



## OPEN ACCESS

## EDITED BY

Philippe M. Oger,  
UMR5240 Microbiologie, Adaptation et  
Pathogenie (MAP), France

## REVIEWED BY

Masao Inoue,  
Ritsumeikan University, Japan  
Likui Zhang,  
Yangzhou University, China  
Wolfgang R. Streit,  
University of Hamburg, Germany

## \*CORRESPONDENCE

Adele Williamson  
✉ adele.williamson@waikato.ac.nz

## SPECIALTY SECTION

This article was submitted to  
Extreme Microbiology,  
a section of the journal  
Frontiers in Microbiology

RECEIVED 01 February 2023

ACCEPTED 23 March 2023

PUBLISHED 14 April 2023

## CITATION

Rzoska-Smith E, Stelzer R, Monterio M,  
Cary SC and Williamson A (2023) DNA repair  
enzymes of the Antarctic Dry Valley  
metagenome.  
*Front. Microbiol.* 14:1156817.  
doi: 10.3389/fmicb.2023.1156817

## COPYRIGHT

© 2023 Rzoska-Smith, Stelzer, Monterio, Cary  
and Williamson. This is an open-access article  
distributed under the terms of the [Creative  
Commons Attribution License \(CC BY\)](#). The  
use, distribution or reproduction in other  
forums is permitted, provided the original  
author(s) and the copyright owner(s) are  
credited and that the original publication in this  
journal is cited, in accordance with accepted  
academic practice. No use, distribution or  
reproduction is permitted which does not  
comply with these terms.

# DNA repair enzymes of the Antarctic Dry Valley metagenome

Elizabeth Rzoska-Smith<sup>1</sup>, Ronja Stelzer<sup>1</sup>, Maria Monterio<sup>2</sup>,  
Stephen C. Cary<sup>2</sup> and Adele Williamson<sup>1\*</sup>

<sup>1</sup>Proteins and Microbes Laboratory, School of Science, University of Waikato, Hamilton, New Zealand,  
<sup>2</sup>Thermophile Research Unit, School of Science, University of Waikato, Hamilton, New Zealand

Microbiota inhabiting the Dry Valleys of Antarctica are subjected to multiple stressors that can damage deoxyribonucleic acid (DNA) such as desiccation, high ultraviolet light (UV) and multiple freeze-thaw cycles. To identify novel or highly-divergent DNA-processing enzymes that may enable effective DNA repair, we have sequenced metagenomes from 30 sample-sites which are part of the most extensive Antarctic biodiversity survey undertaken to date. We then used these to construct wide-ranging sequence similarity networks from protein-coding sequences and identified candidate genes involved in specialized repair processes including unique nucleases as well as a diverse range of adenosine triphosphate (ATP) -dependent DNA ligases implicated in stationary-phase DNA repair processes. In one of the first direct investigations of enzyme function from these unique samples, we have heterologously expressed and assayed a number of these enzymes, providing insight into the mechanisms that may enable resident microbes to survive these threats to their genomic integrity.

## KEYWORDS

Antarctic Dry Valleys, metagenome, sequence similarity network, DNA repair, nuclease, DNA ligase

## Introduction

New areas of exploration and advances in detection methods during the last decade have revealed the diversity of life surviving in regions previously thought to be uninhabitable (Cary et al., 2010; Christner et al., 2014; Yung et al., 2014; Wei et al., 2016). One example are the McMurdo Dry Valley systems of Victoria Land Antarctica, the coldest and driest places on Earth with an environment that has been compared to the surface of Mars (Cary et al., 2010; Stomeo et al., 2012). Despite this, metagenomic surveys have demonstrated that the Dry Valleys support a low-complexity but diverse ecosystem which is dominated by bacteria and appears to be extremely reactive to change (Cary et al., 2010; Tiao et al., 2012; Yung et al., 2014; Chown et al., 2015; Wei et al., 2016; Lee et al., 2018, 2019; Niederberger et al., 2019). Abiotic stressors in this environment present particular challenges to DNA integrity, including multiple freeze-thaw cycles and desiccation leading to lethal double-stranded DNA breaks (Pitcher et al., 2007a; Stomeo et al., 2012), nutrient scarcity which limits DNA synthesis (Gourse and Keck, 2007) and UV radiation which induces base-dimer lesions (Goosen and Moolenaar, 2008).

Bacteria comprising environmental communities and inhabiting extreme environments are often slow-growing, adapted to low-nutrient conditions and include an enormous species diversity with many uncultured representatives (Stomeo et al., 2012; Christner et al., 2014; Le et al., 2016). In contrast, the model organisms on which our present understanding of DNA repair processes is based are predominantly pathogens and mesophiles from habitats of moderate temperature that are able to grow rapidly on rich media in pure culture (Aravind et al., 1999; Kreuzer, 2013). In these well-characterized bacterial repair systems, single-stranded DNA damage such as chemically modified bases, bulky UV-induced lesions and mismatches caused by replication errors are

targeted by the excision-repair pathways that remove and replace the affected nucleotides, or by direct chemical reversal of the damage (Goosen and Moolenaar, 2008; Perugino et al., 2015; Brissett et al., 2020; Wozniak and Simmons, 2022). Double-strand breaks caused by collapsed replication forks, prolonged desiccation or ionizing radiation are repaired using RecA-dependent homologous recombination, which requires the presence of an intact complementary DNA copy (Kuzminov, 1999; Wang et al., 2022). Some species are also able to repair double-strand breaks directly *via* an error-prone non-homologous end-joining system that utilizes the Ku-dependent multifunctional ATP-dependent DNA ligase/polymerase/nuclease enzyme LigD (Bowater and Doherty, 2006; Shuman and Glickman, 2007).

Beyond these canonical mechanisms, studies of cultivatable DNA-damage resistant bacteria have identified novel clade-specific DNA repair systems. These include condensed nucleoid-dependent end-joining by the radiation-resistant extremophile *Deinococcus radiodurans*, which involves the *Deinococcus*-specific PprA DNA-binding protein and a dedicated DNA ligase (Blasius et al., 2007, 2008; Ishino and Narumi, 2015). In the last two decades, exponential growth of whole-genome sequences from cultivated extremophile microorganisms has provided an opportunity to take a sequence-driven approach to identify novel DNA-repair mechanisms. In one example, subsequent to the discovery of the LigD-dependent non-homologous end-joining pathway in bacteria, ATP-dependent DNA ligases have been annotated in the genomes of numerous bacteria in addition to their replicative NAD-dependent forms (Wilkinson et al., 2001; Williamson et al., 2016). Some of these, the LigC ligases, are involved in alternate base-excision pathways (Płociński et al., 2017) while for others their roles in specific repair pathways and possible interaction partners *in vivo* remain to be determined. In another recent example, a novel mismatch-specific endonuclease NucS was identified in *Mycobacteria* as well as species of Archaea, which appears to compensate for the lack of key proteins in the highly-conserved mismatch repair pathway (Castañeda-García et al., 2017; Ishino et al., 2018; Wozniak and Simmons, 2022).

Given the diversity of known mechanisms present in culturable, predominantly mesophilic, species investigated to date it is highly likely that divergent or entirely novel mechanisms operate in more taxonomically distant or niche-selected bacteria, such as Dry Valleys microbiota. DNA repair pathways operating at low temperature are particularly interesting, as psychrophilicity and psychrotolerance generally imply slow rates of replication and rationing of cellular resources (De Maayer et al., 2014; Nunn et al., 2014), meaning a second replicating chromosome may not be available to for homologous recombination-mediated repair (Kuzminov, 1999). Hyperthermophile-specific DNA modifying enzymes have been identified that are thought to be adaptations to DNA repair and protection at high temperatures (Perugino et al., 2009), and it is likely that specific mechanisms also exist for cold-tolerance.

The samples analyzed in this work were collected as part of the most comprehensive terrestrial survey of Antarctic bacterial ecology undertaken to date including extensive data on topographic features and physicochemical parameters in addition to biological composition (Lee et al., 2019; Bottos et al., 2020). These previous studies, which evaluated bacterial richness and community structure through sequencing ribosomal RNA, have shown that environmental factors such as salinity and moisture content are the main drivers influencing taxonomic richness and community composition; however biotic

factors, in particular carbon fixation by cyanobacteria, remain essential determinants of ecosystem complexity (Lee et al., 2019; Bottos et al., 2020). In the present work we have undertaken further exploration of the genomic content of these samples by sequencing metagenomes from a sub-set of sites with a view to understanding adaptation to this harsh climate at the molecular level. Here we report the first detailed systematic exploration of DNA repair pathways from Antarctic metagenomes, and one of the first attempts at recombinant protein production from these unique samples.

## Methods

### Sample collection, DNA extraction and sequencing

Samples were collected from the Miers, Marshall, and Garwood valleys during the New Zealand Terrestrial Antarctic Biocomplexity Survey (nzTABS, <https://ictar.aq/nztabs-science/>) as part of a wide-ranging study into the influence of geochemistry, geology, climate, and biotic factors on the complexity of terrestrial ecosystems. Sample collection, geological characterization and environmental metadata of the sample site have been described previously (Lee et al., 2019; Bottos et al., 2020). Briefly, the site includes a predominantly ice-free area over 200 km<sup>2</sup> with a mean annual temperature of −20°C and precipitation of less than 10 cm per year. Sample sites were selected as representatives of geographical tiles delineated by their topographic and geologic attributes and samples were collected over two successive austral summers (January 2009 and 2010). For DNA extraction, soil samples were resuspended in ultra-pure water or Tris HCl and DNA was extracted as described using either a modified Ctab method or *via* MoBio's Power Soil DNA Isolation protocol (Qiagen; Lee et al., 2019; Bottos et al., 2020). Purified DNA was shipped to the Joint Genome Institute (JGI) for sequencing, subsequent to passing quality controls as recommended by guidelines provided on the JGI website.<sup>1</sup> Sequencing was carried out according to the standard JGI metagenomic workflow. Briefly, whole genome shotgun sequences were sequenced from 300 bp paired-end libraries using an Illumina HiSeq 2,500 and reads were assembled using MEGAHIT v. 1.0.6 and annotated using the IMG Annotation Pipeline v.4. 13.0 (Markowitz et al., 2014). Final minimal drafts of 31 assembled metagenomes ranging in size from 46,280–5,011,667 genes (59,690,366 total) are publicly available through the JGI Genomes Online Database (GOLD). GOLD analysis project identifiers are given in the table in [Supplementary 1](#) along with gene counts at each site for proteins discussed in detail here.

