



HAL
open science

In silico prediction of the enzymes involved in the degradation of the herbicide molinate by *Gulosibacter molinativorax* ON4T

A. Lopes, E. Bunin, A. Viana, H. Froufe, A. Muñoz-Merida, D. Pinho, J. Figueiredo, C. Barroso, I. Vaz-Moreira, X. Bellanger, et al.

► To cite this version:

A. Lopes, E. Bunin, A. Viana, H. Froufe, A. Muñoz-Merida, et al.. In silico prediction of the enzymes involved in the degradation of the herbicide molinate by *Gulosibacter molinativorax* ON4T. *Scientific Reports*, 2022, 12 (1), pp.15502. 10.1038/s41598-022-18732-5 . hal-03779503

HAL Id: hal-03779503

<https://hal.univ-lorraine.fr/hal-03779503>

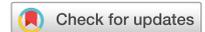
Submitted on 4 Oct 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



OPEN

In silico prediction of the enzymes involved in the degradation of the herbicide molinate by *Gulosibacter molinivorax* ON4^T

A. R. Lopes^{1,2}, E. Bunin¹, A. T. Viana¹, H. Froufe², A. Muñoz-Merida³, D. Pinho², J. Figueiredo¹, C. Barroso^{2,4}, I. Vaz-Moreira⁵, X. Bellanger⁶, C. Egas^{2,4} & O. C. Nunes¹✉

Gulosibacter molinivorax ON4^T is the only known organism to produce molinate hydrolase (MolA), which catalyses the breakdown of the thiocarbamate herbicide into azepane-1-carboxylic acid (ACA) and ethanethiol. A combined genomic and transcriptomic strategy was used to fully characterize the strain ON4^T genome, particularly the *molA* genetic environment, to identify the potential genes encoding ACA degradation enzymes. Genomic data revealed that *molA* is the only catabolic gene of a novel composite transposon (Tn6311), located in a novel low copy number plasmid (pARLON1) harbouring a putative T4SS of the class FATA. pARLON1 had an ANI value of 88.2% with contig 18 from *Agrococcus casei* LMG 22410^T draft genome. Such results suggest that pARLON1 is related to genomic elements of other *Actinobacteria*, although Tn6311 was observed only in strain ON4^T. Furthermore, genomic and transcriptomic data demonstrated that the genes involved in ACA degradation are chromosomal. Based on their overexpression when growing in the presence of molinate, the enzymes potentially involved in the heterocyclic ring breakdown were predicted. Among these, the activity of a protein related to caprolactone hydrolase was demonstrated using heterologous expression. However, further studies are needed to confirm the role of the other putative enzymes.

Biocatalysis is a crucial component of white or industrial biotechnology, relying on the broad catalytic activity of enzymes leading to the production of value-added compounds (e.g., detergents, cosmetics), or catalyzing the degradation of several environmental contaminants (e.g., pesticides, hydrocarbons)^{1–3}. Genome and transcriptome sequencing projects are critical to the burst of the biocatalysis field, enabling the identification of genes encoding proteins putatively involved in the production/degradation of the target compound(s), which can be further characterized to develop a biocatalytic solution^{1,4,5}.

The biotechnological potential of *Actinobacteria* is known^{1,3}, yet there is still much to unveil⁶. *Gulosibacter molinivorax* ON4^T, an actinobacterium of the family *Microbacteriaceae*⁷, is an example. It was isolated from a five-membered enrichment culture (strains ON1–ON5) able to mineralize the thiocarbamate herbicide molinate^{8,9}. Strain ON4^T produces molinate hydrolase (MolA), responsible for breaking molinate into azepane-1-carboxylic acid (ACA) and ethanethiol^{8,10}. Strain ON4^T does not further use ethanethiol, but strains ON1 and ON2 consume this metabolite. In contrast, ACA supports the growth of strain ON4^T and other culture members⁸. Hence, molinate mineralization occurs through the cooperation between strain ON4^T and the culture members able to deplete the sulfur metabolite⁸. Although the putative degradation pathway for molinate has been described⁸, MolA is the only enzyme involved in this pathway identified and characterized so far^{10–12}. Previous studies have shown the capacity of MolA and the recombinant mutant (Arg187Ala) thereof to degrade the thiocarbamate herbicides molinate and thiobencarb^{10–12}. The biocatalytic potential of the enzymes involved in the ACA degradation pathway in *G. molinivorax* ON4^T motivated this study.

¹LEPABE - Laboratory for Process Engineering, Environment, Biotechnology and Energy, Departamento de Engenharia Química, Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal. ²Next Generation Sequencing Unit, Biocant, BiocantPark, Núcleo 04, Lote 8, 3060-197 Cantanhede, Portugal. ³CIBIO – Research Centre in Biodiversity and Genetic Resources, InBIO, University of Porto, Rua Padre Armando Quintas, no. 7, 4485-661 Vairão, Portugal. ⁴Center for Neuroscience and Cell Biology, University of Coimbra, 3004-504 Coimbra, Portugal. ⁵Universidade Católica Portuguesa, CBQF – Centro de Biotecnologia e Química Fina – Laboratório Associado, Escola Superior de Biotecnologia, Rua Diogo Botelho 1327, 4169-005 Porto, Portugal. ⁶Université de Lorraine, CNRS, LCPME, 54000 Nancy, France. ✉email: opnunes@fe.up.pt

Genes encoding enzymes involved in a particular metabolic pathway, as in xenobiotic degradation, tend to be close to each other^{13–16}, often organized in catabolic operons with common regulation mechanisms^{17–20}. Operons encoding enzymes involved in the degradation of xenobiotics can be located either in plasmids or in the chromosomes. Examples of plasmid operons are those encoding enzymes involved in the degradation of naphthalene, toluene, or xylene, described in pseudomonads (e.g., NAH7, pDTG1^{14,17,18}, pWVO^{13,19}). In opposition, the operon involved in 2,4,6-trichlorophenol catabolism (*tcpABC* genes) is located in the chromosome, as suggested by the ability of *R. eutropha* JMP134 (pJP4) and of *R. eutropha* JMP222 (pJP4-cured JMP134 derivative) to degrade this pollutant²¹. Although primarily described as chromosomal^{20,22}, the operon involved in biphenyl catabolism (*bphABC* genes) in pseudomonads, and other distant related *Betaproteobacteria*, is carried by an integrative conjugative element, that can integrate the chromosome and/or plasmid of the host²³. In addition, genes involved in the same catabolic pathway are not always organized in catabolic operons and can be located in different regions of the chromosome²⁴, or sometimes in the chromosome and extrachromosomal replicons (e.g., plasmids). An example of complementarity between chromosome and plasmid genes encoding xenobiotic degradation is given for carbaryl (1-naphthyl N-methylcarbamate) by *Arthrobacter* sp. strain RC100²⁵.

This study aimed at a genomic characterization of *G. molinivorax* ON4^T, particularly at identifying the genes coding for the potential enzymes involved in the ACA degradation pathway. Based on the premise that the genes of this pathway would be in a polycistronic operon unit, we started by searching the *mola* genetic environment in the publicly available draft genome of *G. molinivorax* DSM13485^T (= LMG 21909^T = CCUG 49965^T = CIP 108515^T = ON4^T) generated by the DOE Joint Genome Institute (JGI, GenBank with the accession number AUDX00000000.1). However, it did not report the *mola* gene. We obtained the complete genome sequence of strain ON4^T and, for comparison purposes, the genome of a variant of *G. molinivorax* ON4^T unable to degrade molinate (strain ON4⁽⁻⁾, obtained by successively growing in a culture medium without the herbicide for 12 generations). As the genomic data did not reveal other catabolic genes potentially involved in ACA degradation in the vicinity of the *mola* gene, transcriptomic analyses were further used to identify these genes. Transcripts obtained when strain ON4^T grew in the presence and absence of the herbicide were compared and thoroughly analysed. We have identified a novel mobilizable plasmid and seven candidate enzymes involved in ACA mineralization. The future characterization of their *in vivo* activity will be of great value not only for fundamental microbiology, but to other emerging fields such as biocatalysis and synthetic biology.

Results and discussion

Strain ON4^T genome sequencing. The 454 pyrosequencing of the *G. molinivorax* ON4^T genome (ON4^T₄₅₄) generated 305,068 reads, comprising 133.7 million nucleotides (nt), which were assembled in 163 contigs (PXVE00000000). Interested in disclosing the genes in the vicinity of *mola*, analyses started with a BLAST search of the *mola* gene sequence against the 163 assembled contigs. The complete *mola* gene sequence (1398 bp) was located in a single contig (1957 bp) but revealed no other open reading frame (ORF) in its vicinity. The strain ON4^T genome was resequenced using PacBio (ON4^T_{PacBio}), which generated 102,855 polymerase reads, resulting in 184,306 subreads with a mean subread length of 7254 bp, and 1,337 billion nt. These reads were *de novo* assembled and two contigs of 3,465,658 bp (linear) (contig A—CP028426) and 37,013 bp (circular) (contig B—CP028427) were generated. The assembled genome had 98.6% completeness, 0.6% contamination (Table S1) and 64% average G+C content for each contig. This *in silico* G+C content value agrees with previously published data⁷. The ON4^T_{PacBio} genome assembly disclosed 3243 predicted protein-coding genes, most of them assigned to putative functions (75.9%) and matching 94.3% of the predicted protein-coding genes in the assembly of strain DSM13485^T (JGI, AUDX00000000.1) (Table 1, Fig. 1, Table S1). Furthermore, forty-nine tRNAs, three 16S rRNA, three 23S rRNA, and three 5S rRNA genes were annotated (Table 1). The occurrence of three 16S rRNA gene copies was further confirmed by determining the ratio between the 16S rRNA and the *recA* genes (single-copy gene) in both strains ON4^T and ON4⁽⁻⁾ (Table S2).

As expected, the data retrieved using both sequencing technologies (454 and PacBio) were very similar (Fig. 1). Nevertheless, the long reads obtained by PacBio sequencing allowed to sum up the 163 assembled ON4^T₄₅₄ contigs into two contigs only. In addition, the comparison of the two draft genomes allowed closing the contig A (Fig. S1) and identifying the *mola* gene in the contig B.

***mola* is the passenger gene of Tn6311.** The *mola* gene, located in contig B, was surrounded by two identical insertion sequences (IS) of the IS3 group and family (Fig. 2a), whose original transposase belongs to the DDE transposase family²⁶. The primary sequence of this IS was not identical to any IS described so far and was named IS*Gmo1* (<https://www-is.biotoul.fr/>). This structure, where a passenger gene (*mola*), which does not code for a transposition enzyme, is flanked by two IS (IS*Gmo1*), is typical of a composite transposon^{27,28}. This organization was confirmed by Sanger sequencing (data not shown). The novel composite transposon, named Tn6311, contained three copies of *mola*, each flanked upstream and downstream by IS*Gmo1* (a total of 4 IS*Gmo1* copies) (Fig. 2a). This tandem organization was primarily observed in specific PacBio raw reads (data not shown). Further, the three *mola* copies were confirmed in strain ON4^T using quantitative PCR, as indicated by the gene copy ratio between *mola* and a neighbour single-copy gene (*parA*) (Table S2).

