity of ≥95% were retained. In cases where two or more replicon hits were found at overlapping positions on a contig, the one with the higher percentage identity was retained. Identified plasmid types were used to cluster contigs into broader plasmid groups: IncX3, IncF, IncC, IncN2, IncHI1B, IncHI2, and other (Figure 1D).

NDM-positive contigs were also screened against a dataset of complete bacterial plasmids. Bacterial plasmid references were obtained from RefSeq[35] and curated as described in Acman et al.[73] Mash, a MinHash based genome distance estimator[74], was applied with default settings to evaluate pairwise genetic distances between contig sequences and plasmid references. Contig-reference hits with less than 0.05 Mash distance and less than 20% difference in length were retained. Additional pruning was implemented such that, for each contig analysed, only the 10% of best scoring plasmid reference hits were retained. A table of pairwise genetic distances between contigs and references was represented as a network which was then analysed with the infomap[75] community detection algorithm. Contigs were annotated according to their community membership and the network was visualized using Cytoscape[76] (Supplementary Figure 4).

Resolving structural variants of NDM-positive contigs

A novel alignment-based approach was used to identify stretches of homology (i.e., maximal alignable regions) as well as structural variations across all contigs upstream and downstream of *bla*$_{\mathrm{NDM}}$ gene. A conceptual illustration of the method is presented in Figure 2. First, contigs carrying *bla*$_{\mathrm{NDM}}$ were reoriented such that the *bla*$_{\mathrm{NDM}}$ gene was located on the positive-sense DNA strand (i.e., facing 5' to 3' direction). A discontiguous Mega BLAST (v2.10.1+)[47] search with default settings was then applied against all pairs of retained contigs. This method was selected over the regular Mega BLAST implementation as it is comparably fast, but more permissive towards dissimilar sequences with frequent gaps and mismatches. BLAST hits including a complete *bla*$_{\mathrm{NDM}}$ gene represent maximal stretches of homology around the gene for every pair of contigs. The analysis continues by considering only portions of BLAST hits: (i) the start of *bla*$_{\mathrm{NDM}}$ gene and the downstream sequence or (ii) the end of the *bla*$_{\mathrm{NDM}}$ gene and the upstream sequence depending on the analysis at hand: the downstream or the upstream analysis respectively. This trimming of BLAST hits establishes *bla*$_{\mathrm{NDM}}$ as an anchor and enables comparisons to be made across all samples.

A table of BLAST hits can be considered as a network (graph), where each pair of contigs (i.e., nodes) are connected by the edge weighted by the length of the BLAST hit. The algorithm proceeds with a stepwise network analysis of BLAST hits. For this purpose, a 'splitting threshold' was introduced. Starting from zero which represents the start/end position of *bla*$_{\mathrm{NDM}}$ gene, the threshold is gradually increased by 10 bp. At each step, BLAST hits with a length lower than the value given by the 'splitting threshold' are excluded. Thus, as the 'splitting threshold' increases, a network of BLAST hits is also pruned and broken down into components – groups of interconnected nodes (contigs). It is expected that contigs within each component share a homologous region downstream (or upstream) of *bla*$_{\mathrm{NDM}}$ at least of the length given by the threshold. It is therefore not possible for a single contig to be assigned to multiple components. Components of size <10 are labelled as 'Other Structural Variants'. Also, contigs that are shorter than the defined 'splitting threshold' and share no edge with any other contig are considered as 'cutting short'.

By tracking the splitting of the network as the 'splitting threshold' is increased, one can determine clusters of homologous contigs at any given position downstream or upstream from the anchor gene (here *bla*$_{\mathrm{NDM}}$), as well as the homology breakpoint. The precision of the algorithm is directly influenced by the step size, in this case 10 bp, and the alignment algorithm, in this case discontiguous Mega BLAST. We assessed the precision of the algorithm on the tree of structural variations downstream of *bla*$_{\mathrm{NDM}}$ (Figure 3). To this end, we compared extended 50bp sequence fragments of each branching point in the tree checking for missed homologies and comparing Mash distances between pairs of branched-out contigs. We found no similarities among 50bp fragments of any split branches. The described algorithm is available at https://github.com/macman123/track_structural_variants.

Analysing the contig overhanging reads

To investigate the reasons behind a number of distinctively short *bla*$_{\mathrm{NDM}}$-carrying contigs, we mapped 781 raw Illumina paired-end sequencing reads (originally downloaded from SRA) back to their matching contigs. The mapping was done using BBMap[77] (v38.59; *maxindel = 0* and *minratio = 0.2* settings). Within the output SAM file, only the overhanging reads with the CIGAR string matching the "*[0-9]*M[0-9]*S"* regular expression were selected. All overhangs of reads ≥50bp were screened against ISFinder database[49].

Molecular tip-dating analysis.

The 112 complete Tn*125* and 73 complete Tn*3000* contigs with a known collection date and harbouring *bla*$_{\mathrm{NDM}}$ were sequentially aligned (--pileup flag) using Clustal Omega (v1.2.3)[78] specifying the *bla*$_{\mathrm{NDM\text{-}1}}$ sequence (FN396876.1) as a profile. Each alignment was manually inspected using UGENE (v38.0)[79]. The ancestral (i.e., root) sequence was determined by evaluating SNP frequencies over time (Supplementary Figure 9 and 10). Due to a short sampling time span and relatively few mutations present, it is unlikely that any one non-ancestral SNP has become dominant in the population. Therefore, we expect the ancestral sequence to have a higher SNP frequency in earlier years.

We find that, in all but two cases, the consensus sequence of an alignment displaying this behaviour. The first exception is the consensus sequence allele of Tn*125* at the variable position 441 (Supplementary Figure 9). This allele has a low frequency in 2009. However, by inspecting the allele frequency table, we observe the low frequency is based on a single sample. Leaving out this early sample restores the desired frequency pattern; hence the consensus allele is considered ancestral in this case. The second exception is the variable position 449 in the case of the Tn*3000* alignment (Supplementary Figure 10). The consensus allele 'a' is not found in the early sample from 2009. Both allele 't', present in the early sample, and allele 'a' were found equally frequent in more recent samples. Thus, due to lack of other evidence, allele 't' was considered ancestral. Determined ancestral sequences were used to evaluate temporal signal in the alignment, and in the subsequent rooting of phylogenetic trees.

Date randomization (10,000 iterations) and linear regression analyses were employed to estimate the presence of temporal signal in the alignment[80–82] (Supplementary Figures 11 and 12). Tn*125* and Tn*3000* showed significant temporal signal using simple regression (*p*=0.0356 and *p*=0.0456 respectively) and date randomization (true evolutionary rate quantiles >0.95).

Bayesian based molecular dating approaches were implemented in BEAST2 (v2.6.0)[83] and BactDating[84] to infer the date of the emergence of the two transposons. Both BEAST2 and BactDating were run specifying a strict prior on the molecular clock. For BEAST2, the generalised time reversible (GTR) substitution model prior was used together with three population size models: Coalescent Constant population, Coalescent Exponential population, and Coalescent Bayesian Skyline. In addition, all BEAST2 and BactDating runs were supplied with a maximum likelihood (ML) phylogenetic tree (starting tree prior) constructed from both transposon alignments using RAxML (v8.2.12)[85] with specified GTRCAT substitution model and rooted using the inferred ancestral sequences. The chosen MCMC chain lengths for BactDating and BEAST2 runs were 10^7 and $1.5\text{x}10^9$ respectively to ensure convergence. We evaluated effective sample sizes (ESS) of the posterior distributions using *effectiveSize* function implemented in *coda*[86] R package after discarding the first 20% of burn-in (Supplementary Figures 13 and 14). All BEAST2 and BactDating runs successfully converged with ESS of the posteriors close to or above 200. BEAST2 input files are available as xml files in Supplementary Data 3.