### Profile searching and retrieval

Raw Hidden Markov Model (HMM) profiles for Pfam searches were downloaded from the European Molecular Biology Laboratory (EMBL)'s Pfam site (<http://pfam.xfam.org/>, accessed May 2019). The

<sup>1</sup> <https://jgi.doe.gov/user-programs/pmo-overview/project-materials-submission-overview/>

TABLE 1 Numbers of DV-metagenome sequences retrieved from HMM searches for each Pfam domain.

First search (catalytic domain)			Second search (appending domains)		
Family name	Pfam identifier (Interpro identifier)	# Sequences <sup>a</sup>	Family name	Pfam identifier (Interpro identifier)	# Sequences <sup>b</sup>
NucS	PF01939 (IPR002793)	8,119	-	-	-
Hjc	PF01870 (IPR002732)	2,819	-	-	-
Rad52_Rad22	PF04098 (IPR041247)	1,621	-	-	-
DdrB	PF12747 (IPR024305)	271	-	-	-
UvdE	PF03851 (IPR004601)	2,743	-	-	-
DNA_ligase_A_M	PF01068 (IPR012310)	23,640	-	-	20,044
			DNA_ligase_A_N	PF04675 (IPR012308)	1791
			LigD_N +/- PrimaseS	PF13298 +/- PF01896 (IPR014144 +/- IPR002755)	1805
DNA_PhotoLyase	PF00875 (IPR006050)	9,373	-	-	9,373
HhH-GPD	PF00730 (IPR003265)	24,895	-	-	23,450
			AlkA_N	PF06029 (IPR010316)	1,445
MutS_I	PF01624 (IPR007695)	3,947	-	-	1942
			MutS_II +/- MutS_II +/- I MutS_V	PF05188 +/- PF05192 +/- PF00488 (IPR007860 +/- IPR007696 +/- IPR000432)	2005

<sup>a</sup>Total number of sequences with the catalytic domain used in the first search; <sup>b</sup>Number of sequences in each sub-group with a particular appending domain (where present).

list of Pfam domains used in this initial search is given in Table 1. Predicted protein-coding sequences from each Antarctic metagenome were downloaded from the JGI's IMG/MER web portal as FASTA sequences<sup>2</sup> and searched with the HMM profiles using the hmmsearch program from the HMMER suite 3.2.1.<sup>3</sup> Sequences above the inclusion threshold ( $E$ -value <1) were downloaded in FASTA format. Where the Pfam family used in the initial search was known to be part of one or more multi-domain proteins, a second search was performed to create sub-sets with up to three additional domains (Table 2).

As Pfam has now been merged with and assimilated into the Interpro database, equivalent Interpro identifiers are included in the table.

## Sequence similarity network construction and analysis

Sequence similarity networks were constructed for each set of sequences identified by hmmsearch using the EFI-EST server<sup>4</sup> 'Option C, Generate SSN from provided FASTA sequence' (Zallot et al., 2019). To distinguish groups of sequences that were unique to the Antarctic metagenomes, a proportion of randomly-selected NCBI sequences was included in the SSN using the 'Protein Family Addition Option'.

A fraction value was set to give 25–35% NCBI sequences from the UniRef database in the total SSN, or, for particularly large families a fraction giving >500 UniRef sequences. The initial SSN used a permissive  $E$ -value of five for edge calculation, and alignment scores were set during the finalization step to give between 2 and 4 million edges. Minimum length thresholds were set based on the model length of the HMM profile used for the initial search, or in the case of multi-domain proteins, on the sum of the model lengths. No maximum length was set.

The full network for each SSN was downloaded for further processing in Cytoscape 3.8.2.<sup>5</sup> The yFiles organic layout was applied to optimize node layout and visualize separate clusters, while sequences were colored by source as being of metagenome origin or from UniRef. The network was refined by progressively increasing the percentage similarity edge threshold until nodes resolved into distinct clusters. Refinement was considered complete when one of the following criteria was met: nodes had resolved into clusters where metagenomic sequences and UniRef sequences were separated; the edge threshold exceeded 50% identity. Sequence similarity networks where metagenome-only clusters were evident prior to these latter two conditions being met were considered to contain potentially novel metagenome DNA repair candidate genes and were analyzed further. SSNs where metagenome and UniRef sequences remained mixed to

2 <https://img.jgi.doe.gov/cgi-bin/mer/main.cgi>

3 <http://hmmer.org/>

4 <https://efi.igb.illinois.edu/efi-est/>

5 [https://cytoscape.org/release\\_notes\\_3\\_8\\_2.html](https://cytoscape.org/release_notes_3_8_2.html)

TABLE 2 Nodes used in Sequence Similarity Network construction for different Pfam assignments.

Family names	Length threshold (aa) <sup>a</sup>	Total nodes in SSN <sup>b</sup>	# MG nodes <sup>c</sup>	# NCBI nodes (as %) <sup>d</sup>	Sequence identity (%) <sup>e</sup>	# Connected (Fragmentation %)
NucS	200	2094	1,578	516 (24.6)	28	366 (17)
Hjc	45	2,844	1976	868 (30.5)	26	427 (15)
Rad52_Rad22	150	1,340	665	675 (51.8)	26	75 (6)
Rad52_Rad22 (high ID threshold)					50	151 (11)
DdrB	80	302	210	92 (30.5)	25	151 (30)
UvdE	270	819	422	397 (48.5)	54	92 (21)
DNA_ligase_A_M	300	2,998	2,477	521 (17.4)	64	190 (7)
DNA_ligase_A_M + DNA_ligase_A_N	450	1,145	700	445 (38.9)	50	203 (7)
DNA_ligase_A_M + LigD_N +/- PrimaseS	500	1,005	650	355 (35.3)	54	78 (7)
DNA_PhotoLyase	160	6,556	3,933	2,623 (40.0)	40	679 (10)
HhH-GPD	110	25,680	18,321	7,359 (28.7)	25	69 (17)
HhH-GPD + AlkA_N	200	7,444	1,030	6,414 (86.2)	40	4,357 (3)
MutS_I	100	3,662	953	2,709 (74.0)	50	223 (11)
MutS_I + MutS_II +/- MutS_II +/- I MutS_V	200	4,510	2094	2,416 (53.6)	50	413 (21)

<sup>a</sup>Length cut-off applied to sequences before SSN construction; <sup>b</sup>Number of nodes above the length threshold used in final SSN; <sup>c</sup>Number of DV-metagenome nodes in the SSN; <sup>d</sup>Number of NCBI nodes and their percentage in the SSN; <sup>e</sup>Percentage identity threshold for edges retained in the SSN; <sup>f</sup>Number of connected components (i.e., clusters or single nodes) and percentage fragmentation calculated as total # nodes/# connected components.

the point of network fragmentation or at higher edge values were considered unlikely to include unique metagenomic representatives.

## Selection of candidate genes for recombinant expression

Where defined clusters comprised primarily metagenome representatives, the sequences in these clusters were investigated further for their potential participation in novel or as-yet undescribed DNA repair pathways. FASTA sequences within metagenome clusters were first analyzed by hmmscan (HMMER suite 3.2.1; <http://hmmer.org/>) to exclude any spurious hits which were included due to the permissive threshold used in the initial hmmscan. Sequences from unique clusters were then aligned using ClustLW 2.1 and partial sequences lacking probable start or stop codons or those that aligned poorly were discarded. For the three groups of protein selected for recombinant expression, the longest contig in the group was submitted to the PHASTER server (<http://phaster.ca/>) to detect genes that were part of bacteriophage rather than bacterial chromosomes (Arndt et al., 2016). From the shortlisted genes, priority for cloning was given to sequences from large contigs to enable future investigation of adjacent genes for related functions.

## Recombinant protein expression and purification

Clonal genes encoding proteins of interest were ordered from Twist Biosciences in the pTwist-ENTR vector with codon optimization

for *Escherichia coli*. Ordered constructs included an N-terminal His-tag and cleavage recognition site for the tobacco mosaic virus protease (TEV-protease) for tag removal. Genes were sub-cloned into the expression vectors pDEST17 (Invitrogen) and pHMGWA (GenBank #Eu680841) using the Gateway™ LR reaction kit (ThermoFisher) according to the manufacturer's instructions. Resulting expression constructs were transformed into chemically-competent DH5α cells for propagation.

To generate truncated DV-1-1-lig and DV-1-1-nuc genes, insert *att* sites for gateway cloning and to add the TEV cleavage site to the DV-1-1-lig product, a two-step PCR protocol was used with primer sequences given in Supplementary 2. In the first round the pDONR221 plasmid containing the DV-1-1-Nuc-Lig gene insert was used as a template to amplify both DV-1-1-Nuc segment (primer pair DV1-1Nuc Forward and DV1-1Nuc Back) and the DV-1-1-Lig segment (primer pair DV1-1Lig Forward and DV1-1Lig Back). PCR products were visualized on a 1% agarose gel and bands, corresponding to the correct PCR product size (1.73 bp for DV-1-Lig and 1.248 bp for DV-1-1-Nuc), were excised and extracted using the QIAquick®, Gel extraction kit (QIAGEN). The second PCR step added *att*-sites to DV-1-1-Nuc (primer pair DV1-1 Forward-2, DV1-1Nuc Back) and DV-1-1-Lig (primer pair DV1-1 Forward-2, DV1-1Lig Back) generating products suitable for Gateway cloning (1.765 bp for DV-1-Lig and 1.279 bp for DV-1-1Nuc). The PCR products were cloned into the Gateway donor vector, pDONR221 (Invitrogen), using the Gateway™ BP reaction kit (ThermoFisher) and transformed into chemically-competent DH5α cells, then further sub-cloned into the expression vector pDEST17 as described above.