As described for other IS3-like sequences, IS*Gmo1* has a total length of 1287 bp and contains two partially overlapping open reading frames, *orfA* (291 bp) and *orfB* (1080 bp), which encode the DNA-binding and the catalytic domain of the transposase, respectively^{27,29–31}. Some IS3 family members might assemble different functional protein domains from one DNA segment, using a mechanism known as recoding programmed ribosomal frameshifting. This mechanism involves a ribosomal slippage of 1 nt, occurring at the heptanucleotide sequence A AAA AAG position, between *orfA* and the downstream fragment, *orfB*^{26,31}. The slippage generates a –1 frameshift resulting in a fused product of the upstream frame (DNA-binding domain) and the downstream

| Feature/attribute | Value |
|-----------------------------------|--------------|
| Genome size (bp) | 3,502,671 |
| G+C content (%) | 64 |
| DNA coding region (bp) | 3,152,404 |
| Total genes (No.) | 3301 |
| Total contig number (No.) | 2 |
| Contig A (CP028426) | 3,465,658 bp |
| 5S rRNA | 3 |
| 16S rRNA | 3 |
| 23S rRNA | 3 |
| tRNA | 49 |
| Protein-coding genes | 3196 |
| Proteins with function prediction | 2459 |
| Proteins with gene name | 2347 |
| Proteins with enzyme (EC number) | 1390 |
| Proteins with GO terms | 2406 |
| Proteins assigned to a pathway | 277 |
| Contig B (CP028427) | 37,013 bp |
| Protein-coding genes | 47 |
| Proteins with function prediction | 27 |
| Proteins with gene name | 20 |
| Proteins with enzyme (EC number) | 7 |
| Proteins with GO terms | 23 |
| Proteins assigned to a pathway | 0 |

Table 1. Strain ON4^T_{PacBio} genome assembly and annotation metrics.

frame (catalytic domain), ORFAB^{26,31}. Thus, ORFAB (396 aa) from ISGmo1 is a putative protein predicted in silico³² based on this mechanism. The three proteins of ISGmo1, ORFA, ORFB, and ORFAB, shared the highest amino acid sequence identity with the respective open reading frames of ISFsp8 (45–49%) found in *Frankia* sp. and ISAar25 and ISAar4 (43–45% and 44–46%, respectively) both found in *Arthrobacter arilaitensis*, also belonging to the phylum *Actinobacteria* (Fig. S2). Although DDE transposases have a highly conserved catalytic structure, they show overall low sequence conservation³³, which may explain the low identities found between the ISGmo1 transposase and those found in the ISfinder database (<https://www-is.biotoul.fr/>). Despite the low identities, ISGmo1 clusters with IS sequences of other *Actinobacteria*, suggesting a phylogenetic relationship (Fig. S2).

In Tn6311, the inverted repeats on the left side (IRL) includes the start codon of the *orfA/orfAB* ISGmo1 genes and also two putative transposase promoters according to in silico identification (<http://genome2d.molgenrug.nl/>) (Fig. 2a). In contrast, the repeats on the right side (IRR) of ISGmo1, just upstream the *molA* gene, includes the last three codons of *orfB* and the –35 region of the *molA* promoter. The structure of the *molA* gene regulatory region, i.e., the sequence of the regions –35, –10, and the ribosome binding sequence and their relative position towards the *molA* start codon (Fig. S3), suggests a strong promoter^{15,34}. The existence of such a strong promoter combined with the tandem organization of *molA* gene in Tn6311 explains the high rate of molinate breakdown by strain ON4^T, suggesting a high gene expression response.

Interestingly, the ISGmo1 sequence was also found in contig A (ON4^T_{PacBio}), but without any passenger gene (Fig. 2b) and with different direct repeats sequences. Evidence of such a copy of ISGmo1 was also given by the ON4^T₄₅₄, DSM13485^T (JGI, AUDX00000000.1), and ON4⁽⁻⁾ sequenced genomes and confirmed by primer design and Sanger sequencing (data not shown).

Characterization of contig B: pARLON1. Contigs A and B were presumed, based on size, to be a chromosome and a plasmid, respectively. The detection of a band with approximately 30 kbp by Pulsed-Field Gel Electrophoresis (PFGE) (Fig. S4) supported this hypothesis. Moreover, no reads mapping to contig B were found in the draft genomes of strains DSM13485^T (JGI, AUDX00000000.1) and ON4⁽⁻⁾ (Fig. 1), an ON4^T variant unable to degrade molinate (Fig. S5). Altogether these data corroborate the existence of a mobile element enabling the degradation of molinate, i.e., a plasmid carrying the *molA* gene, which, most probably, was cured in strain ON4⁽⁻⁾ after 12 generations in culture medium without the herbicide. In turn, the absence of reads mapping to contig B in strain DSM13485^T (JGI, AUDX00000000.1) suggests that the conditions used to grow the culture collection strain did not favour the maintenance of the plasmid that is quite unstable. Hence, contig B was named pARLON1.

Plasmids are described as having modular structures, where each “module” is dedicated to a particular function (replication, stability, conjugation, establishment and adaptation)³⁵. pARLON1 contained 47 identified ORFs from which 36 ORFs (~27 kbp) belong to the plasmid backbone, and 11 ORFs belong to the adaptive module (Tn6311, ~10 kbp) (Fig. 3). pARLON1 did not share significant nucleotide identity with other described plasmids

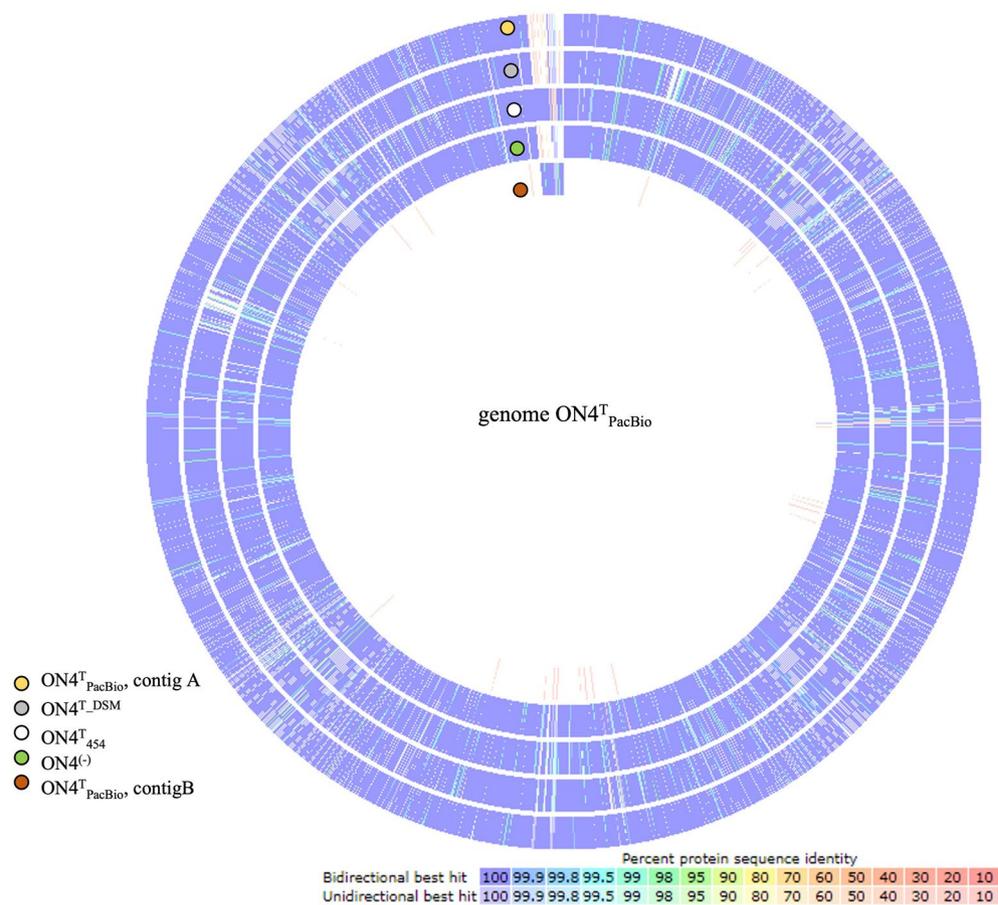


Figure 1. Sequence-based comparison of the reference $ON4^{T}_{PacBio}$ (CP028426 and CP028427) with each of the assembled genomes of *G. molinativorax* strains ($ON4^{T}$ —PXVE00000000, DSM13485^T (= $ON4^{T}$) AUDX00000000.1 and $ON4^{(-)}$ —PXVD00000000) using RAST.

(< 70%), not even with non-conjugative plasmids harboured by other *Microbacteriaceae* members (e.g., plasmids pCM1 and pCM2 of *Clavibacter michiganensis* subsp. *michiganensis*, closely related with *G. molinativorax*)³⁶. However, pARLON1 shared a high nucleotide identity (88.2%) with the contig 18 of *Agrococcus casei* LMG 22410^T draft genome (Contig FUHU01000018.1, GenBank FUHU00000000.1), as determined by ANI³⁷ (Table S3). LMG 22410^T is the type strain of *Agrococcus casei*, also a member of the family *Microbacteriaceae*, with a G+C content (65 mol%) closer to that of *G. molinativorax* (64 mol%) than to other *Agrococcus* species^{7,38}. The pARLON1 genes sharing > 70% identity with genes of the *A. casei* contig were those belonging to the plasmid backbone (36 predicted genes), representing 46.3 and 53.8% of the full coverage of contig B and contig FUHU01000018.1, respectively. In contrast, the predicted genes belonging to the adaptive module of each of these genetic elements were different (Fig. 3, Table S4).