Estimating Shannon entropy among NDM-positive contigs

We estimated Shannon entropy ('diversity') for several categorizations of *bla*_{NDM}-containing contigs: country of sampling, bacterial host genera, plasmid backbones (determined by mapping to plasmid reference sequences), and ISs flanking the *bla*_{NDM} gene. To estimate entropy of the population and to provide confidence intervals around our estimates, we use bootstrapping with replacement (1000 iterations). At each iteration, entropy was estimated for a sampled set of contigs (*X*) classified into *n* unique categories according to the following formula:

$$H(X) = -\sum_{i=1}^{n} P(x_i) log\, P(x_i),$$

where the probability $P(x_i)$ of any sample belonging to any particular category x_i (e.g., country or plasmid backbone) is approximated using the category's frequency. Accordingly, higher entropy values indicate an abundance of equally likely categories, while lower entropy indicates a limited number of categories.

Estimating correlation between genetic and geographic distance

Geographic distance between pairs of samples was determined using their sampling coordinates and the *geodist*[87] R package. Exact Jaccard distance (JD) was used as a measure of the genetic distance calculated using the tool Bindash (v0.2.1)[88] with *k*-mer size equal to 21 bp. The JD is defined as the fraction of total *k*-mers not shared between two contigs. JD between all pairs of contigs was first calculated on a 1000bp stretch of DNA downstream of *bla*$_{\mathrm{NDM}}$ start codon continuing with a 500bp increments. At each increment, the two distance matrices (genetic and geographic) were assessed using the *mantel* function (Spearman correlation and 99 permutations) from the *vegan*[89] package in R. The correlation between genetic and geographic distance, was plotted as a function of distance from *bla*$_{\mathrm{NDM}}$ gene (Supplementary Figure 17).

Acknowledgements

M.A. was supported by a Ph.D. scholarship from University College London. H.W. is supported by National Natural Science Foundation of China (81625014). L.v.D., H.W. and F.B. acknowledge financial support from the Newton Fund UK-China NSFC initiative (MRC Grant MR/P007597/1 and 81661138006). L.v.D. and F.B. are supported from a Wellcome Institutional Strategic Support Fund (ISSF3) – AI in Healthcare (19RX03). F.B. additionally acknowledges support from the BBSRC GCRF scheme and the National Institute for Health Research University College London Hospitals Biomedical Research Centre. L.v.D is supported by a UCL Excellence Fellowship. M.A., L.v.D and F.B. acknowledge UCL Biosciences Big Data equipment grant from BBSRC (BB/R01356X/1). L.P.S. is a Sir Henry Wellcome Postdoctoral Fellow funded by Wellcome (Grant 220422/Z/20/Z). The funders had no role in study design, data collection, interpretation of results, or the decision to submit the work for publication. Lastly, M.A. would like to thank Nicola de Maio for informal discussions which led to the idea for the algorithm used to track structural variants.

Contributions

M.A., F.B., L.v.D. and H.W. conceived the project and designed the experiments. M.A., L.v.D., L.P.S., and N.L. collected data from online repositories. R.W., Y.Y., Q.W., S.S, and H.C sequenced samples from Chinese hospitals. M.A., L.v.D, and R.W. *de novo* assembled all the genomes. M.A. performed all the analyses under the guidance of L.v.D and F.B. M.A., L.v.D. and F.B. take responsibility for the accuracy and availability of the results. M.A. wrote the paper with contributions from L.P.S., L.v.D., and F.B. All authors read and commented on successive drafts and all approved the content of the final version.

Competing interests

The authors declare no financial or non-financial competing interests.

Data availability

The accession numbers of bacterial genomes obtained from the RefSeq, Enterobase and SRA databases are given in the Supplementary Data 1. One hundred and four paired-end Illumina and PacBio sequencing data from China are available on SRA under the BioProject accession number PRJNA761884. Whole genome *de novo* assemblies are available on GenBank under the same BioProject accession number. Filtered dataset of 7148 *bla*$_{\mathrm{NDM}}$ bearing contigs is available on Figshare: 10.5522/04/16594784

Code availability

All software used in this research are listed in Methods. An implementation of the algorithm used to track structural variations around *bla*$_{\mathrm{NDM}}$ is available at https://github.com/macman123/track_structural_variants.

References

1. Wu, W. *et al.* NDM metallo-β-lactamases and their bacterial producers in health care settings. *Clinical Microbiology Reviews* vol. 32 (2019).
2. Yong, D. *et al.* Characterization of a new metallo-β-lactamase gene, bla NDM-1, and a novel erythromycin esterase gene carried on a unique genetic structure in Klebsiella pneumoniae sequence type 14 from India. *Antimicrob. Agents Chemother.* **53**, 5046–5054 (2009).
3. Struelens, M. J. *et al.* New Delhi metallo-beta-lactamase 1–producing Enterobacteriaceae: emergence and response in Europe. *Eurosurveillance* **15**, 19716 (2010).
4. Kumarasamy, K. K. *et al.* Emergence of a new antibiotic resistance mechanism in India, Pakistan, and the UK: A molecular, biological, and epidemiological study. *Lancet Infect. Dis.* **10**, 597–602 (2010).
5. Poirel, L., Dortet, L., Bernabeu, S. & Nordmann, P. Genetic Features of bla NDM-1-Positive Enterobacteriaceae. *Antimicrob. Agents Chemother.* **55**, 5403–5407 (2011).
6. Castanheira, M. *et al.* Early dissemination of NDM-1- and OXA-181-producing Enterobacteriaceae in Indian hospitals: Report from the SENTRY Antimicrobial Surveillance Program, 2006-2007. *Antimicrob. Agents Chemother.* **55**, 1274–1278 (2011).
7. Jones, L. S. *et al.* Plasmid carriage of blaNDM-1in clinical Acinetobacter baumannii isolates from India. *Antimicrob. Agents Chemother.* **58**, 4211–4213 (2014).
8. Roca, I. *et al.* Molecular characterization of NDM-1-producing Acinetobacter pittii isolated from Turkey in 2006. *J. Antimicrob. Chemother.* **69**, 3437–3438 (2014).
9. Poirel, L., Bonnin, R. A. & Nordmann, P. Analysis of the resistome of a multidrug-resistant NDM-1-producing Escherichia coli strain by high-throughput genome sequencing. *Antimicrob. Agents Chemother.* **55**, 4224–4229 (2011).
10. Toleman, M. A., Spencer, J., Jones, L. & Walsha, T. R. bla NDM-1 is a chimera likely constructed in Acinetobacter baumannii. *Antimicrob. Agents Chemother.* **56**, 2773–2776 (2012).
11. Partridge, S. R. & Iredell, J. R. Genetic Contexts of bla NDM-1. *Antimicrobial Agents and Chemotherapy* vol. 56 6065–6067 (2012).
12. Partridge, S. R. & Iredell, J. R. Genetic Contexts of bla NDM-1. *Antimicrobial Agents and Chemotherapy* vol. 56 6065–6067 (2012).
13. Toleman, M. A., Bennett, P. M. & Walsh, T. R. ISCR Elements: Novel Gene-Capturing Systems of the 21st Century? *Microbiol. Mol. Biol. Rev.* **70**, 296–316 (2006).
14. Ilyina, T. S. Mobile ISCR elements: Structure, functions, and role in emergence, increase, and spread of blocks of bacterial multiple antibiotic resistance genes. *Molecular Genetics, Microbiology and Virology* vol. 27 135–146 (2012).
15. Poirel, L. *et al.* Tn125-related acquisition of blaNDM-like genes in Acinetobacter baumannii. *Antimicrob. Agents Chemother.* **56**, 1087–1089 (2012).
16. Sekizuka, T. *et al.* Complete Sequencing of the blaNDM-1-Positive IncA/C Plasmid from Escherichia coli ST38 Isolate Suggests a Possible Origin from Plant Pathogens. *PLoS One* **6**, e25334 (2011).
17. McArthur, A. G. *et al.* The comprehensive antibiotic resistance database. *Antimicrob. Agents Chemother.* **57**, 3348–3357 (2013).

18. Basu, S. Variants of the New Delhi metallo-β-lactamase: New kids on the block. *Future Microbiology* vol. 15 465–467 (2020).