To determine optimal production conditions, expression from each plasmid was tested in the *E. coli* expression strains BL21(DE3)-pLysS (Novagen), Origami (DE3; Novagen) and ArcticExpress (DE3; Aligent) at 15, 18, and 20°C. For small-scale expression trials, 50 mL cultures of cells were grown in conical flasks on Terrific Broth, 50 µg/mL ampicillin at 37°C with shaking at 180 rpm. Upon reaching an OD<sub>600</sub> between 0.3 and 0.4 the temperature was adjusted and cells were equilibrated for 30 min before addition of 0.5 mM IPTG to induce expression. Cells were harvested after 18 h by centrifugation, resuspended in 2 mL lysis buffer (50 mM Tris pH 8.0, 750 mM NaCl, 1 mM MgCl<sub>2</sub>, 5% v/v glycerol) and lysed by sonication on ice. Insoluble material was pelleted by centrifugation and the soluble fraction was incubated with 20 µL of nickel beads (Cytiva) for 15 min and recovered by centrifugation followed by two washes with lysis buffer. Protein-bound nickel beads, insoluble fractions and soluble fractions were electrophoresed on 12% SDS-PAGE gels with successful expression being indicated by a strong band at the expected molecular weight in the nickel-bead fraction and, in some cases, the soluble sample. Expression as inclusion bodies was indicated as a band at this molecular weight in the insoluble sample.

Optimal conditions (listed in [Supplementary 3](#)) were used for large-scale cultivation for purification. Cultures (1–4L) were grown at the appropriate temperatures as identified from the small-scale cultures and harvested by centrifugation. Cells were lysed by sonication in lysis buffer and clarified by centrifugation and passage through a 0.45 µm filter. Protein purification was carried out at 4°C on an Äkta Prime FPLC as described previously ([Williamson and Pedersen, 2014](#)). Briefly, clarified lysate with an additional 10 mM imidazole was purified by immobilized affinity chromatography (IMAC) using 5 mL His-Trap HP column (Cytiva) and washed with 5–10 column volumes of buffer A (50 mM Tris pH 8.0, 750 mM NaCl, 5% v/v glycerol, 10 mM imidazole). His-tagged proteins were eluted with Buffer B (50 mM Tris pH 8.0, 750 mM NaCl, 5% v/v glycerol, 500 mM imidazole) on a gradient from 0 to 100% over 75 mL. For tag removal, fractions containing the target protein were exchanged into Buffer C (50 mM Tris pH 8.0, 100 mM NaCl, 1 mM DTT, 5% v/v glycerol) using a HiPrep 26/10 Desalting column (Cytiva) and incubated at 4°C overnight with 1 mg of TEV-protease which had been produced in-house according to published protocols ([Tropea et al., 2009](#)). Cleaved proteins were subjected to reverse IMAC by re-application to the His-Trap column. To increase purity and remove high molecular-weight aggregates, the flow-through which contained the untagged target protein was up-concentrated to less than 5 mL volume and loaded onto a either a Superdex200 16/600 (S200) or Superdex75 16/600 (S75) column. In cases where the tag was not removed by TEV cleavage, fractions from the initial gradient elution in Buffer B were pooled and used directly this gel-filtration polishing step. Final purified protein was up concentrated to 0.5–5 mg mL<sup>-1</sup>, mixed 50:50 v/v with glycerol and stored at –80°C.

## Assay of DNA binding and nuclease activity

Enzymatic activity on DNA-damage substrates ([Supplementary 4B](#)), flapped/splayed substrates ([Supplementary 4C](#)) and DNA ligase ([Supplementary 4D](#)) were analyzed by denaturing gel electrophoresis, while nuclease substrates ([Supplementary 5B](#)) and substrate binding experiments were analyzed by native

PAGE. Substrates were generated from oligonucleotides (IDT) with synthetically-incorporated DNA damages and fluorescent labels which are listed in [Supplementary 6](#). These were annealed in the combinations given in [Supplementary 7](#) as described previously ([Sharma et al., 2020](#)) using final concentrations of 80 nM for the 5' FAM-labeled probe strand and 112 nM for the unlabeled strand (damage, nuclease and double-strand break ligase substrates) or 400 nM unlabeled strands (nicked ligase substrates). All assays were carried out in 50 mM Tris pH 8.0, 50 mM NaCl, 10 mM DTT. Enzyme activity assays also included 10 mM Mg<sup>2+</sup> or Mn<sup>2+</sup> as specified, while for binding experiments this was replaced with 5 mM EDTA. For DNA ligase assays, 1 mM of nucleotide cofactor (ATP, ADP, or NAD<sup>+</sup>) was used as specified. Reactions were initiated by addition of protein and were incubated at the temperatures and times indicated. For the DNA ligase nucleotide preference experiments, enzyme was pre-incubated with unlabeled DNA for 2 h at 20°C to ensure any enzyme purified in the pre-adenylated state was turned over. After this time, labeled DNA substrate and 1 µM of cofactor (ATP, ADP, or NAD) was added to the reaction and incubated for indicated times.

Activity assays for damage, flapped/splayed and ligase substrates were quenched by addition of Quench Buffer to give a final concentration of 25% formamide, 20 mM EDTA, 0.05% bromophenol blue and heated to 95°C for 5 min before electrophoresis on 20% denaturing TBE urea gels (20% acrylamide/Bis-acrylamide 29:1, 7 M urea, 1x TBE). For binding and nuclease assays Loading buffer was added to give a final concentration of 20 mM EDTA, 0.05% bromophenol blue, 5% glycerol before the samples were electrophoresed on a 10% TBE gel (10% acrylamide/Bis-acrylamide 29:1, 1x TBE). Gels were visualized on an iBright imager (Invitrogen) using the fluorescein setting.

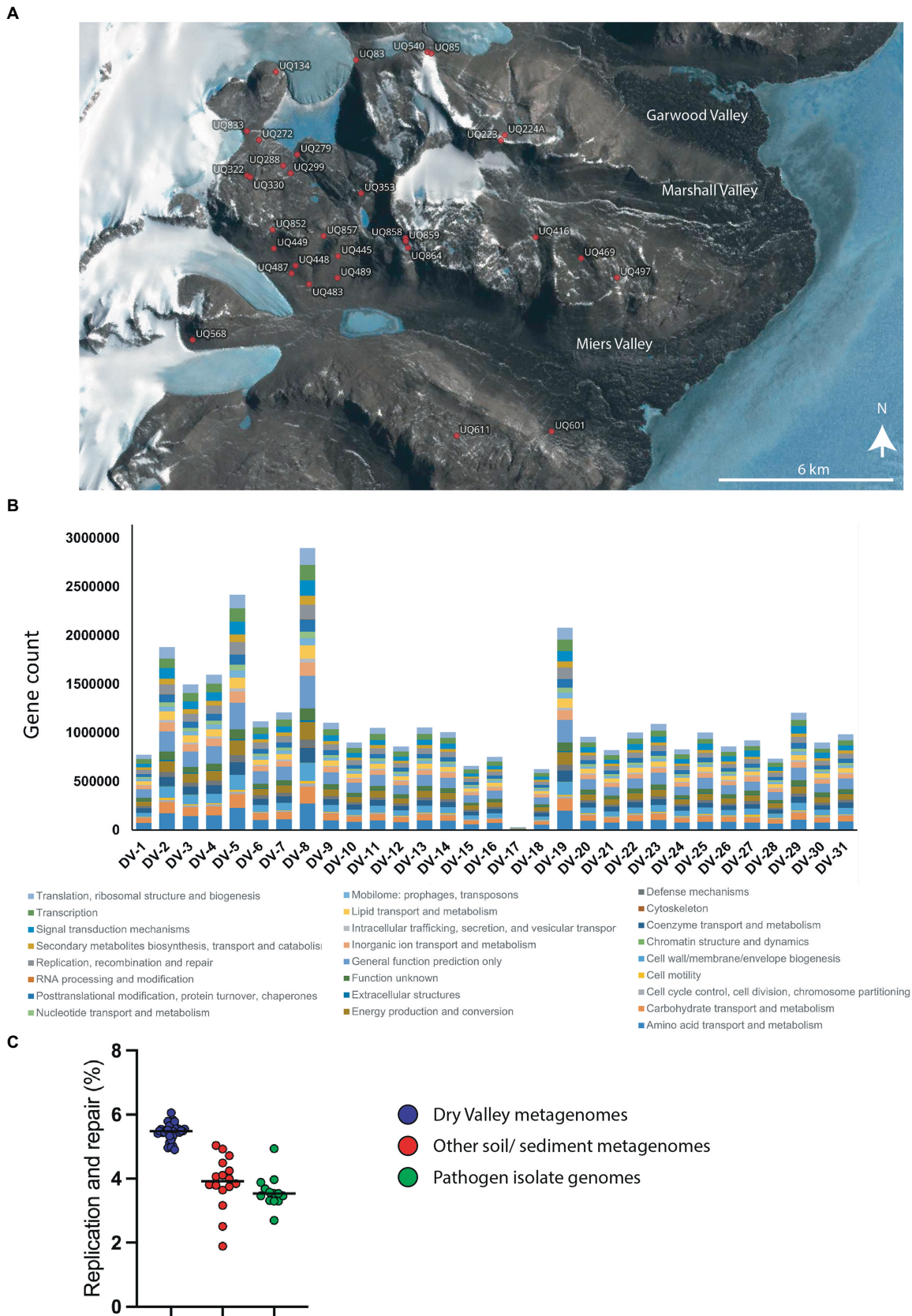
## Differential scanning fluorimetry

Thermal stability measurements of DV-Hjc were carried out by differential scanning fluorimetry (DSF) as previously described ([Ericsson et al., 2006](#)). The pH was adjusted by dilution into Britton-Robinson universal buffers over a range of 5.0–9.5 to give a final DV-Hjc concentration of 0.2 mg/mL and final SYPRO dye concentration of 1.2x. Unfolding over a temperature range from 25 to 98°C was measured using a RotoGene Q thermocycler (Qiagen) and the  $T_m$  was determined by plotting the first derivative of the melt data.