Some ORFs of both pARLON1 and contig FUHU01000018.1 are related to the functional modules of transmissible plasmids, according to the annotation of pARLON1 and the in silico predictions performed by CONJscan⁴⁰ and by oriTDB⁴¹. These analyses predicted the existence of two essential proteins of the single-strand DNA conjugation machinery, the relaxase (MOB) and the type IV coupling protein (T4CP). The former is a critical component of the relaxosome responsible for initiating the DNA transfer in both conjugative and mobilizable plasmids, and the latter couples the relaxase-DNA nucleoprotein complex to the type IV secretion system (T4SS)⁴². A T4SS is a large complex of proteins spanning from the cytoplasm to the cell exterior constituting the mating-pair formation (MPF) module^{43,44}. ORF44 and ORF12 from strains $ON4^{T}$ and *A. casei*, respectively, showed homology with MOB proteins, whereas ORF46/*trsK* and ORF10 were related to the TraG/TraD family protein whose members are homologs to T4CP^{45,46} (Fig. 3, Table S4). However, only CONJscan predicted the existence of an ORF related to VirB4 ATPase (ORF4 and ORF7 from strains $ON4^{T}$ and *A. casei*, respectively), a marker of a T4SS presence^{43,44} (Fig. 3, Table S4). Hence, according to CONJscan, both $ON4^{T}$ and *A. casei* harbour a putative T4SS of the class FATA⁴⁰, which is a MPF family specific to *Firmicutes*, *Actinobacteria*, *Tenericutes* and *Archaea*^{43,44}, agreeing with the taxonomic affiliation of both $ON4^{T}$ and *A. casei*. Because members of these microbial groups lack an outer membrane, some of the core complex T4SS components (e.g., VirB7/10) of *Agrococcus tumefaciens* are not identified in all the well-characterized conjugation systems of MPF_{FATA} type (e.g., pCF10 and pGO1)⁴⁴. However, proteins homologous to VirB3, VirB6 and VirB8, proteins of the inner-membrane pore,

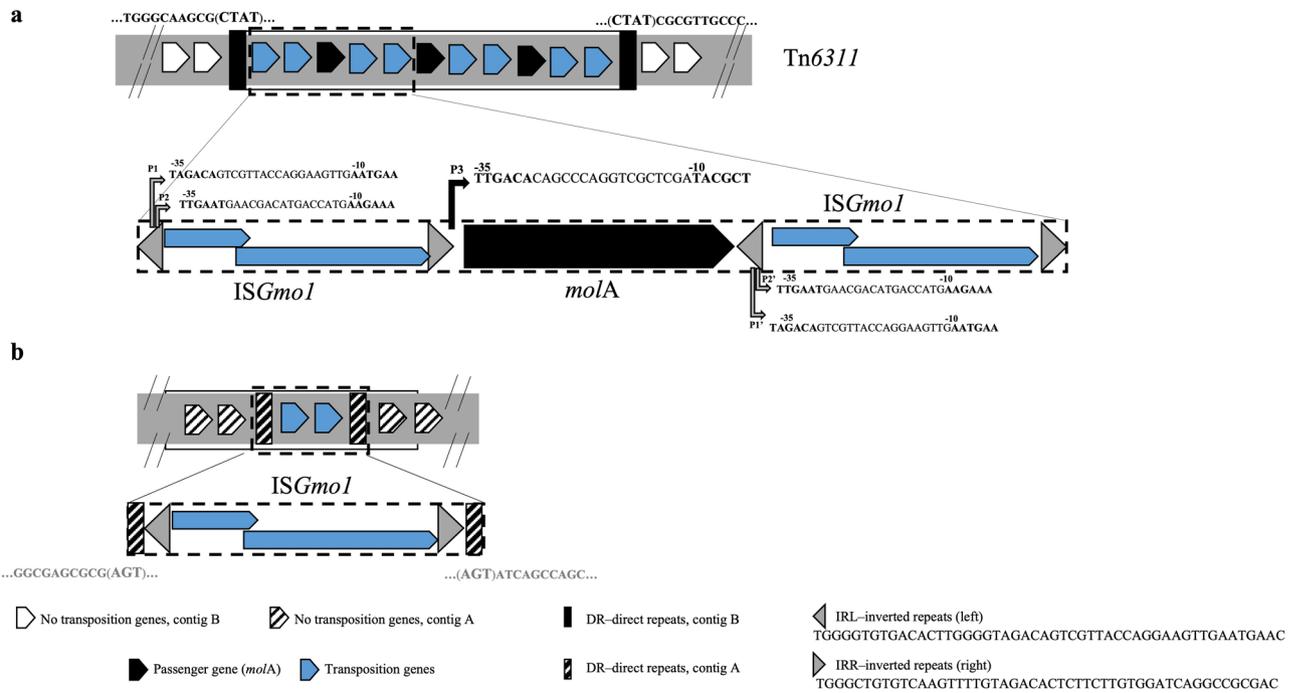


Figure 2. Mapping of the *molA* gene genetic vicinity. **(a)** Representation of the composite transposon Tn6311 in strain ON4^T (ON4^T_{PacBio}—CP028427), as well as the in silico predicted promoters (P) close to *molA* gene. **(b)** Schematic representation of the ISGmo1 without passenger gene in strain ON4^T and ON4⁽⁻⁾ (ON4⁽⁻⁾_{PacBio}—CP028426; and ON4^T₄₅₄—PXVE0000000; DSM13485^T (=ON4^T)—AUDX0000000.1 and ON4⁽⁻⁾—PXVD00000000, <http://www.ncbi.nlm.nih.gov/>).

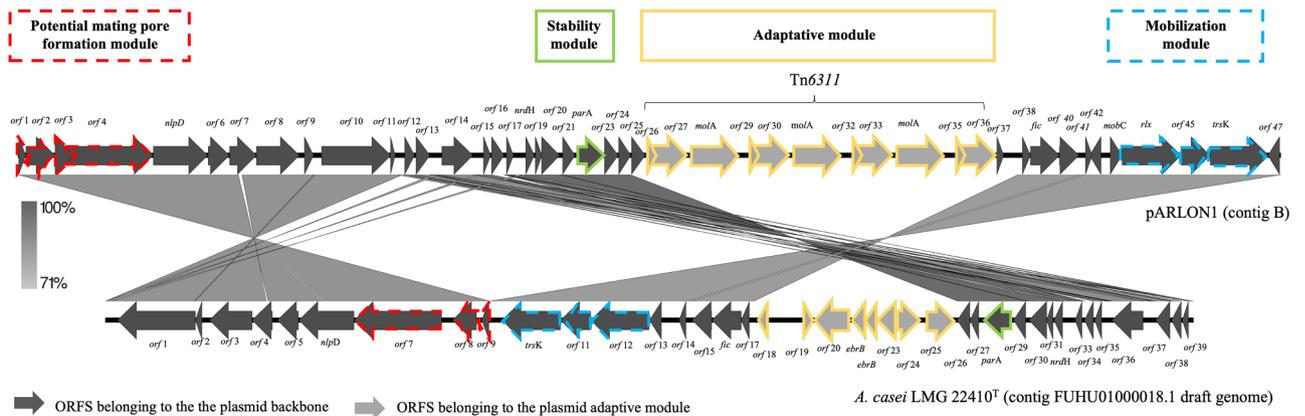


Figure 3. Genetic organization of pARLON1 and scheme of the BLASTn search results between pARLON1 (contig B) and *A. casei* LMG 22410^T (contig FUHU01000018.1 draft genome) using Easyfig software³⁹. Plasmid genes predicted by CONJscan⁴⁰ and/or oriTDB⁴¹ as related to a putative T4SS of the type_{FATA} are represented by dash lines.

are common within this system type⁴⁴. Among these MPF proteins, CONJscan predicted the existence of PrgH in ON4^T (ORF2) and *A. casei* (ORF8), which show homology with VirB6⁴⁴. Also, PrgI, which is suggested to have homology with VirB3⁴⁴, was predicted in ON4^T (ORF3). In addition, CONJscan predicted the presence of two other ORFs (1 and 45, and 9 and 11 from strains ON4^T and *A. casei*, respectively) related to accessory genes detected at frequencies higher than 50% within the MPF_{FATA} family (FATA_{prgF} and FATA_{cd411}). Although these predictions support the existence of a MPF_{FATA} type, the presence of a functional MPF module in pARLON1 needs to be experimentally validated.

In addition, the pARLON1 backbone seems to carry a stability module. This module is particularly important in low copy number plasmids, where mechanisms, such as the partition (Par) systems, are required to ensure the even partitioning of the DNA molecules among progeny (vertical transmission)^{35,47,48}. ORF22 and ORF28 from strains ON4^T and *A. casei*, respectively, were predicted to be related to the plasmid partition motor protein ParA. However, no homologous to other proteins of this system (e.g., ParB and ParS) were predicted. These results

| Symbol | Predicted gene product | RNA-Seq data | RT- qPCR data |
|--------------------------|--|---------------------|---|
| | | LOG2FC ^a | Average _N relative fold expression ^b ± SE |
| <i>molA</i> [#] | Molinate hydrolase | 15.21 | 393.27 ± 14.91 |
| | | 15.21 | |
| | | 15.21 | |
| <i>metE</i> | 5-methyltetrahydropteroyltriglutamate-homocysteine methyltransferase | 8.80 | 103.38 ± 11.35 |
| <i>chrC</i> | Caprolactone hydrolase | 3.56 | 387.49 ± 78.21 |
| <i>cinC</i> | Nitric oxide synthase | 5.85 | 5.40 ± 0.18 |
| <i>adhP</i> | Alcohol dehydrogenase | 5.85 | 37.59 ± 2.70 |
| <i>fprA</i> | NADPH-ferredoxin reductase FprA | 7.44 | 6.32 ± 0.20 |
| <i>hyuA</i> | Hydantoin utilization protein A | 3.59 | 2.85 ± 0.15 |
| <i>parA</i> [*] | Plasmid partition protein homolog ParA | N.D | 579.00 ± 69.16 |

Table 2. Validation of transcriptomics results by RT-qPCR. ^aFold-change was calculated as the ratio of the expression level of strain ON4^T growing in MMM by the expression level of strain ON4^T growing in LB. Log2FC > 2 was considered as different and indicative of overexpressed genes in ON4^T_{MMM}. For *parA* gene, as the expression was 0 for ON4^T_{LB} (i.e., reads mapping this gene were not detected), it was not possible to obtain the FC value (N.D.); ^bDefined as the ratio of the fold change of each target gene ($E_{\text{target}}^{\Delta} [C_{\text{LB}} - C_{\text{MMM}}]$) by the fold change of reference gene (geomean of $E_{\text{ref}}^{\Delta} [C_{\text{LB}} - C_{\text{MMM}}]$, based on the results obtained from each of the 3 reference genes selected in the present study); [#]There are 3 copies of this gene; ^{*}Ct values obtained for strain ON4^T growing in LB are close to 40, the lowest value for qPCR, which corroborates the data from cDNA libraries where the detection of this transcript is only possible in strain ON4^T growing in MMM. N.D. not detected, SE standard error.

may indicate that the Par system is incomplete/not functional (which could explain the existence of the cured strains DSM13485^T and ON4⁽⁻⁾), or that the existing DNA binding proteins have low homology to known Par proteins (lower homology was found for proteins of the ParB group than for group ParA, difficulting the classification of the former⁴⁹), or that another maintenance mechanism is present. At the population level, conjugative plasmids are described as low copy number, while mobilizable plasmids tend to be high copy number^{35,50}. The ratio between the plasmid (*parA*) and chromosomal (*recA*) single-copy genes in strain ON4^T was 3:1 (Table S2), suggesting that pARLON1 is a low copy number plasmid⁵¹. The presence of ORFs potentially associated with the maintenance and conjugation modules, together with the plasmid size (> 30 kbp) and its low copy number, suggest that pARLON1 can be a conjugative plasmid³⁵. However, further studies are necessary to validate its conjugative activity.