19. Baraniak, A. *et al.* NDM-producing Enterobacteriaceae in Poland, 2012–14: inter-regional outbreak of *Klebsiella pneumoniae* ST11 and sporadic cases. *J. Antimicrob. Chemother.* **71**, 85–91 (2016).

20. Rahman, M. *et al.* Prevalence and Molecular Characterization of New Delhi Metallo-Beta-Lactamases in Multidrug-Resistant *Pseudomonas aeruginosa* and *Acinetobacter baumannii* from India. *Microb. Drug Resist.* **24**, 792–798 (2018).

21. Hu, H. *et al.* Novel plasmid and its variant harboring both a bla(NDM-1) gene and type IV secretion system in clinical isolates of Acinetobacter lwoffii. *Antimicrob. Agents Chemother.* **56**, 1698–702 (2012).

22. Yang, Q. *et al.* Dissemination of NDM-1-producing Enterobacteriaceae mediated by the IncX3-type plasmid. *PLoS One* **10**, (2015).

23. Wailan, A. M. *et al.* Genetic contexts of blaNDM-1 in patients carrying multiple NDM-producing strains. *Antimicrob. Agents Chemother.* **59**, 7405–7410 (2015).

24. Rasheed, J. K. *et al.* New Delhi Metallo-β-Lactamase–producing Enterobacteriaceae, United States. *Emerg. Infect. Dis.* **19**, 870 (2013).

25. Poirel, L. *et al.* Tn125-related acquisition of blaNDM-like genes in Acinetobacter baumannii. *Antimicrob. Agents Chemother.* **56**, 1087–1089 (2012).

26. Campos, J. C. *et al.* Characterization of Tn3000, a Transposon Responsible for blaNDM-1 Dissemination among Enterobacteriaceae in Brazil, Nepal, Morocco, and India. *Antimicrob. Agents Chemother.* **59**, 7387–95 (2015).

27. Feng, Y., Liu, L., McNally, A. & Zong, Z. Coexistence of two blaNDM-5 genes on an IncF plasmid as revealed by nanopore sequencing. *Antimicrob. Agents Chemother.* **62**, (2018).

28. Zhao, Q.-Y. *et al.* IS 26 Is Responsible for the Evolution and Transmission of bla NDM -Harboring Plasmids in Escherichia coli of Poultry Origin in China . *mSystems* (2021) doi:10.1128/MSYSTEMS.00646-21.

29. Lynch, T. *et al.* Molecular evolution of a klebsiella pneumoniae st278 isolate harboring blandm-7 and involved in nosocomial transmission. *J. Infect. Dis.* **214**, 798–806 (2016).

30. Huang, T. W. *et al.* Copy Number Change of the NDM-1 Sequence in a Multidrug-Resistant Klebsiella pneumoniae Clinical Isolate. *PLoS One* **8**, 1–12 (2013).

31. Datta, S. *et al.* Spread and exchange of bla NDM-1 in hospitalized neonates: role of mobilizable genetic elements. *Eur. J. Clin. Microbiol. Infect. Dis.* **36**, 255–265 (2017).

32. Krahn, T. *et al.* Intraspecies transfer of the chromosomal Acinetobacter baumannii blaNDM-1 carbapenemase gene. *Antimicrob. Agents Chemother.* **60**, 3032–3040 (2016).

33. González, L. J. *et al.* Membrane anchoring stabilizes and favors secretion of New Delhi metallo-β-lactamase. *Nat. Chem. Biol.* **12**, 516–522 (2016).

34. Chatterjee, S., Mondal, A., Mitra, S. & Basu, S. Acinetobacter baumannii transfers the blaNDM-1 gene via outer membrane vesicles. *J. Antimicrob. Chemother.* **72**, 2201–2207 (2017).

35. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-45 (2016).

36. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **41**, (2013).

37. Zhou, Z., Alikhan, N. F., Mohamed, K., Fan, Y. & Achtman, M. The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. *Genome Res.* **30**, 138–152 (2020).

38. Souvorov, A., Agarwala, R. & Lipman, D. J. SKESA: Strategic k-mer extension for scrupulous assemblies. *Genome Biol.* **19**, 153 (2018).

39. Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K. & Tolstoy, I. RefSeq microbial genomes database: New representation and annotation strategy. *Nucleic Acids Res.* **42**, D553–D559 (2014).

40. Chavda, K. D. *et al.* Comprehensive genome analysis of carbapenemase-producing Enterobacter spp.: New insights into phylogeny, population structure, and resistance mechanisms. *MBio* **7**, (2016).

41. Ashton, P. M. *et al.* Identification of Salmonella for public health surveillance using whole genome sequencing. *PeerJ* **2016**, e1752 (2016).

42. Sahl, J. W. *et al.* Phylogenetic and genomic diversity in isolates from the globally distributed Acinetobacter baumannii ST25 lineage. *Sci. Rep.* **5**, (2015).

43. Bonnin, R. A. *et al.* Dissemination of New Delhi metallo-β-lactamase-1-producing Acinetobacter baumannii in Europe. *Clin. Microbiol. Infect.* **18**, E362–E365 (2012).

44. Carattoli, A. *et al.* In Silico Detection and Typing of Plasmids using PlasmidFinder and Plasmid Multilocus Sequence Typing. **58**, 3895–3903 (2014).

45. Roach, D. *et al.* Whole Genome Sequencing of Peruvian Klebsiella pneumoniae Identifies Novel Plasmid Vectors Bearing Carbapenem Resistance Gene NDM-1. *Open Forum Infect. Dis.* **7**, (2020).

46. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

47. Ma, B., Tromp, J. & Li, M. PatternHunter: Faster and more sensitive homology search. *Bioinformatics* **18**, 440–445 (2002).

48. Schnetz, K. & Rak, B. IS5: A mobile enhancer of transcription in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 1244–1248 (1992).

49. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. & Chandler, M. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.* **34**, D32–D36 (2006).

50. Harmer, C. J., Pong, C. H. & Hall, R. M. Structures bounded by directly-oriented members of the IS26 family are pseudo-compound transposons. *Plasmid* vol. 111 102530 (2020).

51. Harmer, C. J., Moran, R. A. & Hall, R. M. 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* **5**, (2014).

52. Li, J. *et al.* Sequential Isolation in a Patient of Raoultella planticola and Escherichia coli Bearing a Novel ISCR1 Element Carrying blaNDM-1. *PLoS One* **9**, e89893 (2014).

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

54. Harmer, C. J. & Hall, R. M. An analysis of the IS6/IS26 family of insertion sequences: Is it a single family? *Microb. Genomics* **5**, (2019).

55. He, S. *et al.* Insertion sequence IS26 reorganizes plasmids in clinically isolated multidrug-resistant bacteria by replicative transposition. *MBio* **6**, 1–14 (2015).

56. He, D. D. *et al.* Antimicrobial resistance-encoding plasmid clusters with heterogeneous MDR regions driven by IS26 in a single Escherichia coli isolate. *J. Antimicrob. Chemother.* **74**, 1511–1516 (2019).

57. Wang, R. *et al.* The global distribution and spread of the mobilized colistin resistance gene mcr-1. *Nat. Commun.* **9**, 1–9 (2018).

58. Acman, M., van Dorp, L., Santini, J. M. & Balloux, F. Large-scale network analysis captures biological features of bacterial plasmids. *Nat. Commun.* **11**, 1–11 (2020).

59. Shaw, L. *et al.* Niche and local geography shape the pangenome of wastewater- and livestock-associated Enterobacteriaceae. 1–23 (2020) doi:10.1101/2020.07.23.215756.

60. Redondo-Salvo, S. *et al.* Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nat. Commun. 2020 111* **11**, 1–13 (2020).

61. Sheppard, A. E. *et al.* Nested Russian Doll-Like Genetic Mobility Drives Rapid Dissemination of the Carbapenem Resistance Gene blaKPC. *Antimicrob. Agents Chemother.* **60**, 3767–3778 (2016).

62. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**, D61–D65 (2007).

63. Luhmann, N., Holley, G. & Achtman, M. BlastFrost: Fast querying of 100,000s of bacterial genomes in Bifrost graphs. *bioRxiv* 1–24 (2020) doi:10.1101/2020.01.21.914168.