## Results and discussion

### Dry Valley metagenomes contain numerous DNA replication and repair proteins

Of the almost 60 million genes predicted in the 30 sequenced metagenomes from Dry Valleys, 58% (34,784,791 genes) were assigned to Clusters of Orthologous Groups (COG) categories ([Figures 1A,B](#)). Consistent with our hypothesis that these microbes will possess effective or specialized DNA repair systems to cope with environmental stressors, 5.5% of these annotated Dry Valley metagenome genes belonged to the 'Replication, Recombination and Repair' COG category ([Figure 1C](#)). This is higher than the proportion



**FIGURE 1**  
**(A)** Map of sample sites for metagenome sequences. The location of each sample was plotted using Quantarctica (v3.1) (Matsuoka et al., 2021).  
**(B)** Counts of genes in COG categories for metagenome sequences from each site. **(C)** Percentage of genes in the COG category 'DNA replication and repair' from Dry Valley metagenomes (blue), other soil/ sediment metagenomes (red) or genomes from isolated pathogens (green). Values for datapoints are a percentage of the total genes in COG categories for each sample and are given in [Supplementary 1](#) together with GOLD identifiers and sample type. Numbers and percentages of genes in all COG categories for all samples are shown in [Supplementary 9, 10](#), respectively.

found in other environmental metagenomes including forest soils (3.9%), other soil communities (3.6%) and freshwater sediments (4.3%). It is also considerably higher than the proportion of DNA repair genes found in the genomes of isolates of human pathogens which are the model organisms on which most of our understanding of bacteria DNA repair is based (Figure 1C).

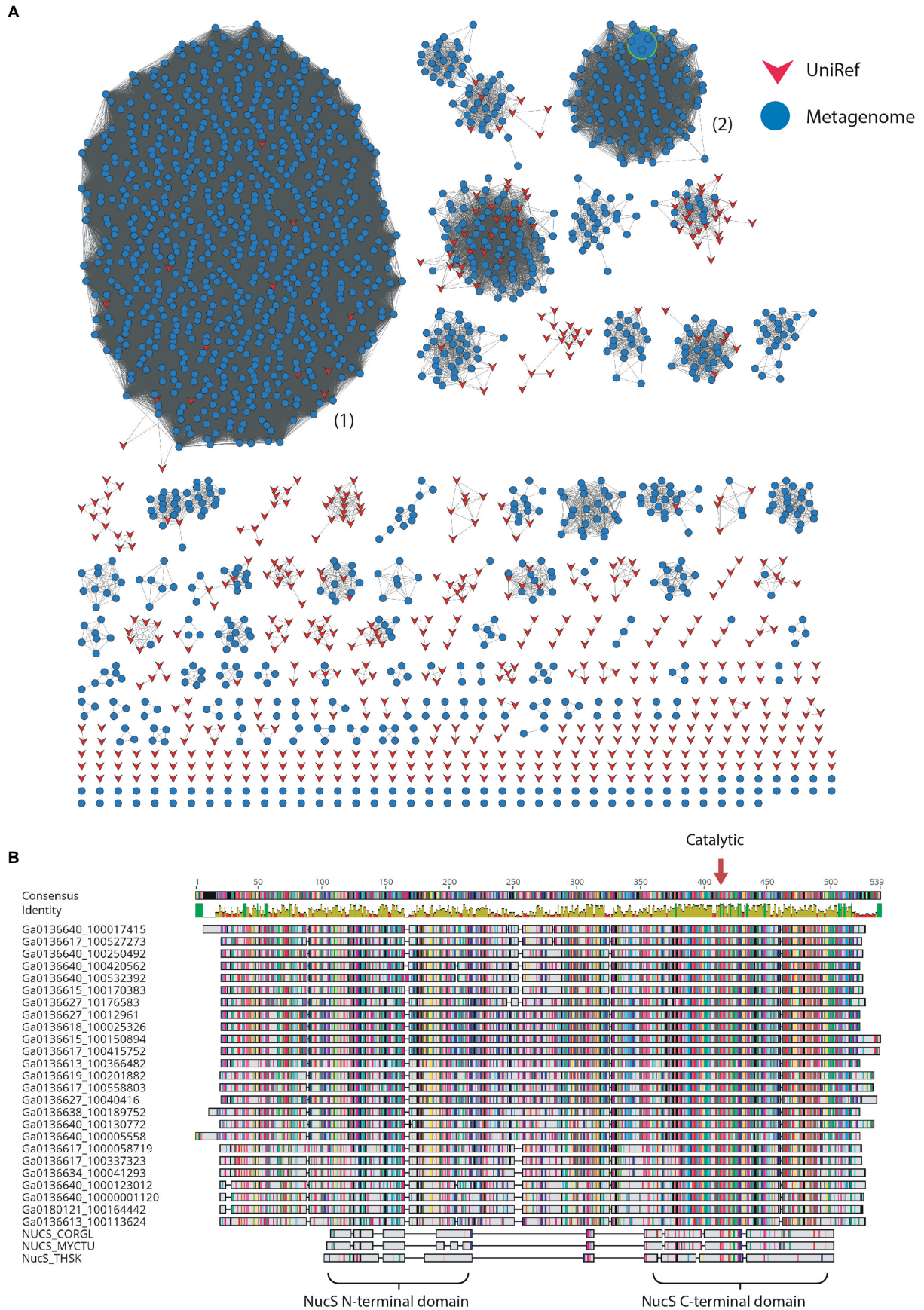
To further explore whether these sequences represent novel or divergent DNA repair enzymes, translated predicted coding genes were searched for distant homology to known DNA repair proteins using *hmmsearch*. Probe domains used in the first *hmmsearch* were chosen on the basis that (i) they comprise a large family with many representatives involved in DNA repair, or (ii) they represent a recently-described or specialist DNA repair mechanism (Pavlopoulou et al., 2016; Lilley, 2017; Wozniak and Simmons, 2022). This included single-domain proteins of the NucS/ EndoMS endonucleases (NucS; Castañeda-García et al., 2017; Ishino et al., 2018), Holliday junction resolvase-type endonucleases (Hjc; Lestini et al., 2010; Huang et al., 2015), UV-damage excision endonucleases (UvdE; Lim et al., 2019), light-activated photolyase proteins for direct reversal of UV-induced pyrimidine-pyrimidine dimers (DNA\_photolyase; Yi and He, 2013), *Deinococcus* radiation resistance proteins DdrA (Rad52\_Rad22) and DdrB (DdrB; Lim et al., 2019), as well as multi-domain families such as ATP-dependent DNA ligases (DNA\_ligase\_A\_M), excision-repair glycosylases (HhH\_GPD) and mismatch-recognition proteins (MutS\_I; Bruner et al., 2000; Otani et al., 2013; Pidugu et al., 2021). The former single-domain proteins were searched with a primary catalytic HMM profile only, while the latter multi-domain proteins were sub-divided by a second set of search criteria using up to three additional HMM domain searches. Hits were retrieved for all primary domains and sub-domains used in this search, indicating that DV-metagenomes possess a diverse array of DNA repair enzymes (Table 1). In particular, large numbers of NucS homologs (>8,000) and ATP-dependent DNA ligases (>23,000) were identified which is notable as neither pathway is ubiquitous among bacteria and both are absent from many standard mesophilic model organisms such as *E. coli* (Wilkinson et al., 2001; Creze et al., 2011; Williamson and Leiros, 2020; Zhang et al. 2020a,b). Sequences above the inclusion threshold (E-value <1) were downloaded and used to build SSNs. To minimize inclusion of partial or small fragments of proteins, length thresholds corresponding to minimum domain lengths of the search model were applied during SSN construction. For some families, this decreased the proportion of sequences remaining in the final SSN considerably, for example <20% of Nuc S and <15% of DNA ligase DNA\_ligase\_A\_M-only proteins were retained (Table 2). The final SSNs also included a proportion of NCBI sequences retrieved from UniRef which served as outgroups allowing clusters of DV-metagenome sequences with unique features to be distinguished (Table 2). The target proportion of NCBI sequences was 30–50% of the total nodes in the SSN; however, for functionally-diverse families with a large number of representatives such as proteins containing AlkA and MutS\_I domains, the number was increased to ensure that the range of sequence diversity within the family was accurately reflected. Conversely, the application of the length threshold in the DNA\_ligase\_A\_M dataset meant a large number of NCBI sequences were excluded and the final SSN included fewer than 25%.

Full SSNs with permissive (low) edge threshold values were downloaded and manually refined until either distinct DV metagenome clusters resolved, or until the percentage threshold

retained exceeded 50% sequence identity. An approximate measure of SSN cohesiveness was determined by the percentage fragmentation, calculated as the proportion of connected components (clusters or single nodes) relative to the total number of nodes. This remained under 20% for all SSNs except DdrB (30%), UvsE and MutS\_I +/-II +/-III +/-V (both 21%). Five domain families resolved unique DV-metagenome clusters at edge thresholds below 30%; NucS, Hjc, Rad52\_Rad22, DdrB and HhH-GPD, suggesting that some of these sequences may encode specialized adaptations. Four additional families, DNA\_photolyase, UvsE and the DNA ligase groups DNA\_ligase\_A\_M+A\_N and DNA\_ligase\_A\_M+LigD\_N+primase resolved unique clusters before or just above 50% sequence identity. In contrast the DNA ligase DNA\_ligase\_A\_M network was refined to more than 60% before resolution of predominantly metagenome clusters. No significant clusters of unique sequences were resolved for the MutS\_I sequences at the 50% threshold, either with or without additional domain searches, and this family was not investigated further (Supplementary 11).