As referred to above, the adaptive module of pARLON1 differed from that of contig 18 from strain *A. casei* LMG 22410^T. However, the similarity of the plasmid backbone and the identification of a putative MPF_{FATA} in strains ON4^T and *A. casei* LMG 22410^T strengthens the existence of a common actinobacterial ancestor that, depending on the environment, acquired different adaptive modules. Moreover, plasmid evolution might have been mediated via the activity of the insertion sequences flanking the adaptive module, IS6100 in *A. casei* LMG 22410^T and ISGmo1 in Tn6311 of *G. molinivorax* ON4^T.

Identification of candidate genes involved in ACA mineralization. Given the absence of any other catabolic gene in the vicinity of *molA*, a transcriptomic approach was used to identify the potential genes involved in ACA mineralization. The gene expression profiles of strain ON4^T growing in molinate (mineral medium with the herbicide as the single source of energy, carbon and nitrogen, herein referred to as ON4^T_{MMM}) and without molinate (control medium LB, herein referred to as ON4^T_{LB}) were compared to identify the genes overexpressed in ON4^T_{MMM}.

After quality control, 14,380,135 and 15,330,180 reads in each library (ON4^T_{MMM} and ON4^T_{LB}, respectively) were retrieved. Analysis of the reads revealed that subtractive hybridization only removed a small amount of the ribosomal RNA, since 86–90% of reads mapped to rRNA. Consequently, approximately 1.4 and 1.0 million mRNA sequence reads were obtained from ON4^T_{MMM} and ON4^T_{LB}, respectively. After mapping the mRNA reads against the ON4^T genome sequence, 95% of the 3243 expected transcripts were expressed under the tested growth conditions (ON4^T_{MMM} and/or ON4^T_{LB}) (Table S5). From the 3095 expressed transcripts, 2726 were expressed under both growth conditions and 335 were overexpressed in ON4^T_{MMM} compared to ON4^T_{LB} (Tables S6 and S7). Of these, 149 transcripts were exclusive in ON4^T_{MMM}, i.e., no reads were detected in ON4^T_{LB}. RT-qPCR validated expression results in eight overexpressed genes in ON4^T_{MMM}. The trend of expression level was similar to that of the transcriptomics experiment (Table 2).

Metabolic pathways for *Gulosibacter* sp. are not available in KEGG. Nevertheless, the predicted gene products from the 335 overexpressed transcripts in ON4^T_{MMM} were mapped to the KEGG database⁵². Some of the overexpressed predicted gene products in ON4^T_{MMM} mapped to KEGG Orthology (KO) categories (Fig. S6), namely to metabolism (e.g., amino acid, carbohydrates), environmental information processing (e.g., membrane transport) and Brite hierarchies (e.g., signaling and cellular processes), respectively. However, no complete KEGG pathway appeared to be overexpressed in ON4^T_{MMM}, in particular, none that could be related to ACA mineralization.

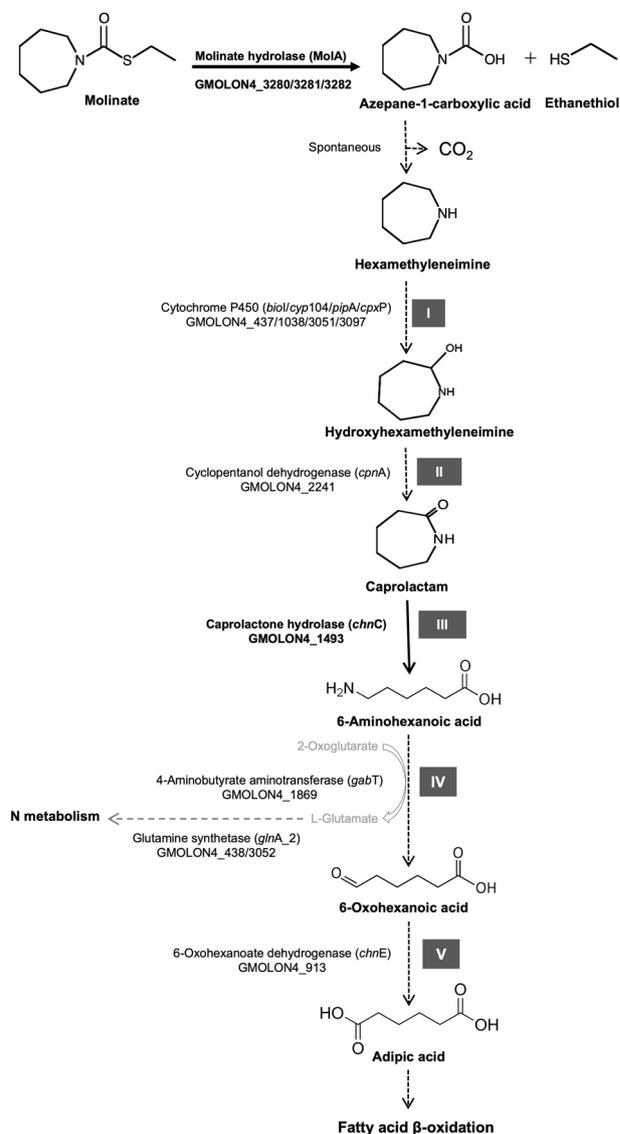


Figure 4. Candidate enzymes involved in the five steps (I–V) of the putative molinate degradation pathway by strain ON4^T adapted from Barreiros and collaborators⁸.

Most of the predicted gene products mapped to more than one metabolic pathway and/or different predicted gene products mapped to the same metabolic function.

Thereby, genes potentially involved in ACA mineralization were identified based on homology to previously described enzymes involved in the degradation of other xenobiotics^{53–56} and on the structural similarity of the putative substrates (Fig. 4). Potential candidates involved in the breakdown of the heterocyclic ring containing metabolites are cytochromes P450 (CYP), which act as monooxygenases and are among the most versatile biological catalyst⁵⁷, being, among others, involved in the degradation of different xenobiotics^{57,58}. In typical class I P450 systems, the electrons required for oxygen activation are indirectly provided by NAD(P)H. The co-enzyme generally reduces a ferredoxin reductase (FDR), which in turn transfers the electrons to the second component of the system, usually a ferredoxin⁵⁷. Finally, the ferredoxin mediates the electron transport between the FDR and the CYP⁵⁷, further inducing the incorporation of one oxygen atom into the substrate via the reductive cleavage of oxygen.

Strain ON4^T harbours four ORFs with homology to cytochromes P450, namely *bioI*, *cyp104*, *pipA* and *cpxP* (Fig. 5, Table S8). Among these, *bioI* and its contiguous ferredoxin reductase (GMOLON4_437 and GMOLON4_436, respectively) were highly expressed in ON4^T_{MMM} (Log2FC > 6, Table S8). Also, *cyp104* (GMOLON4_1038) and *cpxP* (GMOLON4_3097) were overexpressed in ON4^T_{MMM} (Log2FC > 3, Table S8), whereas *pipA* and its contiguous ferredoxin reductase (GMOLON4_3051 and GMOLON4_3050, respectively) were not significantly overexpressed in ON4^T_{MMM} (Log2FC < 2, Table S8). Noticeably, the cytochromes P450 *bioI* and *pipA* of strain ON4^T have an amino acid identity of 72% and have a similar genetic neighbourhood (Fig. 5). Indeed, the upstream and downstream enzymes gene sequences seem to be similar, particularly the contiguous ferredoxin

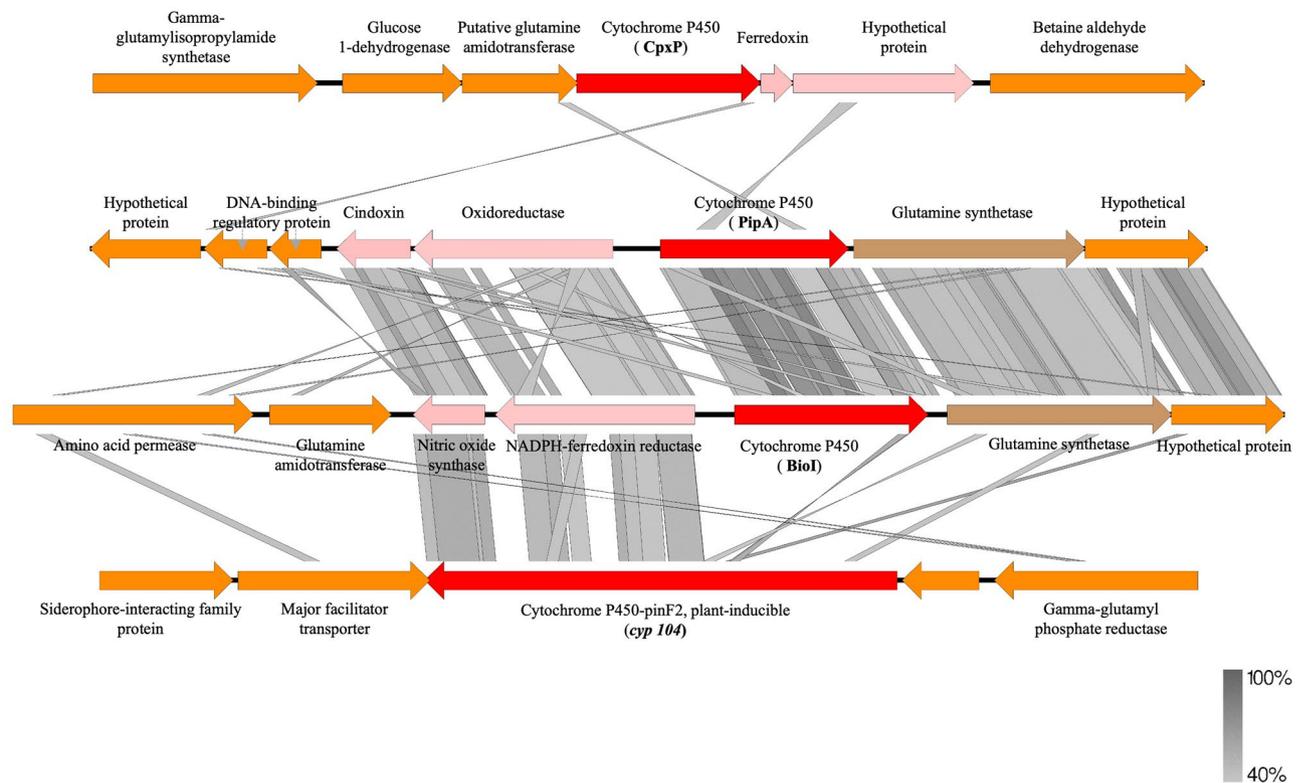


Figure 5. BLASTx between genes in the vicinity of cytochrome P450 (*cpxP*—GMOLON4_3097, *pipA*—GMOLON4_3051, *bioI*—GMOLON4_437 and *cyp104*—GMOLON4_1038, respectively) from strain ON4^T (Accession Number CP028426).

reductase GMOLON4_436 and GMOLON4_3050, respectively, which have an amino acid identity of 48% with each other. Also, the proteins potentially involved in the transfer of the electrons from the ferredoxin reductases to the cytochromes P450 *bioI* and *pipA* (GMOLON4_435 and GMOLON4_3049, respectively) share an amino acid identity of 69%. In contrast, these cytochromes (*bioI* and *pipA*) share less than 26% and 30% amino acid identity with the cytochromes *cyp104* and *cpxP*, respectively, which are also distinct from each other (24% amino acid identity) (Fig. 5).