64. Bradley, P., den Bakker, H. C., Rocha, E. P. C., McVean, G. & Iqbal, Z. Ultrafast search of all deposited bacterial and viral genomic data. *Nat. Biotechnol.* **37**, 152–159 (2019).

65. Wang, R. *et al.* The prevalence of colistin resistance in Escherichia coli and Klebsiella pneumoniae isolated from food animals in China: coexistence of mcr-1 and blaNDM with low fitness cost. *Int. J. Antimicrob. Agents* **51**, 739–744 (2018).

66. Wang, Q. *et al.* Phenotypic and Genotypic Characterization of Carbapenem-resistant Enterobacteriaceae: Data from a Longitudinal Large-scale CRE Study in China (2012-2016). *Clin. Infect. Dis.* **67**, S196–S205 (2018).

67. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**, 1–22 (2017).

68. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–77 (2012).

69. Chamberlain, S. A. & Szöcs, E. taxize: taxonomic search and retrieval in R. *F1000Research* **191**, 1–28 (2013).

70. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

71. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).

72. Orlek, A. *et al.* Plasmid classification in an era of whole-genome sequencing: Application in studies of antibiotic resistance epidemiology. *Frontiers in Microbiology* vol. 8 1–10 (2017).

73. Acman, M., van Dorp, L., Santini, J. M. & Balloux, F. Large-scale network analysis captures biological features of bacterial plasmids. *Nat. Commun.* **11**, 1–11 (2020).

74. Ondov, B. D. *et al.* Mash : fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 1–14 (2016) doi:10.1186/s13059-016-0997-x.

75. Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 1118–1123 (2008).

76. Shannon, P. *et al.* Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

77. Bushnell, B. BBMap: A Fast, Accurate, Splice-Aware Aligner. https://www.osti.gov/biblio/1241166 (2014).

78. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).

79. Okonechnikov, K., Golosova, O., Fursov, M. & team, the U. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* **28**, 1166–1167 (2012).

80. Rieux, A. & Balloux, F. Inferences from tip-calibrated phylogenies: A review and a practical guide. *Mol. Ecol.* **25**, 1911–1924 (2016).

81. Rambaut, A., Lam, T. T., Carvalho, L. M. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, 1–7 (2016).

82. Duchene, S. *et al.* Bayesian Evaluation of Temporal Signal In Measurably Evolving Populations. *bioRxiv* (2019) doi:10.1101/810697.

83. Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **15**, e1006650 (2019).

84. Didelot, X., Croucher, N. J., Bentley, S. D., Harris, S. R. & Wilson, D. J. Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Res.* **46**, 1–11 (2018).

85. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

86. Plummer, M., Best, N., Cowles, K. & Vines, K. CODA: convergence diagnosis and output analysis for MCMC - Open Research Online. *R News* **6**, 7–11 (2006).

87. Padgham, M. & Sumner, M. D. geodist: Fast, Dependency-Free Geodesic Distance Calculations. (2020).

88. Zhao, X. BinDash, software for fast genome distance estimation on a typical personal laptop. *Bioinformatics* **35**, 671–673 (2019).

89. Oksanen, J. *et al.* vegan: Community Ecology Package. (2019).

Figures

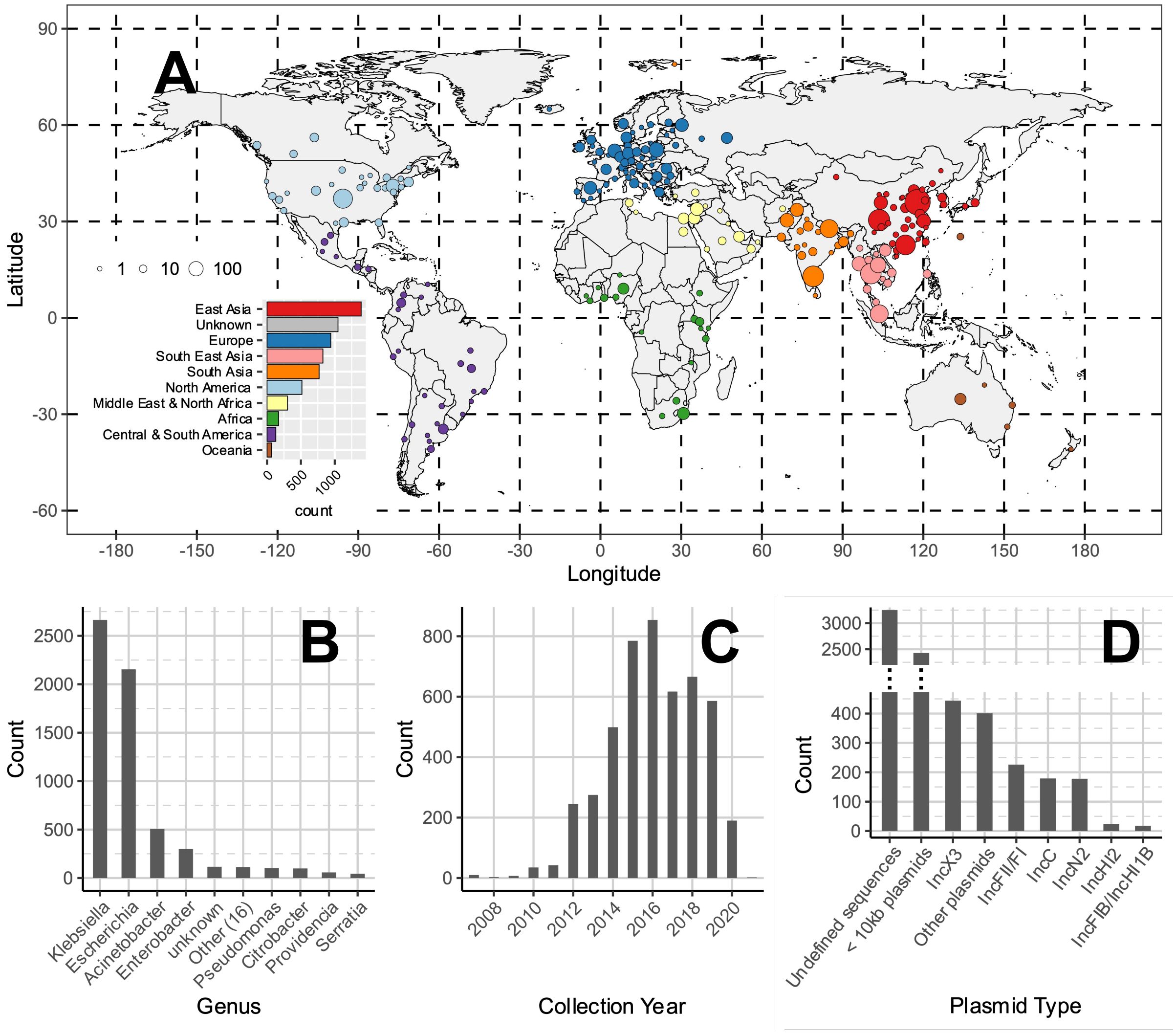


Figure 1. Composition of the global dataset of 6,155 NDM-positive samples. (**A**) Geographic distribution of *bla*$_{\text{NDM}}$-positive assemblies. Points are coloured by geographic region and their size reflects the number of samples they encompass. **(B)** Distribution of host bacterial genera of NDM-positive samples. **(C)** Distribution of sample collection years. **(D)** Distribution of contigs according to the plasmid backbone.