## Dry Valley metagenomes encode unique groups of endonuclease proteins

Nuclease enzymes are ubiquitous participants in almost all known DNA repair pathways with endonucleases playing a key role in excision repair pathways that remove damaged or mis-matched nucleotides within DNA duplexes (Wozniak and Simmons, 2022). The NucS/EndoMS are a recently-discovered family of endonucleases that carry out an alternative mismatch repair pathway in Archaea and Bacteria that lack the canonical MutS-MutL-based system (Ishino et al., 2016; Castañeda-García et al., 2017; Ishino et al., 2018). A large number of sequences with matches to the NucS (PF01939) profile were identified in the DV metagenome, and two major DV-metagenome-containing clusters were resolved at 28% edge identity each with 761 and 115 metagenomic sequences, respectively (Figure 2A). No appending domains were detected by *hmmsearch* for sequences in either cluster, and for both the most significant homology is to the Endonuclease NucS family (Table 3). Interestingly, neither cluster was reliably annotated as NucS/EndoMS by the IMG pipeline (only 3 of the 761 Cluster (1) sequences) indicating these homologs would not have been detected by searching IMG alone. Cluster (1) includes 12 nodes from the UniRef50 family which comprise representative nodes for sequences with >50% identity in UniRef. The largest UniRef50 node represents 2,150 individual sequences from diverse phyla (Acidobacteria, Actinobacteria, Bacteroidetes, Chlamydiae, Firmicutes, Gemmatimonadetes, Proteobacteria and Rhodothermaeota) while the second largest node represents 794 individual sequences from Actinobacteria and Chlamydiae. This suggests that although the number of UniRef50 nodes included is small, this cluster in fact represents the majority of bacterial NucS sequences including *Mycobacterium tuberculosis* and *Corynebacterium glutamicum* both of which have been characterized in the past (Castañeda-García et al., 2017; Ishino et al., 2018). Cluster (2) by contrast includes only DV-metagenome sequences some of which are more than 500 amino acids in length, which is considerably longer than characterized homologs. Alignment of 25 full-length Cluster (2) sequences with sequences of characterized NucS homologs from *C. glutamicum*, *M. tuberculosis* and *Thermococcus kodakarensis*



**FIGURE 2**  
**(A)** Sequence Similarity Network for NucS-type proteins at 28% identity threshold; other network parameters are detailed in [Table 2](#). Dry-Valley metagenome nodes are colored blue, UniRef50 nodes are indicated in red. Cluster numbers discussed in the text are given in parentheses. The node  
*(Continued)*



**FIGURE 2** (Continued)

corresponding to the recombinantly produced homolog from Cluster (2) is indicated with a large green-bordered symbol. **(B)** Alignment of full-length sequences from Cluster (2) with characterized NucS/EndoMS from *Corynebacterium glutamicum* (NUCS\_CORGL), *Mycobacterium tuberculosis* (NUCS\_MYCTU) and *Thermococcus kodakarensis* (NUCS\_THSK). Domains identified in the crystal structure of *T. kodakarensis* NucS/EndoMS are indicated below the alignment and the position of the conserved catalytic aspartic acid residue is indicated with a red arrow.

indicates that the DV metagenome sequences are extended N-terminally relative to the characterized NucS homologs (Figure 2B). The best aligned portion are the final 200 C-terminal residues which correspond to the RecB-like endonuclease domain of the structurally-characterized NucS/EndoMS of *T. kodakarensis*, while there is poor alignment with the N-terminal domain which is involved in mismatch recognition and DNA binding. The 100-residue extension is highly conserved among DV-metagenome NucS homologs and does not have any counterpart in the characterized sequences. The NucS proteins of bacteria provide one recent example where previously-undetected homologs of extremophile archaeal proteins are involved in DNA-damage resistance in bacterial species. Other examples are the LigC DNA ligases and Prim-pol polymerases which participate in an alternate stationary-phase base-excision repair in some bacterial species (Płociński et al., 2017; Brissett et al., 2020). Based on this observation, we considered whether homologs of other Archaeal-type DNA repair proteins could play a role in survival by DV species of bacteria. The archaeal Holliday junction resolvase (Hjc) family includes enzymes with an equivalent function to the bacterial RuvC endonuclease that cleaves Holliday junctions formed during homologous recombination (Shin et al., 2014; Wyatt and West, 2014). These archaeal resolvases possess a restriction endonuclease-type fold which is distinct from the RuvC and bacteriophageT4-endoVII resolvase structures (Nishino et al., 2001; Wyatt and West, 2014; Lilley, 2017). Hmmscan identified almost 3,000 hits to the Hjc domain in the Dry-Valleys dataset, and SSN analysis of these sequences showed they form separate clusters from UniRef at 26% identity with three of the six clusters containing exclusively DV-metagenome sequences (Figure 3A). Hmmscan indicates that clusters 1 and 2 both have hits to unknown protein domain UPF0102 and no highly significant hits to other domains with known enzymatic functions although there are low-probability hits to both Hjc and nuclease family enzymes (Table 3). Cluster 3 has high probability hits to elongation domain (GTP\_EFTU) and is therefore discounted from further analysis. The UPF0102 family is present in a wide variety of bacterial taxa, and as these were not included as UniRef representatives in the Hjc SSN, a new network was built using the 484 reviewed UPF0102 sequences in Interpro. Metagenome-only clusters separated out at threshold of 64% identity (Figure 3B). Several of the Interpro sequences are annotated as the transcription factor YraN, for example the members of the most significant Interpro sequence cluster which contains Enterobacteriaceae representatives (Figure 3B, Cluster (i)). However, neither the Interpro nor the DV-metagenome sequences align well to the characterized YraN protein from *Bacillus subtilis*, nor to known restriction enzymes (data not shown). Alignment of a selection of full-length sequences to the archaeal Holliday junction resolvases shows that the three key catalytic residues required for nucleolytic activity in *Pyrococcus furiosus* Hjc, Asp 33, Glu 46, Lys 48 are fully or functionally conserved; however other positions important for catalysis (Glu 9, Arg 10 and Arg 25) and dimer formation (Val 24, Val35, Phe 68 and Phe 72) are not (Figure 3C; Komori et al., 2000; Nishino et al., 2001).

Although our analysis has not identified any close bacterial equivalents of the archaeal Hjc Holliday junction resolvases in Dry-Valley bacteria, a large number of more distantly-related UPF0102 were found in these metagenomes and are worthy of further investigation (Figure 4).

## Dry Valley metagenomes encode numerous homologs of *Deinococcus* repair proteins

The UvdE endonucleases are involved in the repair of UV-induced cyclobutane pyrimidine dimers (CPD) and 6–4 pyrimidine pyrimidones (6-4PP) and induced bipyrimidine photoadducts as well as oxidative damages. This protein is found in all radiation-resistant *Deinococcus* species sequenced to date, where it functions as a back-up to the nucleotide excision repair pathway (Earl et al., 2002; Timmins and Moe, 2016). Of the more-than 400 DV-metagenome sequences retrieved by searching with the UvdE HMM, 324 formed three distinct predominantly DV-metagenome clusters at 54% identity, all of which are predicted to be UvdE proteins by both hmmscan and IMG annotation (Table 3 and Supplementary 12A). A number of DV-metagenome sequences also group with the smaller UniRef sub-cluster which contains characterized bacterial UvdE sequences. Alignment of full-length representatives of all three clusters shows that despite low sequence identity, none possess any remarkable truncations or extensions relative to *Deinococcus radiodurans* UvsE (Supplementary 12B).

The Rad52/Rad22 family are single-strand annealing proteins and include the DNA Damage Response A (DdrA) single-strand annealing protein which is involved in radiation resistance by species of *Deinococcus*, as well as the Rad52 homologous recombination protein of Eukaryotes. DdrA proteins are specific to *Deinococcus* genomes and function to bind and protect single-stranded DNA ends, likely by self-associating to form heptameric rings which then form higher-order assemblies with single-stranded DNA (Harris et al., 2004; Gutsche et al., 2008; Selvam et al., 2013). In yeast, Rad52 increases the rate of assembly of the Rad51-single-strand DNA filament during homologous recombination and in humans in functions to promote strand annealing (Kowalczykowski, 2015). The 665 DV-metagenome hits to the Rad52\_Rad22 HMM identified by hmmscan formed two major clusters at a 26% identity threshold, with all sequences in both clusters annotated as hypothetical by IMG (Table 3). The first cluster of 227 DV-metagenome sequences (Figure 5A) (1) is merged with 201 UniRef sequences, with these separating from each other at a 50% identity level (Supplementary 13A,B). The majority (96%) of these UniRef sequences are also bacterial and include the DdrA protein from *D. radiodurans*. Two additional predominantly-UniRef clusters at the 26% identity level (Figure 5A) (i) and (ii) are almost exclusively eukaryotic and include the Rad52 DNA repair proteins. The second major metagenome cluster (Figure 5A) (2) comprises 303

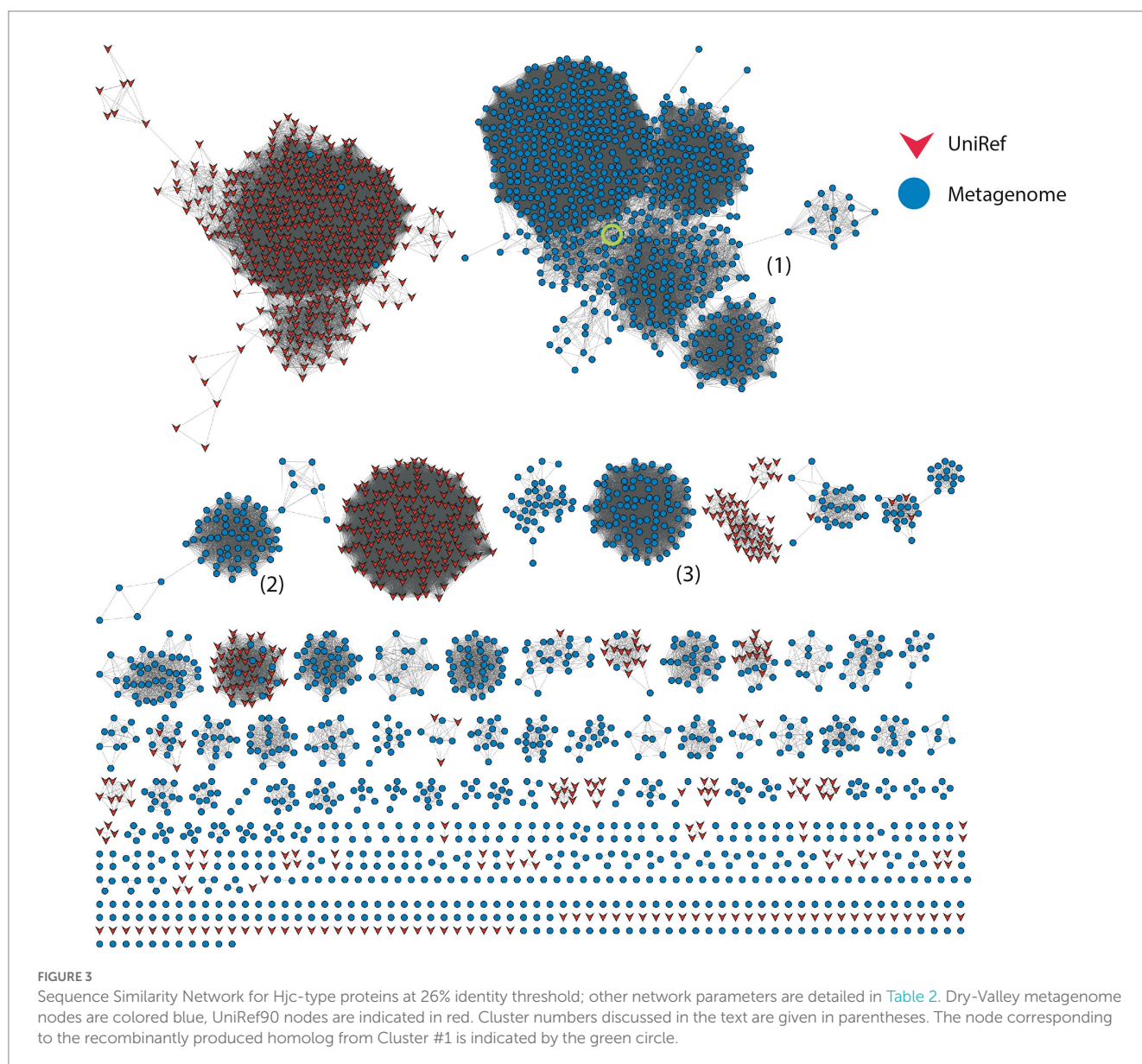
TABLE 3 Analysis of sequence similarity network clusters from Dry-Valley metagenome sequences.