Curiously, the cytochromes P450 *bioI* and *pipA* of strain ON4^T have an amino acid identity of approximately 53 and 52% with that of *Mycobacterium smegmatis* mc²155 (*pipA*, AAD28344.1) and *Mycobacterium* sp. HE5 (*morA*, AAV54064.1), respectively (Table S8). Both are shown to be involved in the hydroxylation of N-heterocycles^{53,59}. *M. smegmatis* *pipA* was shown to be induced when strain mc²155 grew in piperidine as the sole source of energy, carbon, and nitrogen⁵³. *Mycobacterium* sp. HE5 recombinant protein *morA* in combination with the adjacent pair ferredoxin reductase/ferredoxin was able to oxidize morpholine⁵⁹. Although the contiguous *bioI* and *pipA* ferredoxin reductase (GMOLON4_436 and GMOLON4_3050, respectively) were distinct from those found in *Mycobacterium* strains (Fig. S7), the inhibition of ACA mineralization under anaerobic conditions⁸, combined with the similarity between the molecular structure of piperidine/morpholine (six/five membered N-heterocycles), and the predicted seven membered N-heterocycle product from ACA degradation in strain ON4^T (hexamethyleneimine)⁸ (Fig. S7), supports the involvement of one of these systems in the hydroxylation of hexamethyleneimine to hydroxyhexamethyleneimine in strain ON4^T (step I, Fig. 4). However, further experimental validation is necessary.

According to previous studies, the resulting hydroxyhexamethyleneimine is further oxidized to caprolactam⁸, which suggests the activity of a dehydrogenase (step II, Fig. 4). Cyclopentanol dehydrogenase (*cpnA*, GMOLON4_2241) is among the overexpressed dehydrogenases in ON4^T_{MMM} (Log₂FC > 2, Table S8). It shows 40% amino acid identity with cyclopentanol dehydrogenases from *Comamonas* sp. NCIMB 9872⁶⁰ and 31–34% identity with cyclohexanol dehydrogenases from *Rhodococcus* sp. TK6⁶¹ and *Acinetobacter johnsonii*⁶², respectively, known to oxidise cyclic alcohols but not N-heterocyclic alcohols, such as hydroxyhexamethyleneimine^{63–65}. Therefore, the role of the GMOLON4_2241 product in the degradation of molinate in strain ON4^T needs further experimental validation.

Considering the further step of the putative degradation of molinate, the cleavage of the cyclic metabolite (caprolactam) into 6-aminoheptanoic acid (step III, Fig. 4), two possibilities arose from the transcriptomics analysis. These were the gene products of *hyuA*, GMOLON4_3203, and *hyuB*, GMOLON4_3204 (hydantoin utilization protein A and hydantoinase B, respectively), and the gene product of *chnC* (GMOLON4_1493, caprolactone hydrolase). All were overexpressed in ON4^T_{MMM} (Log₂FC > 2, Table S6). GMOLON4_3203 and GMOLON4_3204 showed the highest amino acid identity (72–90%) with proteins of the Hydantoinase-Hydantoinase B/Oxoprolinase family found in different members of the phylum *Actinobacteria*. Whereas GMOLON4_1493

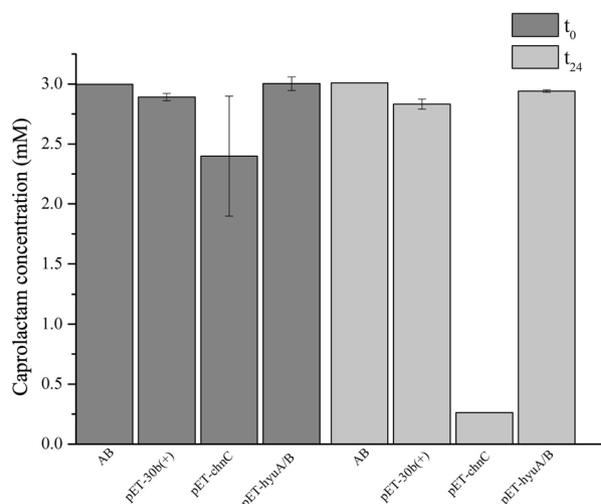


Figure 6. Evaluation of caprolactam degradation in resting cells assays ($DO_{\lambda 600nm} \approx 6$), by *E. coli* BL21(DE3) strain carrying either pET-30b(+), pET-*chnC* or pET-*hyuA/B* over a 24 h incubation period in 50 mM of phosphate buffer (pH 7.2) and 3 mM caprolactam. Values are means \pm SE ($n = 3$, for all biological assays). AB—abiotic control.

showed the highest amino acid identity (75–79%) with *Alpha/Beta* hydrolases, followed by caprolactone hydrolase (72–75%) of *Rhodococcus* sp. strains HI-31 and Phi2^{66,67}. Both hydantoinases/oxoprolinases and caprolactone hydrolase are described as able to open cyclic organic compounds, by hydrolysing C-N bonds in cyclic amides and opening the lactone ring, respectively.

Caprolactam is the raw material for manufacturing Nylon-6 by ring opening polymerization. Attempts to isolate organisms capable of attenuating the environmental pollution caused by caprolactam intensive production and utilization have been made. Up to now, few Gram negative and Gram positive bacterial strains, including *Gulosibacter* sp. BS4, were described as capable of caprolactam degradation^{56,68–70}. Among these organisms, only the enzyme acting upon the cleavage of caprolactam in *Pseudomonas jessenii* GO3 was described^{56,71}. This caprolactamase, which consists of two subunits with high sequence identity to 5-oxoprolinases, hydrolyzes caprolactam into 6-aminohexanoic acid in an ATP-dependent manner^{56,71}. Such information suggested the possible involvement of the GMOLON4_3203 and GMOLON4_3204 gene products in the breakdown of caprolactam by strain ON4^T. However, heterologous expression of both potential candidates demonstrated that only the *chnC* gene product (GMOLON4_1493) degraded caprolactam in resting cells assays (Fig. 6). Proteins affiliated to this family are involved in the cyclic ϵ -caprolactone transformation into a linear organic acid, formed during the oxidation of cyclohexanol or cyclohexanone by both Gram staining positive and negative bacteria^{54,55,66,67}. Hence, as far as we know, this is the first report of their involvement in caprolactam degradation.

Although only 6-oxohexanoate dehydrogenase (GMOLON4_913, *chnE*) was overexpressed in ON4^T_{MMM} (Log₂FC > 3, Table S8), both 4-aminobutyrate aminotransferase (GMOLON4_1869, *gabT*) and 6-oxohexanoate dehydrogenase (GMOLON4_913, *chnE*) are potentially involved in the transformation of 6-aminohexanoic acid to adipic acid (step IV and V, Fig. 4). Indeed, the 4-aminobutyrate aminotransferase (GMOLON4_1869, *gabT*) and the 6-oxohexanoate dehydrogenase (GMOLON4_913, *chnE*) have an amino acid identity of 57 and 80% with the NylD1 6-aminohexanoate aminotransferase and the NylE1 adipate semialdehyde dehydrogenase from the Nylon oligomer-degrading actinobacterium *Arthrobacter* sp. strain KI72²⁴, respectively (Table S8). In strain KI72, NylD1 in combination with the cofactor pyridoxal phosphate (PLP) catalyse the transference of the amino group from 6-aminohexanoate to an amino acceptor (e.g., α -ketoglutarate, pyruvate, and glyoxylate) generating 6-oxohexanoate and an amino-acid (e.g., glutamate, alanine, and glycine), respectively. Then, NylE1 catalyses the oxidation of 6-oxohexanoate to adipate using NADP⁺ as a cofactor, enabling the *Arthrobacter* sp. strain KI72 growth on 6-aminohexanoate as a sole source of carbon and nitrogen²⁴. Similarly, the adipic acid formed through the activity of the 6-oxohexanoate dehydrogenase (GMOLON4_913, *chnE*) may enter the fatty acid β -oxidation providing further energy for strain ON4^T metabolic activity⁷² (Fig. 4). Indeed, the enzymes of this central metabolic pathway were overexpressed in ON4^T_{MMM} (Table S8). The central metabolism activated via succinyl-CoA and acetyl-CoA from fatty acid β -oxidation may be thus related to the ability of strain ON4^T to use ACA as the single source of carbon, nitrogen, and energy⁸.

Also, the overexpression of transcripts related to the metabolism of N-compounds, particularly glutamine synthetase GMOLON4_438/3052 (Fig. 4 and Fig. S6, Table S8), agrees with the ability of strain ON4^T to grow on molinate as the single source of nitrogen⁸. The importance of glutamine synthetase and glutamate synthase cyclic mechanism in ammonium assimilation in prokaryotes is well known. In particular, Harper and collaborators⁷³ reported that either glutamate dehydrogenase or glutamine synthase are very responsive and regulated under high or low nitrogen availability in *M. smegmatis*.

Among the proteins potentially involved in molinate degradation, only those related with the cytochromes P450 BioI and PipA seem to be found in some strains of the genus *Gulosibacter* (amino acid identity of up to 72%

and 91% respectively, Table S9). These results suggest that some *Gulosibacter* strains, such as *G. chungangensis* KCTC 13959^T, may have the ability to partially oxidize N-heterocycle compounds.

Interestingly, the products of the genes putatively involved on hexamethyleneimine degradation in strain ON4^T showed relatively high amino acid identity with proteins involved in the oxidation of 6 membered N-heterocycle compounds (piperidine) and cycloalkane alcohols (cyclopentanol/cyclohexanol) by a wide diversity of strains, namely actinobacteria (Table S8). All these results suggest that small alterations on the structure of these proteins may have allowed the utilization of caprolactam in strains ON4^T and *Gulosibacter* sp. BS4 and ACA in strain ON4^T as energy, carbon, and nitrogen sources.

Conclusions

In summary, the present study, allowed the in silico identification and characterization of a novel transposon (Tn6311), which harbours *molA*, the enzyme responsible for the breakdown of the molinate to ACA and confers the unique degrading ability of *G. molinivorax* ON4^T. In addition, a novel transmissible plasmid (pARLON1), which is present in strain ON4^T and *Agrococcus casei* LMG 22410^T, was characterized in silico. This last finding is of particular importance as the characterization of actinobacterial transmissible plasmids (mobilizable or conjugative) is scarce in the literature, compared to other bacterial phyla, particularly *Proteobacteria*. The future characterization of the plasmid transmission pathway will enlighten the underrepresented type IV secretion system (T4SS) in the FATA group, which includes only 3 representatives⁴⁴. Such knowledge would enrich databases such as CONJscan⁴⁰ and oriTDB⁴¹, contributing to better identifying this system, particularly in genomic mining studies.