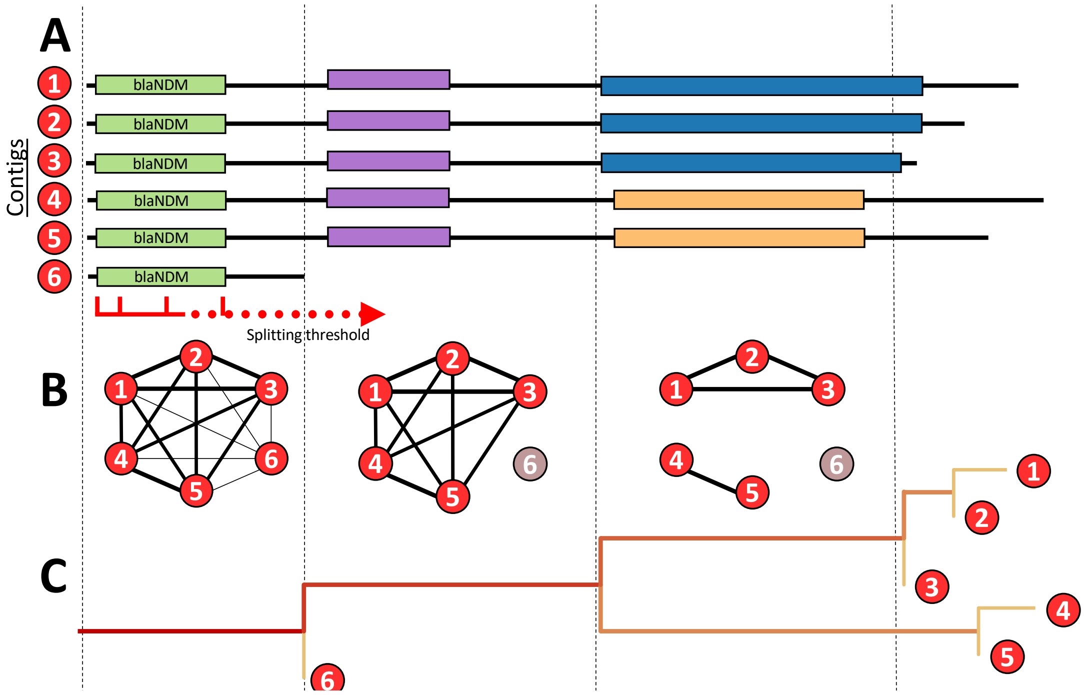
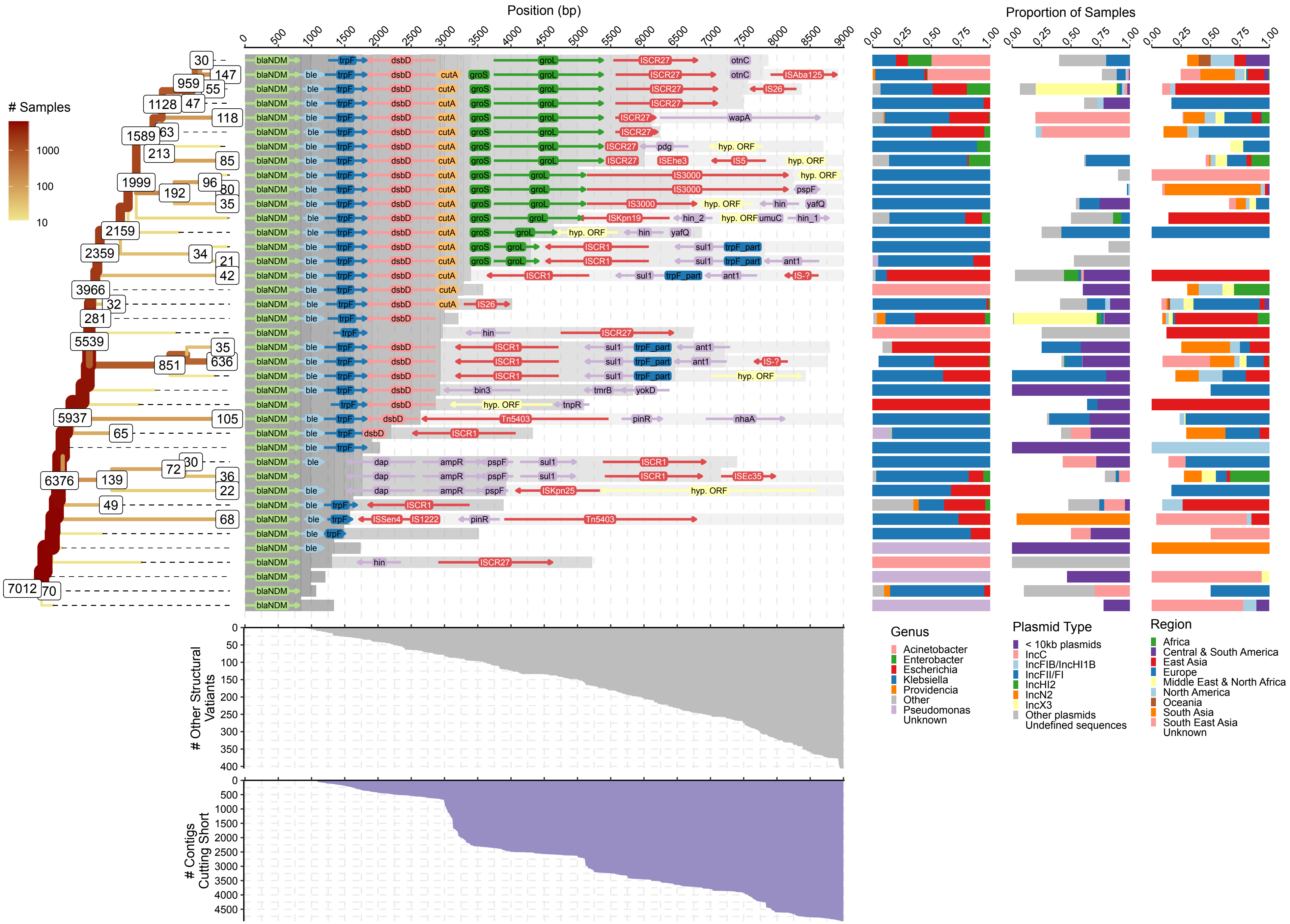


Figure 2. Schematic representation of the tracking algorithm splitting structural variants upstream or downstream of *bla*$_{\text{NDM}}$ gene. (A) A pairwise BLAST search is performed on all NDM-positive contigs. Starting from *bla*$_{\text{NDM}}$ and continuing downstream or upstream, the inspected region is gradually increased using the 'splitting threshold'. **(B)** At each step, a graph is constructed connecting contigs (nodes) that share a BLAST hit with a minimum length as given by the 'splitting threshold'. Contigs which have the same structural variant at the certain position of the threshold belong to the same graph component, while the short contigs are singled out. **(C)** The splitting is visualized as a tree where branch lengths are scaled to match the position within the sequence, and the thickness and the colour intensity of the branches correspond to the number of sequences carrying the homology. For more detailed explanation of the algorithm please refer to the Methods section.



Position (bp)
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
6500
7000
7500
8000
8500
9000
Proportion of Samples
0.00
0.25
0.50
0.75
1.00
# Samples
1000
100
10
30
147
959
55
1128
47
118
1589
63
213
85
1999
96
192
80
35
2159
2359
34
21
42
3966
32
281
5539
35
851
636
5937
105
65
6376
72
30
139
36
22
49
68
7012
70
blaNDM
ble
trpF
dsbD
cutA
groS
groL
ISCR27
otnC
ISAba125
IS26
wapA
pdg
hyp. ORF
ISEhe3
IS5
IS3000
pspF
hin
yafQ
ISKpn19
hin_2
umuC
hin_1
ISCR1
sul1
trpF_part
ant1
IS-?
bin3
tmrB
yokD
tnpR
Tn5403
pinR
nhaA
dap
ampR
ISEc35
ISKpn25
ISSen4
IS1222
# Other Structural Vatiants
0
50
100
150
200
250
300
350
400
# Contigs Cutting Short
0
500
1000
1500
2000
2500
3000
3500
4000
4500
Genus
Acinetobacter
Enterobacter
Escherichia
Klebsiella
Providencia
Other
Pseudomonas
Unknown
Plasmid Type
< 10kb plasmids
IncC
IncFIB/IncHI1B
IncFII/FI
IncHI2
IncN2
IncX3
Other plasmids
Undefined sequences
Region
Africa
Central & South America
East Asia
Europe
Middle East & North Africa
North America
Oceania
South Asia
South East Asia
Unknown

Figure 3. Splitting of structural variants downstream of *bla*$_{\text{NDM}}$*.* The ‘splitting' tree for the most common (i.e., ≥ 10 contigs) structural variants is shown on the left-hand side. The labels on the nodes indicate the number of contigs remaining on each branch. Labels of (yellow) branches with <20 contigs are not shown. The other contigs either belong to other structural variants or were removed due to being too short in length. The number of contigs cutting short is indicated by the area chart at the bottom. Similarly, the number of less common structural variants is indicated by the upper area chart. Genome annotations provided by the Prokka and Roary pipelines of the most common structural variants are shown in the middle of the figure. The homologous regions among structural variants are indicated by the grey shading. Some of the structural variants and branches were intentionally cut short even though their contigs were of sufficient size or longer. This was done to prevent excessive bifurcation and to make the tree easier to interpret. In particular, branches with more than 75% of contigs lost due to variation and short length were truncated. The distribution of genera, plasmid backbones and geographical regions of samples that belong to a each of the common structural variant is shown on the right-hand side.