Catalytic domain <sup>a</sup>	Other domains <sup>b</sup>	Cluster (DV sequence count)	Top hmmscan hit <sup>c</sup> (E-value)	Annotation from IMG	
NucS	-	1 (761)	Endonuclease NucS (7.60E-93)	Hypothetical protein	
		2 (115)	Endonuclease NucS (3.50E-06)	Hypothetical protein	
Hjc	-	1 (722)	Uncharacterised UPF0102 (5.10E-30)	Putative endonuclease	
		2 (71)	Uncharacterised UPF0102 (1.60E-29)	Putative endonuclease	
		3 (88)	GTP_EFTU (7.00E-56)	GTP-binding protein	
Rad52_Rad22	-	1 (227)	Rad52_Rad22 (1.90E-18)	Hypothetical protein	
		2 (303)	Rad52_Rad22 (3.90E-12)	Hypothetical protein	
DdrB	-	1 (9)	DdrB (7.20E-60)	DdrB-like protein	
		2 (8)	DdrB (1.20E-14)	DdrB-like protein	
		3 (17)	AICARFT_IMPCHas (1.00E-113)	Phosphoribosylaminoimidazolecarboxamide formyltransferase / IMP cyclohydrolase	
		4 (15)	GDPD (1.90E-13)	Glycerophosphoryl diester phosphodiesterase	
		5 (9)	DdrB (0.034)	Hypothetical protein	
		6 (7)	GATase_6 (5.10E-30)	Asparagine synthase (glutamine-hydrolysing)	
		7 (9)	DUF150 (1.10E-20)	Ribosome maturation factor RimP	
UvdE	-	1 (146)	UvdE (5.40E-60)	UV DNA damage endonuclease	
		2 (98)	UvdE (3.20E-70)	UV DNA damage endonuclease	
		3 (80)	UvdE (2.20E-53)	UV DNA damage endonuclease	
DNA_ligase_A_M	-	1 (345)	DNA_ligase_A_M (2.00E-39)	Non-homologous end joining protein LigD	
		2 (141)	DNA_ligase_A_M (4.30E-38)	Non-homologous end joining protein LigD	
		3 (70)	DNA_ligase_A_M (1.30E-34)	Non-homologous end joining protein LigD	
		4 (47)	SecA_DEAD (2.30E-130)	Preprotein translocase subunit SecA	
		DNA_ligase_A_N	1 (145)	DNA_ligase_A_M (3.70E-47)	DNA ligase I
		LigD_N +/- PrimaseS	1 (118)	LigD_N (5.60E-37)	Non-homologous end joining protein LigD
			2 (34)	DNA_ligase_A_M (3.20E-37)	Non-homologous end joining protein LigD
			3 (23)	LigD_N (3.40E-38)	Non-homologous end joining protein LigD
			4 (24)	LigD_N (1.30E-38)	Non-homologous end joining protein LigD
5 (14)	LigD_N (7.10E-34)		Non-homologous end joining protein LigD		
DNA_PhotoIyase	-	1 (124)	E1-E2_ATPase (7.1E-58)	Cu <sup>2+</sup> –exporting ATPase	
		2 (80)	HAD_2 (9.20E-14)	Uncharacterized protein	
		3 (46)	SIR2 (6.6E-45)	NAD-dependent deacetylase	
HhH-GPD	-	1 (1766)	HhH-GPD (1.20E-17)	Endonuclease-3	
		2 (1137)	PD40 (4.90E-38)	DNA-3-methyladenine glycosylase II	
		3 (458)	HhH-GPD (2.30E-09)	N-glycosylase/DNA lyase	
		4 (1500)	HhH-GPD (5.90E-18)	HhH-GPD superfamily base excision DNA repair protein	
		5 (488)	HhH-GPD (2.80E-08)	3-methyladenine DNA glycosylase/8-oxoguanine DNA glycosylase	
		6 (507)	HhH-GPD (6.50E-18)	Base-excision DNA repair protein	
		7 (673)	HhH-GPD (3.00E-13)	DNA-3-methyladenine glycosylase II/AraC family transcriptional regulator	
		8 (382)	Adenine_glyco (1.40E-83)	DNA-3-methyladenine glycosylase I	
		AlkA_N	1 (912)	AlkA_N (2.30E-40)	DNA-3-methyladenine glycosylase II/AraC family transcriptional regulator
	2 (54)	HhH-GPD (4.60E-32)	DNA-3-methyladenine glycosylase II/AraC family transcriptional regulator		

<sup>a</sup>First search.

<sup>b</sup>Second search.

<sup>c</sup>Pfam name.



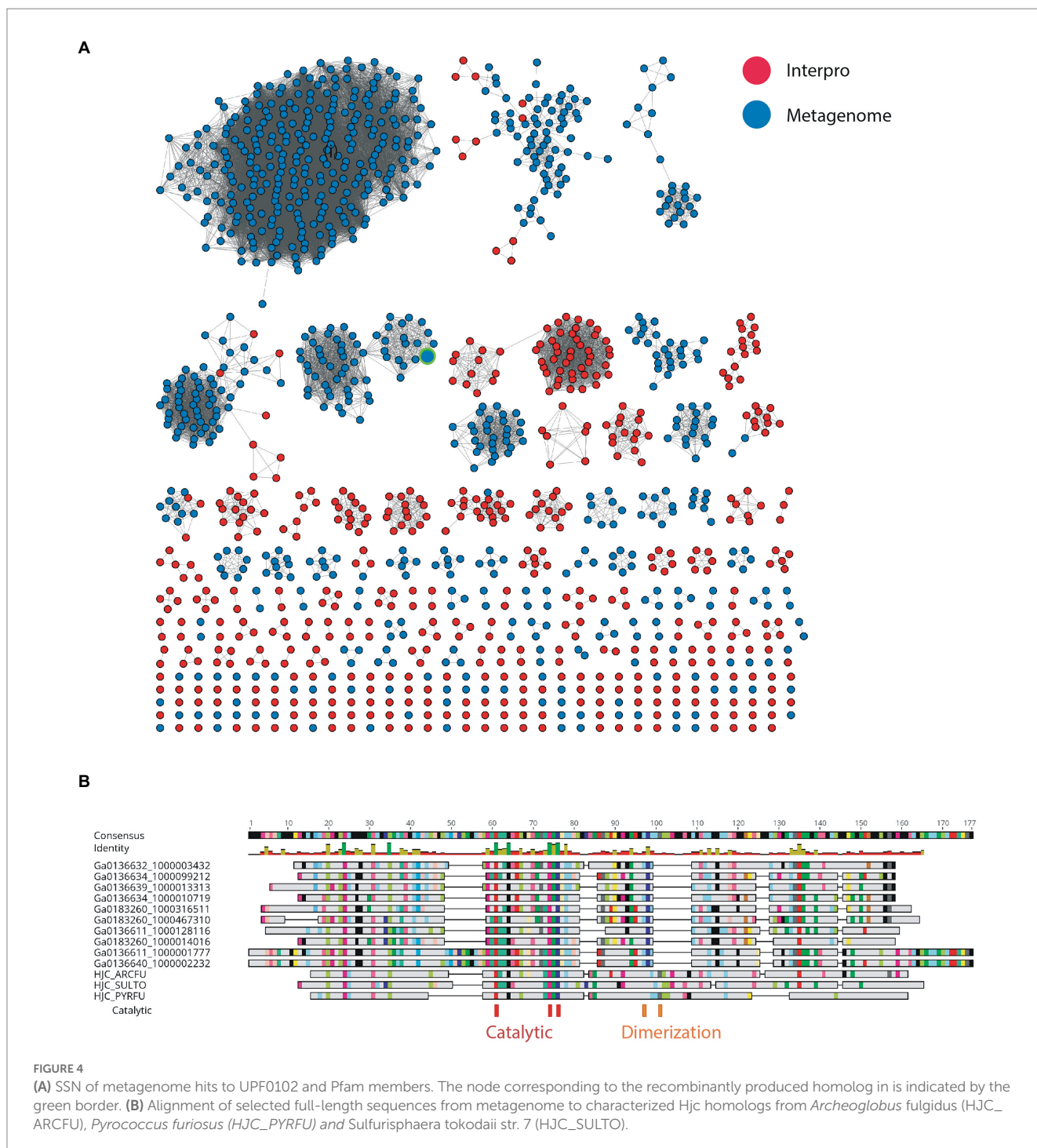
DV-metagenomic sequences and only two UniRef representatives, both of which are from Acidobacteria. Alignment of a representative selection of full-length metagenome sequences with two DdrA sequences from *D. radiodurans* showed that many of these DV-metagenome sequences were significantly longer than DdrA. Several Cluster #1 DV-metagenome sequences include an insertion of about 100 bp in the C-terminal region ([Figure 5B](#)), while the Cluster #2 sequences have a large N-terminal extension which has no counterpart in DdrA ([Figure 5C](#)).

DdrB, together with DdrA is induced in the genome of *D. radiodurans* during radiation damage ([Norais et al., 2009](#)). It is a single-strand binding protein implicated in mediating double-strand break repair *via* a RecA-independent single-strand annealing mechanism ([Norais et al., 2009](#)). Of the 271 sequences retrieved by searches with the DdrB HMM, the majority of clusters at a 26% edge threshold [48 sequences in clusters (3), (4), (6), and (7)] were identified as having non-DNA repair functions by hmmscan ([Table 3](#)). The 17 sequences with highest probability as DdrB homologs group

with sequences from UniRef (Clusters #1 and #2), while the nine sequences of the DV-metagenome-only Cluster #5 have much lower probability hits to DdrB (*E*-value of 34; [Supplementary 14A](#)). All nine sequences from Cluster #5 are essentially identical, and pairwise alignment of the most complete representatives with DdrB from *D. radiodurans* finds few fully-conserved positions ([Supplementary 14B](#)).