Additionally, this study identified seven candidate genes involved in ACA mineralization, and validated the activity of the GMOLON4_1493 product, related to the caprolactone hydrolase *chnC*, in the degradation of caprolactam. The future characterization of these enzymes will contribute to enlightening the molinate degradation pathway in strain ON4^T and strengthening their biocatalytic potential.

Material and methods

Culture conditions. *Gulosibacter molinivorax* ON4^T and the variant strain ON4⁽⁻⁾ were preserved in 15% (v/v) glycerol suspensions at -80°C⁷. Whenever necessary, stored cells were recovered in Luria-Bertani Agar supplemented with 1 mM molinate (LAM). Cultures were incubated at 30 °C for 72 h. For 454 pyrosequencing, strain ON4^T was grown on LAM, and for PacBio sequencing in mineral medium B⁹ supplemented with 0.2 g/L yeast extract and 1.5 mM molinate (MMM). For Illumina sequencing, strain ON4⁽⁻⁾ was grown in LAM. For a detailed description see the Supplementary Material.

Genome analysis (sequencing, assembling and annotation). 454 pyrosequencing and PacBio technologies were used to sequence the genome of strain ON4^T. For 454 pyrosequencing, five hundred nanograms of high-quality genomic DNA were fragmented by nebulization and the sequencing adapters ligated to create double-stranded DNA libraries. The 454 GS FLX sequencing platform (Roche) at GenoInseq (Cantanhede, Portugal) was used following the standard manufacturer's protocols (see Supplementary Material). Sequencing reads were quality trimmed and de novo assembled by GS Assembler, version 2.9 (Roche) using default parameters.

Genome resequencing was performed using SMRT technology PacBio RSII on the RS sequencer at GATC Biotech AG (Constance, Germany), according to the manufacturer's instructions⁷⁴. Genome assembly was performed on the SMRT analysis software (<http://www.pacb.com/products-and-services/analytical-software/smart-analysis/>). The Hierarchical Genome Assembly Process (HGAP, RS HGAP Assembly.3), which includes pre-assembly, de novo assembly with PacBio's AssembleUnitig and assembly polishing with Quiver, was used with the default parameters, except for the minimum polymerase read quality, where 0.80 was used instead of 0.75. Gene prediction was carried out with GeneMarkS⁷⁵ using specific parameters for prokaryotes (genetic code 11) and functional annotation was performed using the Sma3s⁷⁶. The completeness and contamination of the SMRT assembled genome of strain ON4^T were calculated with CheckM⁷⁷.

For the draft genome of strain ON4⁽⁻⁾ a DNA library was prepared from one nanogram of high-quality genomic DNA with the Nextera XT DNA Sample Preparation Kit (Illumina, San Diego, USA), which was sequenced using paired-end (PE) 2 × 300 bp on the MiSeq[®] Illumina[®] platform at GenoInseq (Cantanhede, Portugal). All procedures were performed according to the standard manufacturer's protocols. Following demultiplex and quality-filter procedures, the high quality and adapter free reads were assembled with SPAdes, version 3.7.1⁷⁸, using in-house defined parameters.

Confirmation of the *molA* regulatory region. The 5'-Rapid Amplification of cDNA Ends (5' RACE) was performed to identify the *molA* gene regulatory region using RNA extracted from strain ON4^T grown in MMM. For a detailed description see the Supplementary Material.

Comparison of different assemblies. The assemblies obtained from all the *G. molinivorax* genome sequencing projects (PRJNA245014 present study and PRJNA188907) were compared using MAUVE software⁷⁹ and the RAST server⁸⁰. In addition, the raw reads of genomes ON4^T₄₅₄ (SRR6825088) and ON4⁽⁻⁾ (SRR6825089) were also mapped against the contigs obtained in SMRT analysis software (genome ON4^T_{PacBio}) (<https://www.pacb.com/products-and-services/analytical-software/smart-analysis/>), using the BWA-MEM algorithm⁸¹. The alignments were evaluated using Qualimap v2.2.1 software⁸² and were visualized in Tablet software⁸³.

A BLASTx search of contig B against *Agrococcus casei* LMG 22410^T draft genome (PRJEB19012) was performed (<https://blast.ncbi.nlm.nih.gov/>). In addition, the comparison (through BLASTn) between contig B and contig FUHU01000018.1 of *Agrococcus casei* LMG 22410^T was carried out using the Contiguity software⁸⁴, and

their similarity was measured by the Average Nucleotide Identity (ANI) calculator (<https://www.ezbiocloud.net/tools/ani>).

Growth conditions, RNA extraction and sequencing of cDNA libraries. Given the low biomass yield of strain ON4^T in mineral medium supplemented with a single carbon source other than molinate, strain ON4^T was grown up to the exponential phase in MMM (ON4^T_{MMM}) and Luria–Bertani Broth (ON4^T_{LB}) in 250 mL Erlenmeyer flasks to obtain mRNA libraries. Total RNA was extracted using the RNeasy extraction kit (QIAGEN GmbH) (detailed protocol in the Supplementary Material). The RNA quality was assessed using the RNA 6000 Pico Kit with the Prokaryote Total RNA Pico assay (Agilent 2100 Agilent Technologies, Santa Clara, CA, USA).

The subtractive hybridization protocol described by Stewart et al.⁸⁵ was used (detailed protocol in the Supplementary Material). The resulting rRNA-subtracted products were used to build the cDNA libraries for strand-specific RNA sequencing. The library preparation followed the protocol for whole transcription libraries in the Ion total RNA-Seq kit v2 (Life Technologies), using 500 ng of rRNA-subtracted products, as suggested by the manufacturer. Sequences were obtained in the Ion Proton™ platform at Genoinseq (Cantanhede, Portugal).

Normalization of the data from mRNA libraries. Ion Proton adapter sequences and low-quality bases were trimmed using the Torrent Suite software (Life Technologies). Trimmed reads were aligned to the strain ON4^T genome using TMAP version 4.0.6 (Life Technologies). Reads that mapped to the rRNA regions were excluded. The reads count was accessed using eXpress⁸⁶ and normalized using the Trimmed Mean of M values by edgeR⁸⁷. Transcript expression differences between ON4^T_{MMM} and ON4^T_{LB} were considered when $|\text{Log}_2\text{FC}| > 2$, where Fold Change (FC) is the ratio of normalized expression values of ON4^T_{MMM} over those of ON4^T_{LB} for each transcript. Thus, transcripts with $\text{Log}_2\text{FC} > 2$ are overexpressed in ON4^T_{MMM}, whereas those with $\text{Log}_2\text{FC} > -2$ are overexpressed in ON4^T_{LB}. The transcripts overexpressed in ON4^T_{MMM} were mapped to the KEGG database⁵² to identify potential overexpressed metabolic pathways related to ACA degradation.

Validation of transcriptomic data by RT-qPCR. For validation of libraries data, strain ON4^T was grown in MMM and LB. Total RNA was extracted using kit NucleoSpin™ RNA (Macherey–Nagel), and the cDNA from each sample was obtained with the Maxima H Minus First Strand cDNA Synthesis Kit, (ThermoScientific) (see Supplementary Material for detailed protocol).

The potential reference genes were selected among those without differences in expression between ON4^T_{MMM} and ON4^T_{LB}. Three reference genes (*gyrA*, *gyrB*, *recA*) were selected, and primers were designed using Oligo Explorer (<http://www.genelink.com/tools/gl-oe.asp>). For each reference gene, the primer pair specificity was assessed in silico, followed by qPCR amplification efficiency checking (Tables S10, S11) using different dilutions of DNA amplicons, and the amplification specificity was assessed using melting curves. The target genes were chosen considering differences in the expression values ($|\text{Log}_2\text{FC}| > 2$) between ON4^T_{MMM} and ON4^T_{LB}. The primer choice and design and qPCR amplification efficiency and specificity (Tables S10, S11) for target genes were performed as described for reference genes (see Supplementary Material for detailed protocol).

RT-qPCR was performed in 96-well plates (Thermo Fischer Scientific) on the StepOnePlus system (Applied Biosystems). All RT-qPCR experiments were performed using three biological replicates, two technical replicates, and negative control (without cDNA). For detailed protocol see Supplementary Material.

Evaluation and validation of reference gene expression stability and fold changes of target genes. The candidate reference genes stability was analysed using RefFinder (<https://omictools.com/reffinder-tool>), where a ranking of genes stability is obtained considering the data obtained from geNorm, NormFinder, BestKeeper and deltaCt method^{88–91}. The relative expression profile of each evaluated target gene was determined and normalized using a geometric mean value obtained from the amplification of the three best reference genes in all the samples (ON4^T_{MMM} and ON4^T_{LB}, respectively). Relative fold changes in gene expression were calculated using the Pfaffl method⁹², and data are displayed as mean ± standard error (SE).

Plasmid construction and cloning. All bacterial strains and plasmids used in the conjugants construction are listed in Table S12. The full sequence of the target genes (*chnC* and *hyuA/hyuB*) was amplified using total DNA of strain ON4^T as a template and specific primers with a *NdeI* restriction site (Table S13). An A-tailing was added to the purified PCR products and further inserted into pTZ57R/T vector. *Escherichia coli* JM109 was transformed with pTZ57R/T with each insert and recombinants were identified by blue/white colour selection and confirmed by colony PCR (Table S14). Recombinant pTZ57R/T vectors were extracted and further digested with the respective Speedy Restriction Enzymes (NZYTech), according to manufacturer's instructions (Table S13). The pET-30b(+) vector was restricted according to each target. After ligation, *E. coli* JM109 was transformed with the pET-30b(+) alone or carrying the target genes, and positive clones were selected from LB agar plates containing kanamycin (30 µg/mL) and screened by colony PCR (Table S14). Then, *E. coli* BL21 (DE3) was transformed with pET, pET-*chnC* or pET-*hyuA/B* vectors to further express the targeted genes *r*. See Supplementary Material for detailed protocol.

Heterologous expression. Resting cells of *E. coli* BL21(DE3) strain carrying the pET-*chnC* or pET-*hyuA/B* vectors were used to analyse the ability of the enzymes caprolactone hydrolase and hydantoin utilisation protein A/B, encoded by *chnC* and *hyuA/hyuB*, respectively, to degrade caprolactam. The biomass was resuspended ($\text{DO}_{\lambda_{600\text{nm}}} \approx 12$) in 50 mM of phosphate buffer (pH 7.2) and mixed with 6 mM of caprolactam in 50 mM

of phosphate buffer (1:1 ratio). The resting cells were incubated at 30 °C for 24 h. *E. coli* BL21(DE3) with the pET-30b(+) was used as biotic control. Caprolactam was quantified in cell-free supernatants, collected at 0 and 24 h of incubation, by High Performance Liquid Chromatograph (HPLC, VWR Hitachi Chromaster), following the method previously described⁹³. See Supplementary Material for detailed protocol.