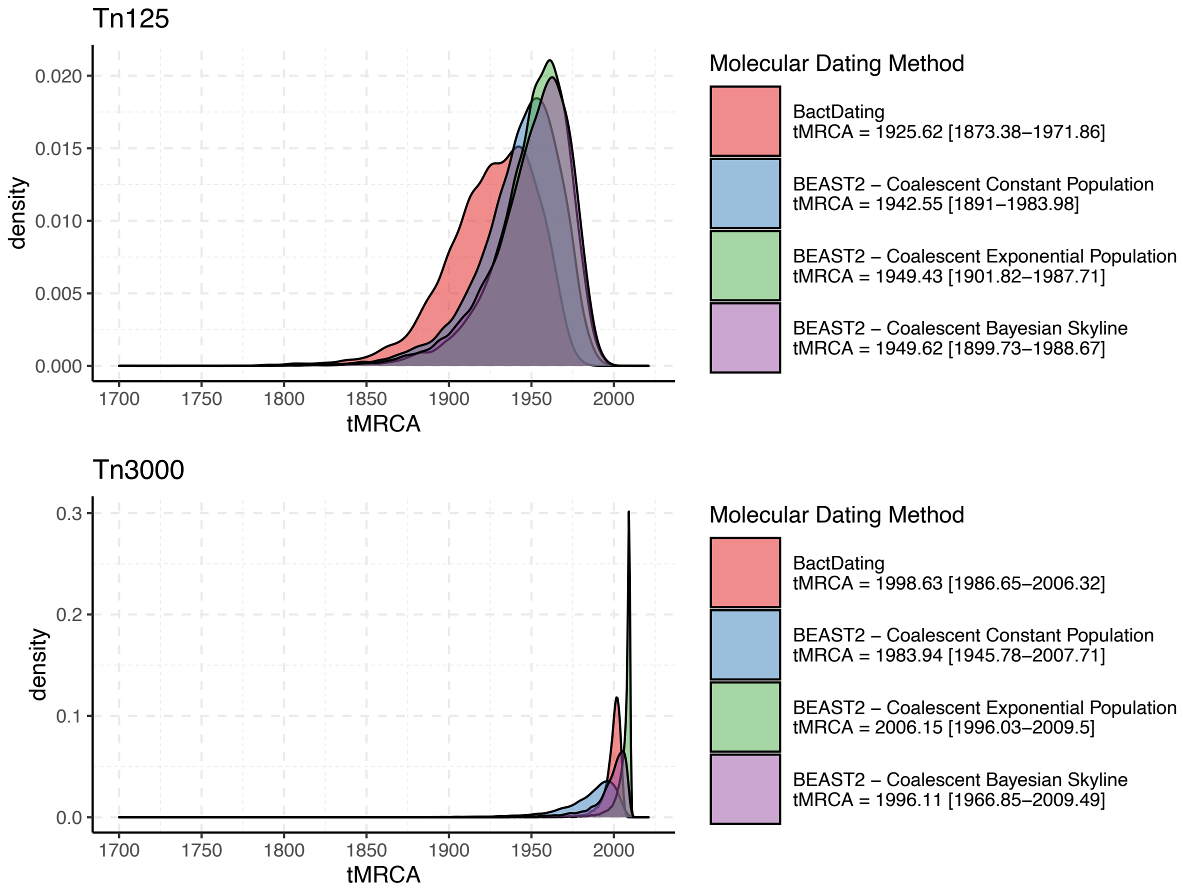
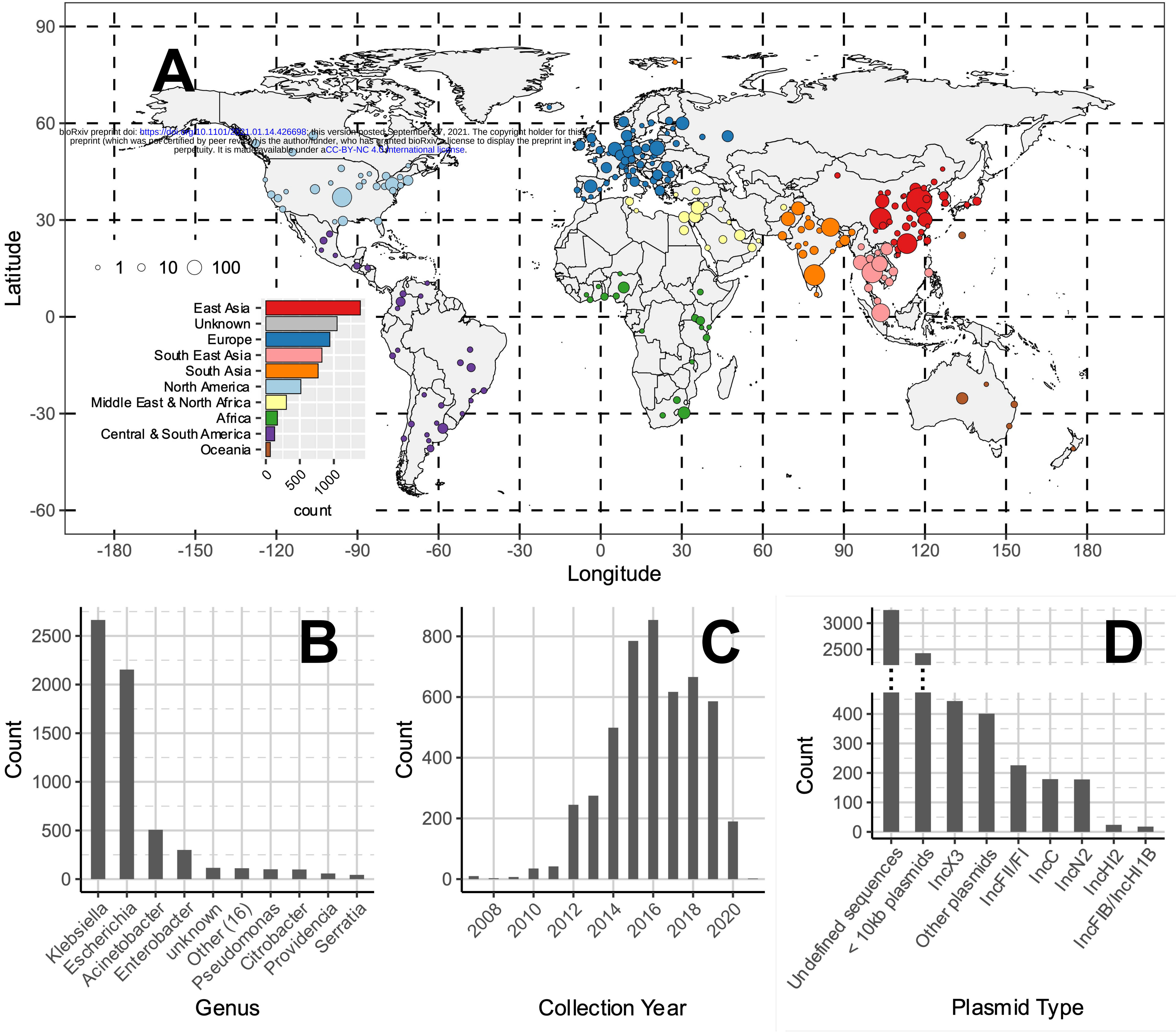
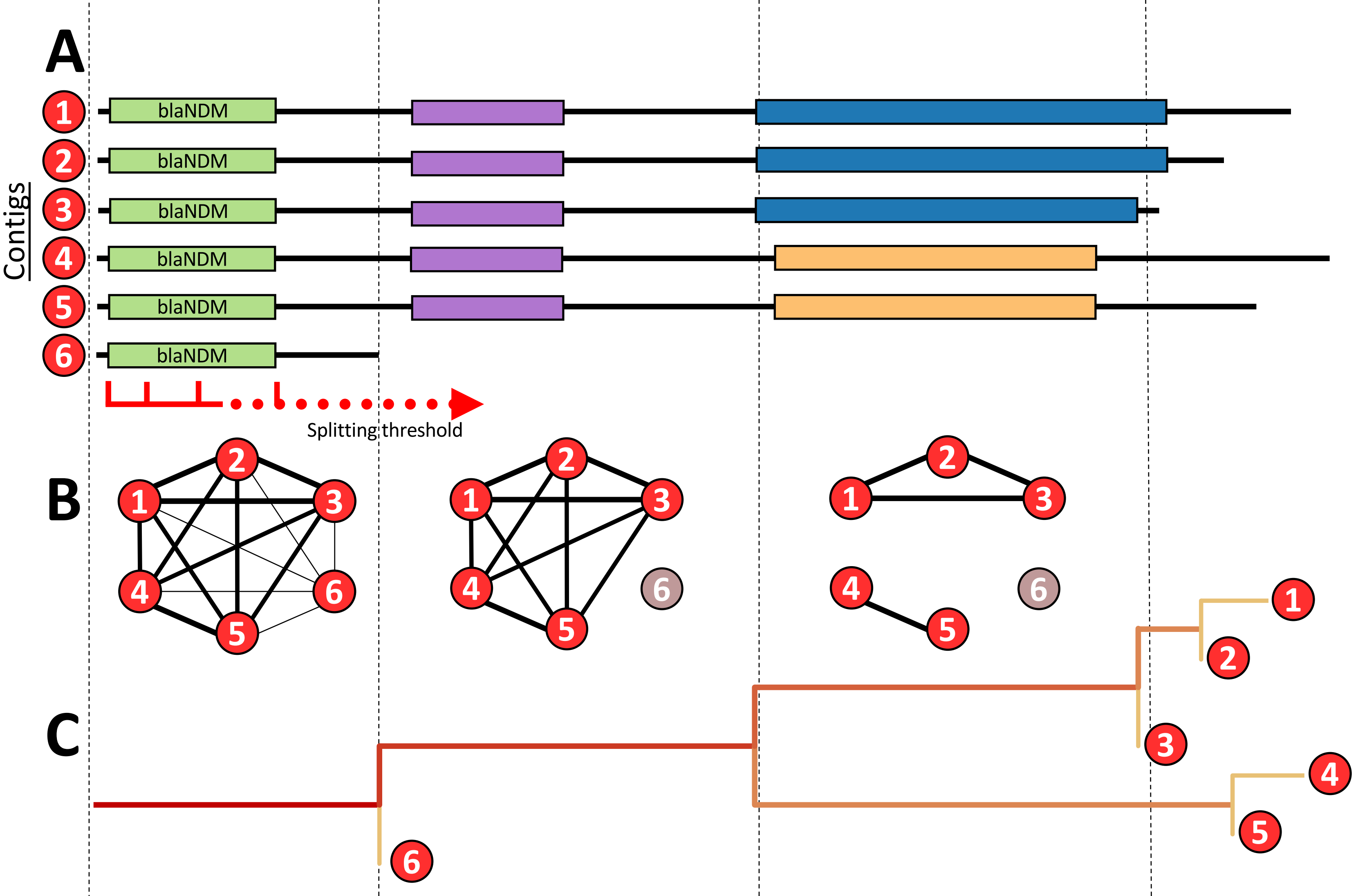
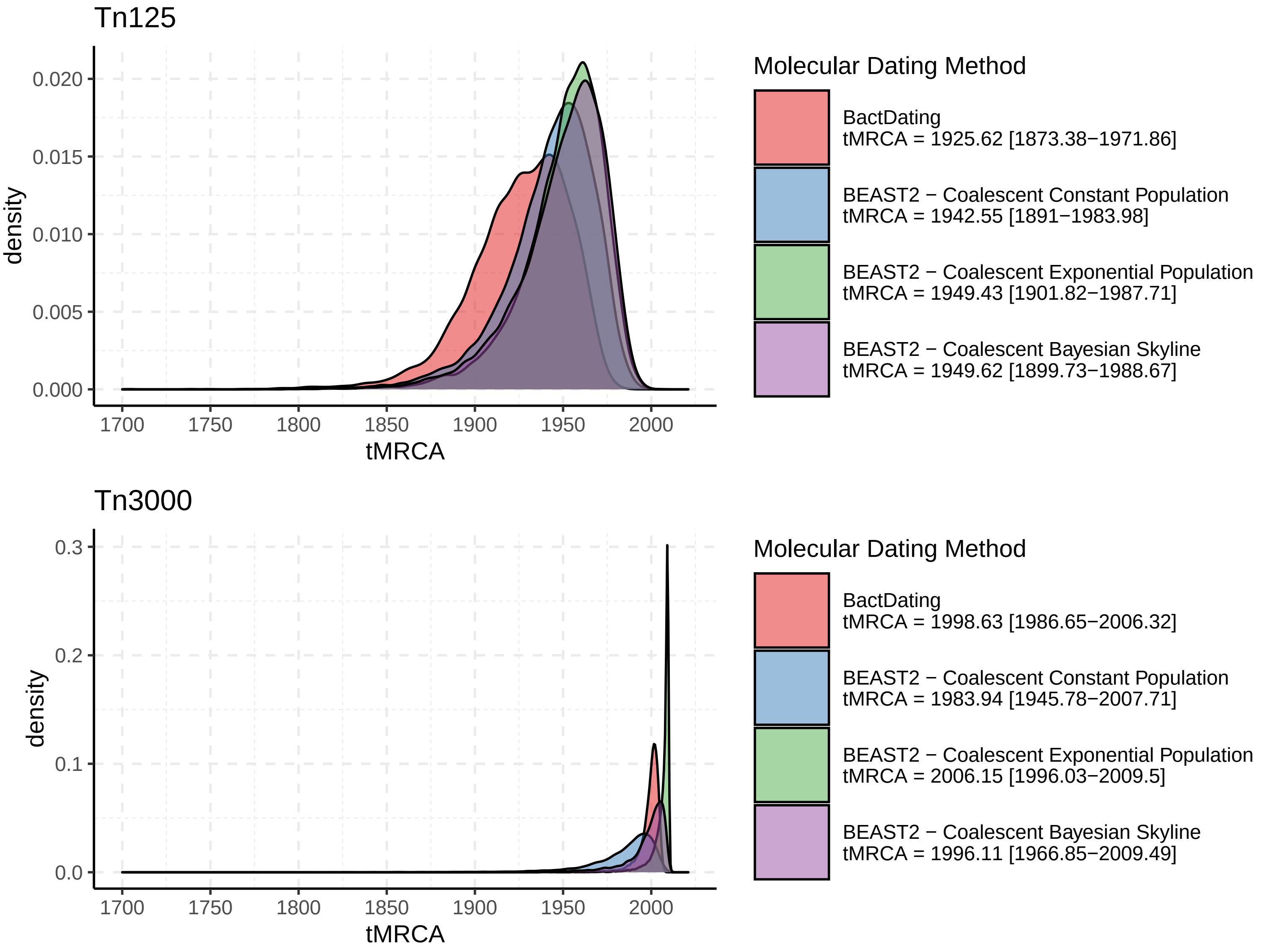