## Dry Valley metagenomes encode a plethora of DNA ligase proteins

DNA ligases are enzymes that join breaks in the phosphodiester backbone of DNA and have an essential role in DNA replication and repair in all organisms ([Williamson and Leiros, 2020](#)). The ATP-dependent sub-class of ligases have dedicated repair functions in bacteria and are often involved in alternative stationary-phase pathways in species that have spore-forming or dormant life-phases

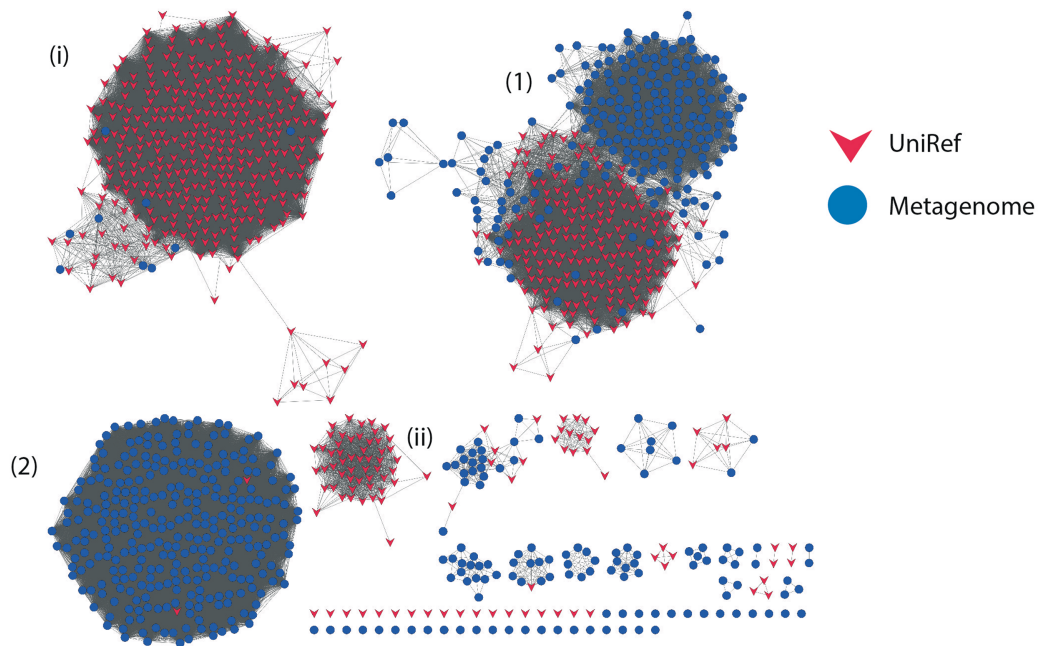


(Pergolizzi et al., 2016; Williamson et al., 2016). Because these enzymes are known to have a range of appending domains that impart specific functions, these sequences were sub-divided by searches for additional domains.

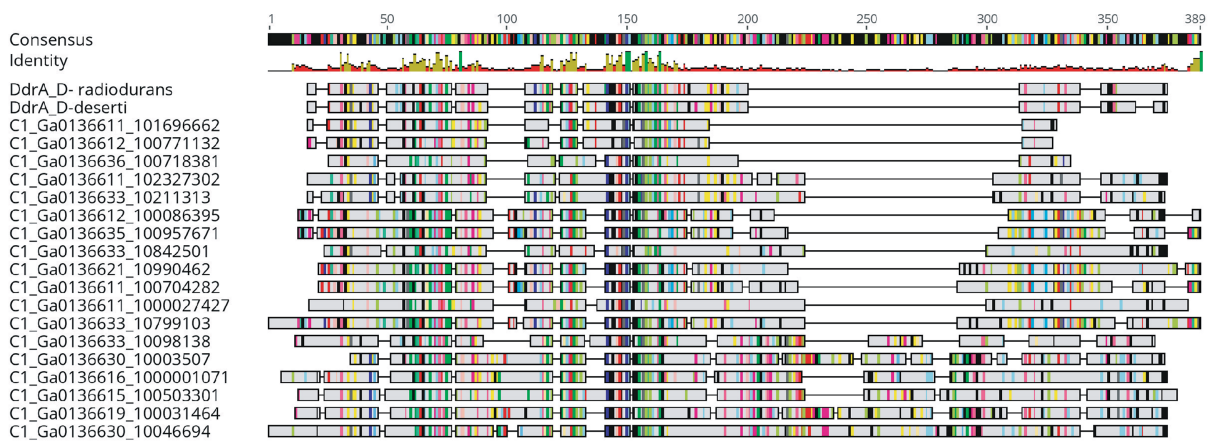
The LigB DNA ligases are the most widespread form of bacterial ATP-dependent DNA ligase, being found in species of *Mycobacterium* and *Pseudomonas* as well as cyanobacteria such as *Prochlorococcus marinus* (Gong et al., 2004; Williamson et al., 2016; Ejaz and Shuman, 2018; Williamson and Leiros, 2019). These ligases are typically found as part of an operon which also includes an Lhr helicase, a phosphodiesterase and a metallo-beta-lactamase-fold protein with

putative nuclease function (Ejaz and Shuman, 2018). The precise biological role and DNA-damage(s) targeted by this pathway are not currently known. Of the 1,791 DV-metagenome proteins which included the N-terminal DNA-binding domain of the LigB class DNA\_ligase\_A\_N (PF04675), most grouped together with a large number of UniRef sequences at the 50% threshold (Figure 6A). Of interest, Cluster #1 has fewer UniRef representatives than other clusters, and includes DV-metagenome sequences that are significantly longer than others in the network (Supplementary 15). Additional searches with hmmscan identified 13 sequences in this cluster that contain a fusion of the Zn-dependent metallo-hydrolase RNA