Data availability

The datasets generated and analysed during the current study are available in the respective databases. The genomes obtained were deposited in GenBank under the accession numbers PXVE00000000 (genome ON4^T₄₅₄), CP028427 and CP028426 (genome ON4^T_{PacBio}) and PXVD00000000 (genome ON4⁽⁻⁾). The IS*Gmo1* sequence was deposited in the ISfinder database (<https://www-is.biotoul.fr/>). The composite transposon carrying *mola* was named Tn6311, according to the Tn number registry (<http://www.ucl.ac.uk/eastman/research/departments/microbial-diseases/tn>). The data from the cDNA libraries obtained in this work have been deposited in the NCBI Gene Expression Omnibus (GEO) under accession numbers GSM3307305 and GSM3307306 for ON4^T_{MMM} and ON4^T_{LB} data, respectively.

Received: 14 January 2022; Accepted: 18 August 2022

Published online: 15 September 2022

References

- Wiltschi, B. *et al.* Enzymes revolutionize the bioproduction of value-added compounds: From enzyme discovery to special applications. *Biotechnol. Adv.* **40**, 107520. <https://doi.org/10.1016/J.BIOTECHADV.2020.107520> (2020).
- Alcalde, M., Ferrer, M., Plou, F. J. & Ballesteros, A. Environmental biocatalysis: From remediation with enzymes to novel green processes. *Trends Biotechnol.* **24**, 281–287 (2006).
- Singh, R., Kumar, M., Mittal, A. & Mehta, P. K. Microbial enzymes: Industrial progress in 21st century. *3 Biotech* **6**, 1–15 (2016).
- Yang, J. W., Zheng, D. J., Cui, B. D., Yang, M. & Chen, Y. Z. RNA-seq transcriptome analysis of a *Pseudomonas* strain with diversified catalytic properties growth under different culture medium. *Microbiologyopen* **5**, 626–636 (2016).
- Miyazaki, R. *et al.* Comparative genome analysis of *Pseudomonas knackmussii* B13, the first bacterium known to degrade chloro-aromatic compounds. *Environ. Microbiol.* **17**, 91–104 (2015).
- Tischler, D., van Berkel, W. J. H. & Fraaije, M. W. Editorial: Actinobacteria, a source of biocatalytic tools. *Front. Microbiol.* **10**, 800. <https://doi.org/10.3389/fmicb.2019.00800> (2019).
- Manaiia, C. M., Nogales, B., Weiss, N. & Nunes O. C. *Gulosibacter molinativorax* gen. nov., a molinate-degrading bacterium, and classification of 'Brevibacterium helvolum' DSM 20419 as Pseudoclavibacter helvolus gen. nov., sp. Nov. *Int. J. Syst. Evol. Microbiol.* **54**, 783–789 (2004).
- Barreiros, L. *et al.* New insights into a bacterial metabolic and detoxifying association responsible for the mineralization of the thiocarbamate herbicide molinate. *Microbiology* **154**, 1038–1046 (2008).
- Barreiros, L. *et al.* A novel pathway for mineralization of the thiocarbamate herbicide molinate by a defined bacterial mixed culture. *Environ. Microbiol.* **5**, 944–953 (2003).
- Duarte, M. *et al.* *Gulosibacter molinativorax* ON4^T molinate hydrolase, a novel cobalt-dependent amidohydrolase. *J. Bacteriol.* **193**, 5810–5816 (2011).
- Sugrue, E. *et al.* Evolutionary expansion of the amidohydrolase superfamily in bacteria in response to the synthetic compounds molinate and diuron. *Appl. Environ. Microbiol.* **81**, 2612–2624 (2015).
- Leite, J. P. *et al.* Structure-guided engineering of molinate hydrolase for the degradation of thiocarbamate pesticides. *PLoS ONE* **10**, 1–18 (2015).
- Greated, A., Lamberts, L., Williams, P. A. & Thomas, C. M. Complete sequence of the IncP-9 TOL plasmid pWW0 from *Pseudomonas putida*. *Environ. Microbiol.* **4**, 856–871 (2002).
- Simon, M. J. *et al.* Sequences of genes encoding naphthalene dioxygenase in *Pseudomonas putida* strains G7 and NCIB 9816-4. *Gene* **127**, 31–37 (1993).
- Rocha, E. P. C. The organization of the bacterial genome. *Annu. Rev. Genet.* **42**, 211–233 (2008).
- Habe, H. *et al.* Phthalate catabolic gene cluster is linked to the angular dioxygenase gene in *Terrabacter* sp. strain DBF63. *Appl. Microbiol. Biotechnol.* **61**, 44–54 (2003).
- Dennis, J. J. & Zylstra, G. J. Complete sequence and genetic organization of pDTG1, the 83 kilobase naphthalene degradation plasmid from *Pseudomonas putida* strain NCIB 9816-4. *J. Mol. Biol.* **341**, 753–768 (2004).
- Sota, M. *et al.* Genomic and functional analysis of the IncP-9 naphthalene-catabolic plasmid NAH7 and its transposon Tn4655 suggests catabolic gene spread by a tyrosine recombinase. *J. Bacteriol.* **188**, 4057–4067 (2006).
- Harayama, S. & Reikik, M. The meta cleavage operon of TOL degradative plasmid pWWO comprises 13 genes. *MGG Mol. Gen. Genet.* **221**, 113–120 (1990).
- Ohtsubo, Y. *et al.* BphS, a key transcriptional regulator of *bph* genes involved in polychlorinated biphenyl/biphenyl degradation in *Pseudomonas* sp. KKS102. *J. Biol. Chem.* **276**, 36146–36154 (2001).
- Matus, V., Sánchez, M. A., Martínez, M. & González, B. Efficient degradation of 2,4,6-trichlorophenol requires a set of catabolic genes related to *tcp* genes from *Ralstonia eutropha* JMP134(pJP4). *Appl. Environ. Microbiol.* **69**, 7108–7115 (2003).
- Furukawa, K., Hayase, N., Taira, K. & Tomizuka, N. Molecular relationship of chromosomal genes encoding biphenyl/polychlorinated biphenyl catabolism: Some soil bacteria possess a highly conserved *bph* operon. *J. Bacteriol.* **171**, 5467–5472 (1989).
- Hirose, J. *et al.* Biphenyl/PCB Degrading *bph* genes of ten bacterial strains isolated from biphenyl-contaminated soil in Kitakyushu, Japan: Comparative and dynamic features as integrative conjugative elements (ICEs). *Genes (Basel)* **10**, 404 (2019).
- Takehara, I. *et al.* Metabolic pathway of 6-aminohexanoate in the nylon oligomer-degrading bacterium *Arthrobacter* sp. KI72: identification of the enzymes responsible for the conversion of 6-aminohexanoate to adipate. *Appl. Microbiol. Biotechnol.* **102**, 801–804 (2018).
- Hayatsu, M., Hirano, M. & Nagata, T. Involvement of two plasmids in the degradation of carbaryl by *Arthrobacter* sp. strain RC100. *Appl. Environ. Microbiol.* **65**, 1015–1019 (1999).
- Siguié, P., Gourbeyre, E. & Chandler, M. Bacterial insertion sequences: Their genomic impact and diversity. *FEMS Microbiol. Rev.* **38**, 865–891 (2014).
- Darmon, E. & Leach, D. R. F. Bacterial genome instability. *Microbiol. Mol. Biol. Rev.* **78**, 1–39 (2014).
- Tsuda, M., Tan, H. M., Nishi, A. & Furukawa, K. Mobile catabolic genes in bacteria. *J. Biosci. Bioeng.* **87**, 401–410 (1999).
- Habe, H. *et al.* Characterization of the upper pathway genes for fluorene metabolism in *Terrabacter* sp. strain DBF63. *J. Bacteriol.* **186**, 5938–5944 (2004).
- Mahillon, J. & Chandler, M. Insertion sequences. *Microbiol. Mol. Biol. Rev.* **62**, 725–774 (1998).