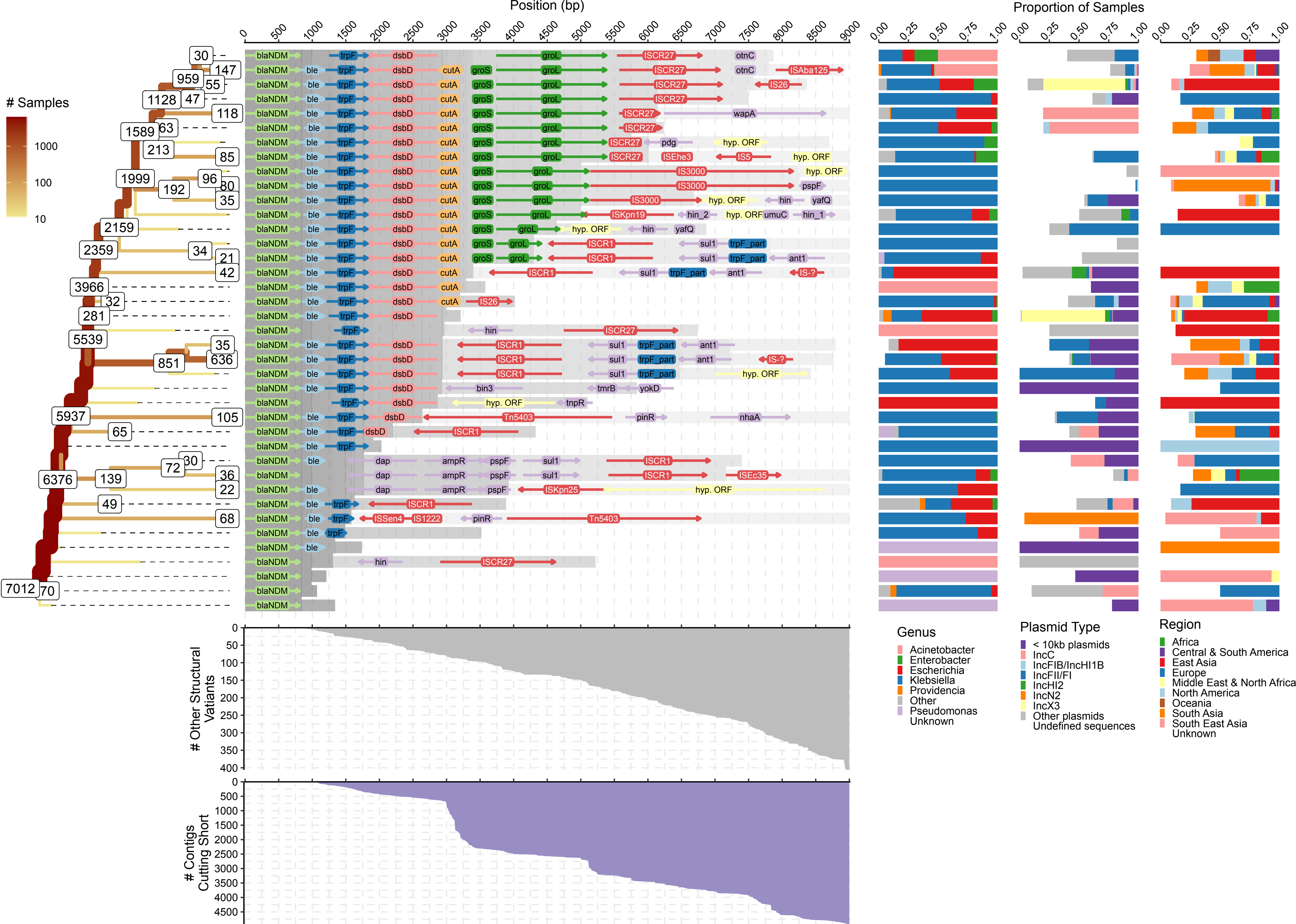
Figure 4. Posterior density distributions of ancestral sequence age (i.e., root height) for the *Tn125* (A) and *Tn3000* (B) transposons. The ancestral sequence emergence was estimated using two Bayesian tip-dating approaches implemented in BactDating and BEAST2. Three different population growth priors were used in case of BEAST2: Coalescent Constant Population, Coalescent Exponential Population, and Coalescent Bayesian Skyline as given by the colour scheme and legend at right. Median estimates with 95% highest density interval (HDI) are provided in the panel legends.


