Supplementary Discussion

Identifying CD-NTases encoded by target organisms is challenging due to high sequence divergence. Below is a guide to locating CD-NTases in a given strain or identifying specific strains/organisms that encode a CD-NTase of interest. Please also see the previous analysis by Burroughs et al., *NAR* (2015), in which the authors carried out initial operon-structure guided bioinformatics. Additionally, below is a guide to classifying the clade association of a newly described CD-NTase.

To identify if a given bacterial strain encodes an already annotated CD-NTase:

Download the supplemental table “Supplementary Table 2-CD-NTases and CD-NTase encoding bacteria.xls.” Search this table for strain names, organism species, or other database identifiers. This spreadsheet can also be searched for a specific CD-NTase to identify bacterial strains encoding that gene. Make sure to search all tabs of data as this document has multiple datasets. Upon locating the bacterial strain/gene, use the associated NCBI accession number to further locate sequence, ordered locus, and genomic information. The three tabs in Supplementary Table 2 are:

- (1) “Type CD-NTases Screened” providing information on specific enzymes screened in Fig. 4a–c and Extended Data Fig. 6c–g.
- (2) “Non-redundant CD-NTase Records” providing information on every CD-NTase sequence used for phylogenetic analysis, protein alignment construction, and Fig. 4a. These data represent the total sequence diversity of CD-NTases but each record may represent multiple bacterial isolates encoding identical proteins. The left-most column represents the order of CD-NTases in the alignment/tree and proceeds clockwise from “G”. For full alignment, see .netwick, FASTA, and .geneious files provided as source data for Figure 4a.
- (3) “Compiled CD-NTase IPG Records” providing a list of all records of CD-NTases found within the Identical Protein Groups tool at NCBI. This list is semi-redundant but contains complete strain and organism specific identifiers.

To identify CD-NTases within any given organism by BLAST:

An additional method of identifying CD-NTase genes encoded by an organism of interest is to BLAST the complete list of type CD-NTases using their NCBI identifiers, found in tab (1) of Supplementary Table 1 using these steps:

1) Copy complete list of type CD-NTase identifiers (below)
2) Using a BLAST-P search (https://blast.ncbi.nlm.nih.gov/Blast.cgi), paste the identifiers in to the box “Enter accession number”
3) Define the organism and strain to search in the query box “Organism”
4) Click “BLAST” button.

To identify CD-NTases within any given organism by conserved domain:

One can also browse for the conserved domains that describe CD-NTase family proteins. Please note that sequences annotated with these conserved domain descriptions may be other pol-β-like nucleotidyltransferases and not specifically CD-NTase family members.

- Pfam domains: Mab-21 protein domain (PF03281), PAP_central domain (PF04928), N-terminal Pol-β-like nucleotidyltransferase core domain (PF14792 and PF01909), C-terminal OAS1_C domain (PF10421), C-terminal tRNA-NucTransf2 domain (PF9249)
- EuKaryotic Orthologous Groups (KOG) database: KOG3963, KOG2245, KOG3792, KOG37933
- Clusters of Orthologous Groups (COG): COG5186, COG1746, COG1665, COG1669
- NCBI conserved domain database: CD05402, CD05400, CD5397

To classify the designated clade/cluster of a new unique CD-NTase:

Align the CD-NTase amino acid sequence of interest to the sequences in tab (2) “Non-redundant CD-NTase Records”. Sequences can also be found in alignment supplementary data files. Compare newly described CD-NTase to neighboring proteins based on alignment. A CD-NTase is considered a member of a clade/cluster if it shares >24.5% amino acid identity with other members of that clade/cluster.

Type CD-NTase identifiers for BLAST analysis:

WP_001901330.1
WP_001593454.1
WP_023121145.1
WP_016849025.1
WP_020363757.1
WP_031517737.1
KDD27955.1
WP_032579276.1
WP_016268104.1
WP_012995826.1
WP_000058223.1
WP_017897513.1
WP_023633898.1
WP_002302472.1
WP_044727581.1
WP_000995828.1
WP_005836899.1
WP_026109030.1
EFJ98156.1
WP_001534692.1
WP_023223657.1
WP_005110610.1
YP_635404.1
WP_003090158.1
WP_001279388.1
WP_000246637.1
WP_031517014.1
WP_000246636.1
EIQ80517.1
WP_044779457.1
WP_008409465.1
WP_002106335.1
WP_000019626.1
WP_000763718.1
WP_003305997.1
WP_009895113.1
WP_043964485.1
EEH69894.1
WP_032676400.1
NP_766712.1
WP_023727438.1
WP_004556387.1
WP_032942206.1
WP_001056752.1
WP_031656629.1
WP_044359458.1
WP_013858317.1
WP_015376200.1
EKS31071.1
WP_006018769.1
WP_008253492.1
WP_012997810.1
WP_023568228.1
WP_000072410.1
WP_017790128.1
WP_006482377.1
WP_001593458.1
WP_042646516.1
WP_041847730.1
WP_000899483.1
WP_062726309.1
WP_054878246.1
WP_009654824.1
EGH79124.1
WP_009929206.1
WP_000102010.1
WP_002347527.1
WP_014072508.1
WP_016200549.1

Supplementary Table 1. Summary of data collection, phasing and refinement statistics

	Rm-CdnE ***Apo*** (6E0K)	**Rm-CdnE** **Upnpp, Apcpp** (6E0L)	**Rm-CdnE** ***Apo*** (Se-SAD)	**Em-CdnE** ***Apo*** (6E0M)	**Em-CdnE** **GTP, Apcpp** (6E0N)	**Em-CdnE** **pppA[3′–5′]pA** (6E0O)	**Em-CdnE** (S-SAD)	**RECON** **cAAG** (6M7K)
Data collection								
Space group	$P\ 2_12_12_1$	$P\ 2_12_12_1$	$P\ 2_12_12_1$	$P\ 2_12_12_1$	$P\ 2_12_12_1$	$P\ 2_12_12_1$	$P\ 2_12_12_1$	$P\ 2_12_12_1$
Cell dimensions								
a, b, c (Å)	52.53, 66.33, 89.21	51.65, 65.65, 88.86	52.70, 66.60, 89.37	57.19, 58.23, 99.45	57.02, 58.24, 99.52	57.07, 58.61, 99.48	56.85, 58.54, 99.60	50.60, 57.07, 110.76
α, β, γ (°)	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0
Wavelength	1.00001	0.97910	0.97940	0.97920	0.97918	0.97918	1.71370	0.97910
Resolution (Å)[a]	37.39–1.60 (1.63–1.60)	36.92–2.25 (2.32–2.25)	45.40–2.29 (2.37–2.29)	49.58–1.52 (1.54–1.52)	37.83–1.50 (1.52–1.50)	37.92–1.25 (1.27–1.25)	37.93–1.99 (2.05–1.99)	46.02–1.10 (1.12–1.10)
R_{pim}	2.4 (12.1)	6.8 (30.4)	2.6 (26.9)	3.3 (62.3)	4.3 (74.5)	2.5 (45.7)	1.5 (6.1)	2.9 (29.8)
$I/\sigma(I)$	18.0 (4.6)	9.8 (3.0)	26.1 (2.9)	13.6 (1.4)	10.6 (1.2)	13.9 (1.6)	35.7 (14.3)	14.0 (2.6)
$CC_{1/2}$	99.9 (96.0)	99.2 (81.7)	99.9 (80.3)	99.9 (54.7)	99.8 (42.6)	99.9 (63.1)	99.9 (97.4)	99.9 (81.3)
Completeness (%)	99.7 (95.0)	99.1 (94.7)	97.5 (79.9)	99.8 (96.1)	99.7 (95.0)	99.9 (98.7)	95.7 (86.3)	99.7 (97.5)
Redundancy	6.7 (5.4)	5.3 (4.4)	81.1 (53.2)	12.9 (8.5)	6.7 (6.0)	6.7 (6.4)	72.8 (62.3)	9.1 (8.2)
Refinement								
Resolution (Å)	37.39–1.60	36.92–2.25		49.58–1.52	37.83–1.50	37.92–1.25		46.02–1.10
No. reflections								
Total	282,982	78,899		669,261	361,682	621,541		1,185,904
Unique	41,925 (1,925)	14,799 (1,263)		52,028 (2,474)	53,857 (2,464)	92,915 (4,497)		130,194 (6242)
Free (%)	5	5		3.9	3.9	3.9		2
R_{work} / R_{free}	16.4 / 18.6	18.2 / 21.7		13.9 / 17.6	15.5 / 18.0	14.2 / 16.4		14.3 / 15.7
No. atoms								
Protein	2457	2426		2256	2239	2328		2644
Ligand		61		9 (PPi)	96	63		83 (cAAG, EtGl)
Water	442	177		326	343	377		537
B factors								
Protein	21.6	23.8		19.8	19.6	17.66		9.9
Ligand		38.9		27.8	38.9	32.54		12.5
Water	33.0	31.4		34.4	33.1	31.96		24.5
r.m.s deviations								
Bond lengths (Å)	0.006	0.008		0.008	0.005	0.007		0.011
Bond angles (°)	0.802	1.06		1.11	1.06	1.30		1.421