31. Polard, P., Prère, M. F., Chandler, M. & Fayet, O. Programmed translational frameshifting and initiation at an AUU codon in gene expression of bacterial insertion sequence IS911. *J. Mol. Biol.* **222**, 465–477 (1991).
32. 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).
33. Guérillot, R., Siguier, P., Goubeyre, E., Chandler, M. & Glaser, P. The diversity of prokaryotic DDE transposases of the mutator superfamily, insertion specificity, and association with conjugation machineries. *Genome Biol. Evol.* **6**, 260–272 (2014).
34. Mccarthy, J. E. G. & Gualerzi, C. Translational control of prokaryotic gene expression. *Reviews* **6**, 78–85 (1990).
35. Garcillán-Barcia, M. P., Alvarado, A. & De la Cruz, F. Identification of bacterial plasmids based on mobility and plasmid population biology. *FEMS Microbiol. Rev.* **35**, 936–956 (2011).
36. Gartemann, K. H. *et al.* The genome sequence of the tomato-pathogenic actinomycete *Clavibacter michiganensis* subsp. *michiganensis* NCPPB382 reveals a large island involved in pathogenicity. *J. Bacteriol.* **190**, 2138–2149 (2008).
37. Yoon, S. H., Ha, S. M., Lim, J., Kwon, S. & Chun, J. A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie van Leeuwenhoek Int. J. Gen. Mol. Microbiol.* **110**, 1281–1286 (2017).
38. Bora, N. *et al.* *Agrococcus casei* sp. nov., isolated from the surfaces of smear-ripened cheeses. *Int. J. Syst. Evol. Microbiol.* **57**, 92–97 (2007).
39. Sullivan, M. J., Petty, N. K. & Beatson, S. A. Easyfig: A genome comparison visualizer. *Bioinformatics* **27**, 1009–1010 (2011).
40. Abby, S. S. *et al.* Identification of protein secretion systems in bacterial genomes. *Sci. Rep.* **6**, 1–14 (2016).
41. Li, X. *et al.* oriFinder: A web-based tool for the identification of origin of transfers in DNA sequences of bacterial mobile genetic elements. *Nucleic Acids Res.* **46**, W229–W234 (2018).
42. Francia, M. V. *et al.* A classification scheme for mobilization regions of bacterial plasmids. *FEMS Microbiol. Rev.* **28**, 79–100 (2004).
43. Guglielmini, J., De La Cruz, F. & Rocha, E. P. C. Evolution of conjugation and type IV secretion systems. *Mol. Biol. Evol.* **30**, 315–331 (2013).
44. Guglielmini, J. *et al.* Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. *Nucleic Acids Res.* **42**, 5715–5727 (2014).
45. Grohmann, E., Muth, G. & Espinosa, M. Conjugative plasmid transfer in Gram-positive bacteria. *Microbiol. Mol. Biol. Rev.* **67**, 277–301 (2003).
46. Alvarez-Martinez, C. E. & Christie, P. J. Biological diversity of prokaryotic Type IV secretion systems. *Microbiol. Mol. Biol. Rev.* **73**, 775–808 (2009).
47. Guynet, C., Cuevas, A., Moncalián, G. & de la Cruz, F. The *stb* operon balances the requirements for vegetative stability and conjugative transfer of plasmid R388. *PLoS Genet.* **7**, e1002073 (2011).
48. Bignell, C. & Thomas, C. M. The bacterial ParA-ParB partitioning proteins. *J. Biotechnol.* **91**, 1–34 (2001).
49. Ceglowski, P. & Zielenkiewicz, U. Mechanisms of plasmid stable maintenance with special focus on plasmid addiction systems. *Acta Biochim. Pol.* **48**, 1003–1023 (2001).
50. Fernández-López, R. *et al.* Dynamics of the IncW genetic backbone imply general trends in conjugative plasmid evolution. *FEMS Microbiol. Rev.* **30**, 942–966 (2006).
51. Münch, K., Münch, R., Biedendieck, R., Jahn, D. & Müller, J. Evolutionary model for the unequal segregation of high copy plasmids. *PLoS Comput. Biol.* **15**, e1006724. <https://doi.org/10.1371/journal.pcbi.1006724> (2019).
52. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAA5: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**, W182–W185. <https://doi.org/10.1093/nar/gkm321> (2007).
53. Poupin, P., Ducrocq, V., Hallier-Soulier, S. & Truffaut, N. Cloning and characterization of the genes encoding a cytochrome P450 (PipA) involved in piperidine and pyrrolidine utilization and its regulatory protein (PipR) in *Mycobacterium smegmatis* mc²155. *J. Bacteriol.* **181**, 3419–3426 (1999).
54. Brzostowicz, P. C., Blasko, M. S. & Rouvière, P. E. Identification of two gene clusters involved in cyclohexanone oxidation in *Brevibacterium epidermidis* strain HCU. *Appl. Microbiol. Biotechnol.* **58**, 781–789 (2002).
55. Cheng, Q., Thomas, S. M., Kostichka, K., Valentine, J. R. & Nagarajan, V. Genetic analysis of a gene cluster for cyclohexanol oxidation in *Acinetobacter* sp. strain SE19 by in vitro transposition. *J. Bacteriol.* **182**, 4744–4751 (2000).
56. Otzen, M., Palacio, C. & Janssen, D. B. Characterization of the caprolactam degradation pathway in *Pseudomonas jessenii* using mass spectrometry-based proteomics. *Appl. Microbiol. Biotechnol.* **102**, 6699–6711 (2018).
57. Hannemann, F., Bichet, A., Ewen, K. M. & Bernhardt, R. Cytochrome P450 systems-biological variations of electron transport chains. *Biochim. Biophys. Acta Gen Subj.* **1770**, 330–344 (2007).
58. Besse, P. *et al.* Degradation of morpholine and thiomorpholine by an environmental *Mycobacterium* involves a cytochrome P450. Direct evidence of intermediates by in situ 1H NMR. *J. Mol. Catal. B Enzym.* **5**, 403–409 (1998).
59. Sielaff, B. & Andreessen, J. R. Kinetic and binding studies with purified recombinant proteins ferredoxin reductase, ferredoxin and cytochrome P450 comprising the morpholine mono-oxygenase from *Mycobacterium* sp. strain HE5. *FEBS J.* **272**, 1148–1159 (2005).
60. Iwaki, H., Hasegawa, Y., Wang, S., Kayser, M. M. & Lau, P. C. K. Cloning and characterization of a gene cluster involved in cyclopentanone metabolism in *Comamonas* sp. strain NCIMB 9872 and biotransformations effected by *Escherichia coli*-expressed cyclopentanone 1,2-monooxygenase. *Appl. Environ. Microbiol.* **68**, 5671–5684 (2002).
61. Kim, A., Tae-Kang, A., Choi, J. H. & Rhee, I. K. Purification and characterization of a cyclohexanol dehydrogenase from *Rhodococcus* sp. TK6. *J. Microbiol. Biotechnol.* **12**, 39–45 (2002).
62. Chen, Y. C., Peoples, O. P. & Walsh, C. T. *Acinetobacter* cyclohexanone monooxygenase: Gene cloning and sequence determination. *J. Bacteriol.* **170**, 781–789 (1988).
63. Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci. Data* **5**, 1–8 (2018).
64. Mehrshad, M. *et al.* The enigmatic SAR202 cluster up close: Shedding light on a globally distributed dark ocean lineage involved in sulfur cycling. *ISME J.* **12**, 655–668 (2018).
65. Tully, B. J., Wheat, C. G., Glazer, B. T. & Huber, J. A. A dynamic microbial community with high functional redundancy inhabits the cold, oxic seafloor aquifer. *ISME J.* **12**, 1–16 (2018).
66. Brzostowicz, P. C., Walters, D. M., Thomas, S. M., Nagarajan, V. & Rouvière, P. E. mRNA differential display in a microbial enrichment culture: Simultaneous identification of three cyclohexanone monooxygenases from three species. *Appl. Environ. Microbiol.* **69**, 334–342 (2003).
67. Mirza, I. A. *et al.* Crystal structures of cyclohexanone monooxygenase reveal complex domain movements and a sliding cofactor. *J. Am. Chem. Soc.* **131**, 8848–8854 (2009).
68. Baxi, N. N., Patel, S. & Hansoti, D. An *Arthrobacter citreus* strain suitable for degrading ϵ -caprolactam in polyamide waste and accumulation of glutamic acid. *AMB Express* **9**, 1–11 (2019).
69. Esikova, T. Z. & Taran, S. A. A novel strain *Gulosibacter* sp. B54 degrading epsilon-caprolactam and Nylon-6 oligomers. *Microbiology (Russian Federation)* **85**, 642–645 (2016).
70. Baxi, N. N. & Shah, A. K. ϵ -caprolactam-degradation by *Alcaligenes faecalis* for bioremediation of wastewater of a nylon-6 production plant. *Biotechnol. Lett.* **24**, 1177–1180 (2002).
71. Marjanovic, A. *et al.* Catalytic and structural properties of ATP-dependent caprolactamase from *Pseudomonas jessenii*. *Proteins Struct. Funct. Genet.* **89**, 1079–1098 (2021).

72. Fujita, Y., Matsuoka, H. & Hirooka, K. Regulation of fatty acid metabolism in bacteria. *Mol. Microbiol.* **66**, 829–839 (2007).
73. Harper, C. J., Hayward, D., Kidd, M., Wiid, I. & van Helden, P. Glutamate dehydrogenase and glutamine synthetase are regulated in response to nitrogen availability in *Mycobacterium smegmatis*. *BMC Microbiol.* **10**, 138. <https://doi.org/10.1186/1471-2180-10-138> (2010).
74. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
75. Besemer, J. GeneMarks: A self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* **29**, 2607–2618 (2001).
76. Casimiro-Soriguer, C. S., Muñoz-Mérida, A. & Pérez-Pulido, A. J. Sma3s: A universal tool for easy functional annotation of proteomes and transcriptomes. *Proteomics* **17**, 1700071 (2017).
77. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
78. Bankevich, A. *et al.* SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
79. Darling, A. E., Mau, B. & Perna, N. T. progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5**, e11147 (2010).
80. Aziz, R. K. *et al.* The RAST server: Rapid annotations using subsystems technology. *BMC Genomics* **9**, 1–15 (2008).
81. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
82. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294 (2016).
83. Milne, I. *et al.* Using tablet for visual exploration of second-generation sequencing data. *Brief. Bioinform.* **14**, 193–202 (2013).
84. Sullivan, M., Zakoura, N., Fordea, B., Stanton-Cook, M. & Beatson, S. Contiguity: Contig adjacency graph construction and visualisation. *PeerJ Prepr.* **3**, e1037v1. <https://doi.org/10.7287/peerj.preprints.63v1> (2015).
85. Stewart, F. J., Ottesen, E. A. & DeLong, E. F. Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. *ISME J.* **4**, 896–907 (2010).
86. Roberts, A. & Pachter, L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat. Methods* **10**, 71–73 (2013).
87. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2009).
88. Pfaffl, M. W., Tichopad, A., Prgomet, C. & Neuvians, T. Determination of most stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper. *Biotechnol. Lett.* **26**, 509–515 (2004).
89. Silver, N., Best, S., Jiang, J. & Thein, S. L. Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR. *BMC Mol. Biol.* **7**, 1–9 (2006).
90. Andersen, C. L., Jensen, J. L. & Ørntoft, T. F. Normalization of real-time quantitative reverse transcription-PCR data: A model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Res.* **64**, 5245–5250 (2004).
91. Vandesompele, J. *et al.* Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* **3**, 1–12 (2002).
92. Pfaffl, M. W. Quantification strategies in real-time PCR in The Real-time PCR Encyclopaedia A–Z of Quantitative PCR (ed. Bustin, S.A.) 87–120 (International University Line, 2004).
93. Yagoubi, N., Guignot, C. & Ferrier, D. Simultaneous determination of polyamide6 monomers: Amino 6 hexanoic acid and ϵ -caprolactam by RP HPLC. *J. Liq. Chromatogr. Relat. Technol.* **21**, 2633–2643 (1998).

Acknowledgements

Authors gratefully acknowledge Mafalda Paiva, Márcia Duarte, Miguel Pinheiro and Miguel Vaz for their technical support. We also acknowledge Ana Catarina Ferreira and Célia Manaia from Universidade Católica Portuguesa, CBQF—Centro de Biotecnologia e Química Fina, Laboratório Associado, Escola Superior de Biotecnologia (ESB) and to Nuno Pereira from Instituto Superior de Engenharia do Porto (ISEP) and António Múrias from CIBIO—Research Centre in Biodiversity and Genetic Resources, InBIO, University of Porto for their valuable help in the PFGE analysis and informatics support, respectively. We are also indebted to Patricia Siguier from the ISFinder database for her valuable help in the identification and characterization of *ISGmo1*. This work was financially supported by: Base Funding LA/P/0045/2020 (ALiCE), and UIDB/00511/2020 (LEPABE) UIDB/00511/2020, funded by national funds through the FCT/MCTES (PIDDAC) and GenomePT: National Laboratory for Genome Sequencing and Analysis (POCI-01-0145-FEDER-022184).

Author contributions

Study concept and design: O.C.N., C.E., A.R.L.; acquisition of data: A.R.L., A.T.V., C.B., D.P., J.F.; data analysis: A.R.L., E.B., A.T.V., C.B., D.P., H.F., A.M.-M.; interpretation of data: O.C.N., C.E., A.R.L., A.T.V., X.B., E.B., I.V.-M.; Wrote the manuscript: O.C.N., C.E., and A.R.L. Critical revision of the manuscript: All authors reviewed the manuscript. Obtained funding: O.C.N.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-18732-5>.

Correspondence and requests for materials should be addressed to O.C.N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022