A
Latitude
90
60
30
0
-30
-60
Longitude
-180
-150
-120
-90
-60
-30
0
30
60
90
120
150
180
1
10
100
East Asia
Unknown
Europe
South East Asia
South Asia
North America
Middle East & North Africa
Africa
Central & South America
Oceania
0
500
1000
count
B
Count
2500
2000
1500
1000
500
0
Klebsiella
Escherichia
Acinetobacter
Enterobacter
unknown
Other (16)
Pseudomonas
Citrobacter
Providencia
Serratia
Genus
C
Count
800
600
400
200
0
2008
2010
2012
2014
2016
2018
2020
Collection Year
D
Count
3000
2500
400
300
200
100
0
Undefined sequences
< 10kb plasmids
IncX3
Other plasmids
IncFII/FI
IncC
IncN2
IncHI2
IncFIB/IncHI1B
Plasmid Type


A
Contigs
1
2
3
4
5
6
blaNDM
blaNDM
blaNDM
blaNDM
blaNDM
blaNDM
Splitting threshold
B
C



Tn125
Molecular Dating Method
BactDating
tMRCA = 1925.62 [1873.38−1971.86]
BEAST2 − Coalescent Constant Population
tMRCA = 1942.55 [1891−1983.98]
BEAST2 − Coalescent Exponential Population
tMRCA = 1949.43 [1901.82−1987.71]
BEAST2 − Coalescent Bayesian Skyline
tMRCA = 1949.62 [1899.73−1988.67]
density
0.020
0.015
0.010
0.005
0.000
1700
1750
1800
1850
1900
1950
2000
tMRCA
Tn3000
Molecular Dating Method
BactDating
tMRCA = 1998.63 [1986.65−2006.32]
BEAST2 − Coalescent Constant Population
tMRCA = 1983.94 [1945.78−2007.71]
BEAST2 − Coalescent Exponential Population
tMRCA = 2006.15 [1996.03−2009.5]
BEAST2 − Coalescent Bayesian Skyline
tMRCA = 1996.11 [1966.85−2009.49]
density
0.3
0.2
0.1
0.0
1700
1750
1800
1850
1900
1950
2000
tMRCA



Position (bp)
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
6500
7000
7500
8000
8500
9000
Proportion of Samples
0.00
0.25
0.50
0.75
1.00
# Samples
1000
100
10
30
147
959
55
1128
47
118
1589
63
213
85
1999
96
192
80
35
2159
2359
34
21
42
3966
32
281
5539
35
851
636
5937
105
65
6376
72
30
139
36
22
49
68
7012
70
blaNDM
ble
trpF
dsbD
cutA
groS
groL
ISCR27
otnC
ISAba125
IS26
wapA
pdg
hyp. ORF
ISEhe3
IS5
IS3000
pspF
hin
yafQ
ISKpn19
hin_2
umuC
hin_1
ISCR1
sul1
trpF_part
ant1
IS-?
bin3
tmrB
yokD
tnpR
Tn5403
pinR
nhaA
dap
ampR
ISEc35
ISKpn25
ISSen4
IS1222
# Other Structural Vatiants
0
50
100
150
200
250
300
350
400
# Contigs Cutting Short
0
500
1000
1500
2000
2500
3000
3500
4000
4500
Genus
Acinetobacter
Enterobacter
Escherichia
Klebsiella
Providencia
Other
Pseudomonas
Unknown
Plasmid Type
< 10kb plasmids
IncC
IncFIB/IncHI1B
IncFII/FI
IncHI2
IncN2
IncX3
Other plasmids
Undefined sequences
Region
Africa
Central & South America
East Asia
Europe
Middle East & North Africa
North America
Oceania
South Asia
South East Asia
Unknown