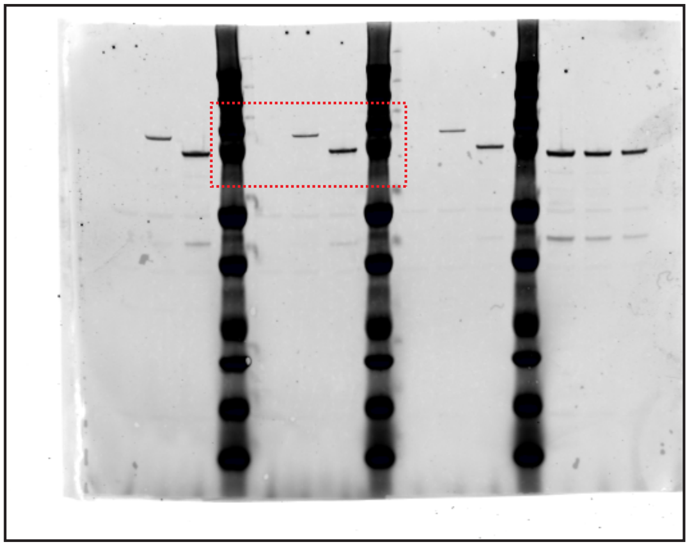
Single crystals were used to collect data for each structure.

[a] Values in parentheses are for highest-resolution shell.

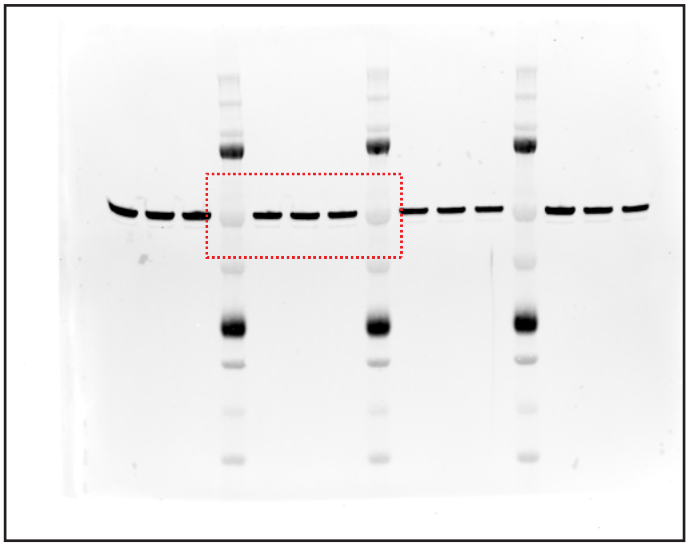
Supplementary Figure 1

Original source images for data obtained by electrophoretic separation
Extended Data Figure 5d

700 channel



800 channel



Western blot of one gel/membrane probed with 2 sets of primary and secondary antibodies simultaneously, visualized in two fluorescence channels

Primary antibodies 1:5,000 Rabbit anti-MBP (Millipore Cat# AB3596, RRID:AB_91531) and 1:10,000 Mouse anti-Tubulin (Millipore Cat# MABT205, RRID:AB_11204167)

Secondary antibodies 1:10,000 IRDye 680RD Goat anti-Rabbit IgG (LI-COR Biosciences Cat# 925-68071, RRID:AB_2721181) and IRDye 800CW Goat anti-Mouse IgG (LI-COR Biosciences Cat# 925-32210, RRID:AB_2687825)

Ladder Bio-rad Precision Plus Protein Dual Color Standard

Approximate location of cropped images
Extended Data Figure 5d