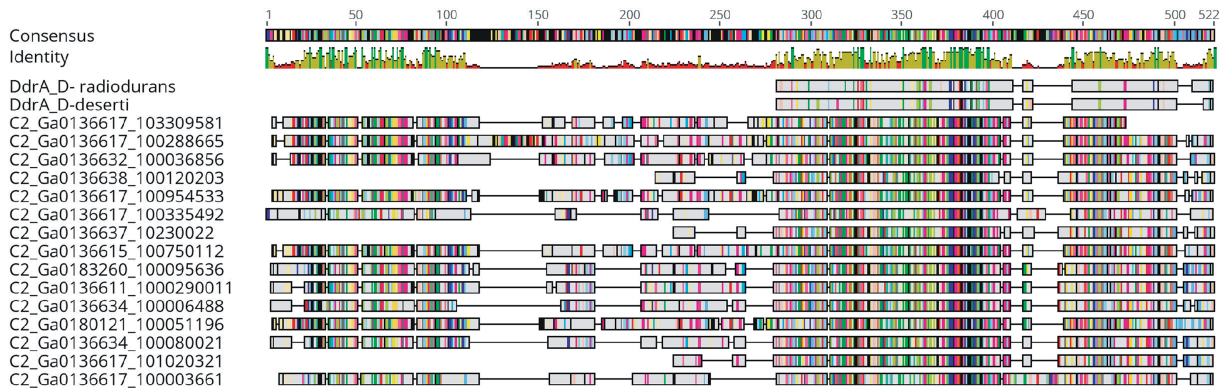
**A**



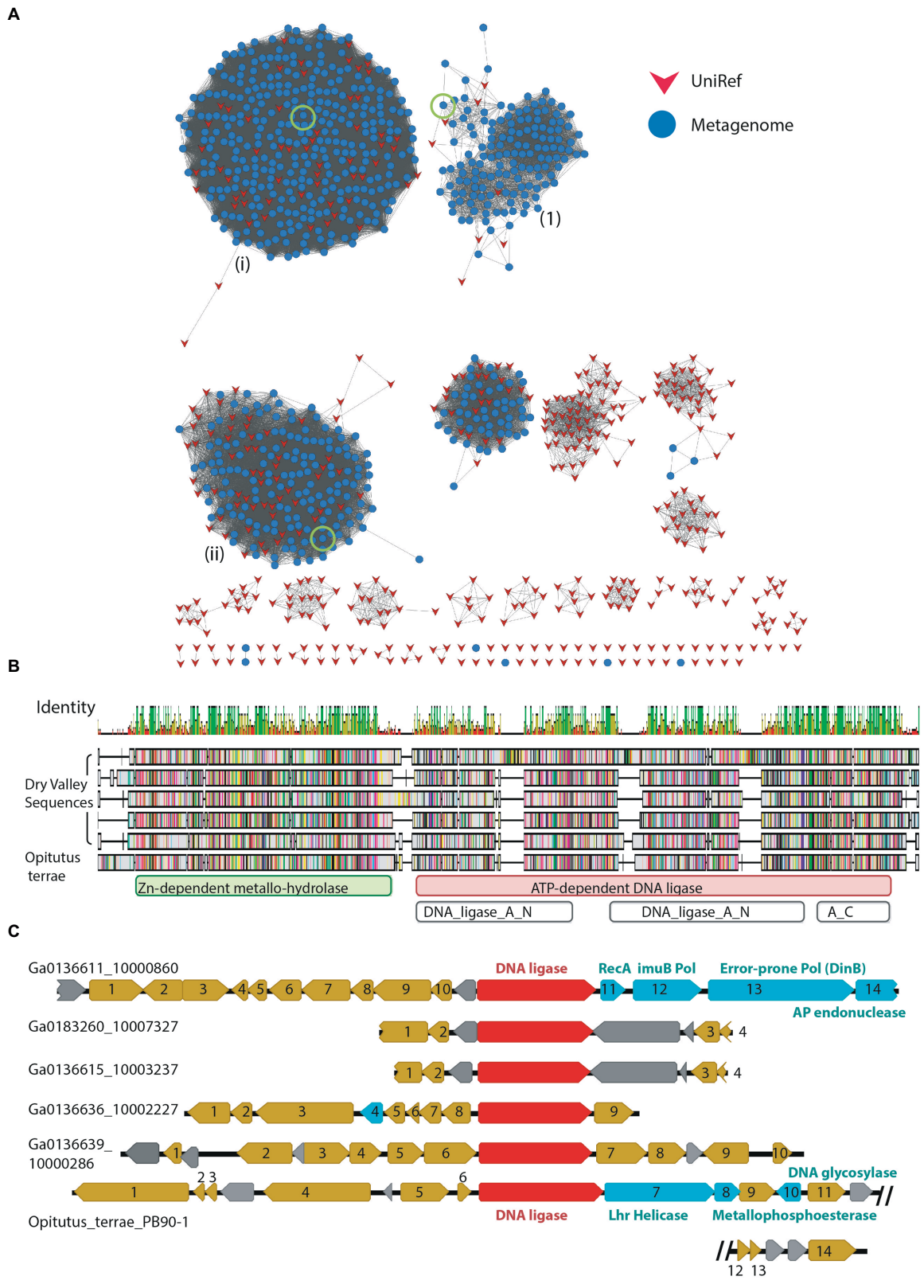
**B**



**C**



**FIGURE 5**  
SSN of metagenome hits to the Rad52/Rad22 family. **(A)** SSN with 26% identity edge threshold. Dry-Valley metagenome nodes are colored blue, UniRef50 nodes are indicated in red. Cluster numbers discussed in the text are given in parentheses. **(B)** Alignment of selected full-length Cluster #1 proteins with DdrA from *Deinococcus radiodurans* and *Deinococcus deserti*. **(C)** Alignment of selected full-length Cluster #2 proteins with DdrA from *D. radiodurans* and *D. deserti*.



**FIGURE 6**  
**(A)** SSN of metagenome hits to LigB-type DNA ligases at 50% identity edge threshold; other network parameters are detailed in Table 2. Domain compositions include the catalytic DNA<sub>ligase</sub>\_A\_M domain together with the N-terminal DNA binding domain DNA<sub>ligase</sub>\_A\_N. Dry-Valley metagenome nodes are colored blue, UniRef50 nodes are indicated in red. Cluster numbers discussed in the text are given in parentheses. The sequences used in further analysis on Clusters #1, ii and iii are indicated by a green circle. **(B)** Sequence alignment of full-length DV-metagenome sequences from Cluster #1 where an N-terminal metallo-nuclease (RMMBL) domain was detected with hmmscan Sequences are aligned to the *O. pituitus terrae* DNA ligase gene OTER\_RS15935 from the genome sequence NC\_010571\_Opitutus\_terrae\_PB90-1. Domain boundaries from Pfam/Interpro sequences searches are shown below the alignment. **(C)** Genomic context of metallo-nuclease-ligase fusion proteins from DV-metagenomes and *O. pituitus terrae*. The nuclease-ligase gene is shown in red, putative DNA-repair genes in cyan, other annotated genes in gold and hypothetical proteins in gray.

specificity domain RMMBL (PF07521) at the N-terminus of the DNA ligase DNA-binding domain (Figure 6B). Such fusions of the operon-associated exonuclease have been previously identified in the genome of the anaerobic bacterium *Opititus terrae* (Ejaz and Shuman, 2018), although the role of this module either as a ligase-fused domain or an autonomous enzyme has not been determined. To investigate whether these fused ligase-nuclease proteins are part of conserved operons, we compared the genomic context of a selection of these proteins where the sequence contig included flanking genes. None of the Dry Valley ligase-nuclease genes were found in the LigB/Lhr helicase/phosphodiesterase/metallo-beta-lactamase configuration seen in the *O. terrae* genome and there was little overall synteny between clusters (Figure 6C). Several genetic regions include genes involved in protein expression such as sigma factors, transcription factors, ribosomal subunit proteins and enzymes involved in tRNA or rRNA modification (Supplementary 16) and none were predicted to be complete or partial phage. Interestingly, contig Ga0136611\_10000860 two predicted error-prone DNA polymerases as well as a putative RecA-domain protein which could cooperate in DNA-repair processes.

LigD DNA ligases are involved in stationary-phase double-strand break repair as an alternative to homologous-recombination under conditions where a second partially-replicated chromosome may not be available (Pitcher et al., 2007b; Shuman and Glickman, 2007). At a bare minimum these modular enzymes have the ATP-dependent DNA ligase adenylation (DNA\_ligase\_A\_M, PF01068) and oligonucleotide binding (DNA\_ligase\_A\_C, PF04679) domains together with a PrimPol (PrimaseS, PF01896) domain, for example the LigD protein of *Bacillus subtilis* (Weller et al., 2002). Others additionally have a phosphodiesterase domain (LigD\_N, PF13298), for example LigD proteins of *Mycobacterium tuberculosis* and *Agrobacterium tumefaciens* (Zhu and Shuman, 2007; Wright et al., 2010). Within the DV-metagenomes LigD sequences were defined as those possessing the PrimaseS and/or the LigD\_N domain in addition to the conserved DNA\_ligase\_A\_M domain. A total of 1,805 such sequences were retrieved from the DV-metagenomes, all of which were annotated as LigD by IMG (Table 2). Of these, 650 were retained after applying length cutoffs and were used to build the SSN. DV-metagenome-only clusters that formed at a 50% threshold were analyzed using hmmscan which, in addition to the PrimaseS and DNA\_ligase\_A\_M domains, detected the oligo-nucleotide binding domain DNA\_ligase\_A\_C that forms part of the DNA ligase catalytic core (Table 3). Clusters comprise combinations of either the DNA\_ligase\_A\_M -DNA\_ligase\_A\_C core domains with PrimaseS [Supplementary 17, Cluster (1)], or these core domains with LigD\_N [Supplementary 17, Cluster (2)–(5)]. The absence of additional domains or sequence features, coupled with the known primary-sequence diversity inherent in the DNA-ligase family (Williamson et al., 2016) suggests these metagenome-only clusters represent partial sequences of LigD proteins, rather than unique Dry Valley variants. Full-length metagenome sequences which include both PrimaseS and LigD\_N modules cluster together with UniRef sequences.

ATP-dependent DNA ligases that lacked either the Lig-B or Lig-D type domains were processed as a single dataset. The majority of sequences still clustered together with UniRef sequences at the 54% threshold (Supplementary 18). In the largest cluster [Supplementary 18. Cluster (i)] both UniRef and metagenome sequences appeared to be the LigC type DNA ligases including only the core catalytic adenylation (DNA\_ligase\_A\_M, PF01068) and oligonucleotide

binding domain (DNA\_ligase\_A\_C, PF04679; Table 3). In other mixed clusters, UniRef ligases were the LigD or LigB-type, indicating that metagenome sequences in these clusters represented incomplete gene fragments for these enzymes. Hmmscan of the largest predominantly-metagenome clusters revealed that Clusters (1) to (3) were likewise catalytic core-only proteins (DNA\_ligase\_A\_M and DNA\_ligase\_A\_C) without any accessory domains, while the fourth cluster (Cluster (4)) had highest probability as a protein translocase subunit. This suggests that the majority of proteins in this dataset were LigC type proteins, some of which have considerable sequence divergence from highly-represented proteins in the UniRef database, while the remainder represent fragments of other previously-characterized DNA ligase classes.

## Dry Valley metagenomes include diverse excision-repair and direct damage-reversal proteins

The Pfam HhH-GPD comprises an expansive superfamily including various enzymes involved in DNA repair and mismatch correction such as the 8-oxoguanidine DNA glycosylases, methyl-CPG-binding proteins and methyladenine glycosylases (Bruner et al., 2000; Otani et al., 2013; Pidugu et al., 2021). The separation of distinct clusters in this network at a relatively low identity threshold (25%) likely reflects this diversity with the two largest clusters being 1,766 and 1,137 nodes and including a mixture of both metagenome and UniRef sequences [Supplementary 19 (i) and (ii)]. Separate clusters of mainly DV-metagenomic sequence composition were also identified, all of which had IMG annotations consistent with known DNA repair functions (Table 3). In an attempt to reduce the complexity of the HhH-GPD network, enzymes involved in alkylation damage repair and possessing the additional AlkA\_N domain as well as the HhH-GPD domain were analyzed in a separate network (Supplementary 20). Most DV-metagenomic sequences in this network fell into the large mixed cluster together with many bacterial UniRef representatives [Supplementary 20 (1)] while a smaller cluster containing most of the remaining DV-metagenomic sequences included only two UniRef nodes [Supplementary 20 (2)]. Both clusters were annotated as DNA-3-methyladenine glycosylases by IMG (Table 3). The clustering of unique DV-metagenome sequences separate from other representatives in the UniRef database suggests that these proteins may possess unique structural feature which may be explored further.

The majority of DV-metagenome photolyases grouped with UniRef sequences clusters (i) and (ii) including 2,303 and 430 DV-metagenome sequences, respectively, (Supplementary 21). Both clusters include a combination of eukaryotic and prokaryotic sequences which are predicted to be photolyases, and most DV-metagenome sequences in these clusters are annotated in IMG as photolyases. Hmmscan analysis of the three DV-metagenome-only clusters Cluster #1-#3 indicates that these have non-DNA repair functions as copper ion transporters, haloacid dehalogenases and regulatory NAD-dependent protein deacetylases (Table 3). The large number of photolyase genes annotated in the DV-metagenome sequences indicates that the direct reversal mechanism of removing UV-induced pyrimidine-pyrimidine lesions is important in these organisms which is consistent with the findings of previous studies

TABLE 4 Summary of proteins recombinantly expressed from DV-metagenome.

Name	Gene ID	Domains	Predicted molecular weight (kDa)	Expression strategy	Enzyme activity
DV-NucS	Ga0136640_100017415	NucS	58.8	MBP fusion, TEV-cleaved	Endonuclease (uracil and abasic), exonuclease single-stranded
DV-Hjc	Ga0136632_1000003432	UPF0102/Hjc	14.0	MBP fusion, TEV-cleaved	Not detected
DV-Lig-2	Ga0136636_1000055115	DNA_ligase_A_N, DNA_ligase_A_M, DNA_ligase_A_C	59.3	MBP fusion	DNA ligase
DV-Lig-5	Ga0136613_1000000468	DNA_ligase_A_N, DNA_ligase_A_M, DNA_ligase_A_C	55.0	MBP fusion, TEV-cleaved	DNA ligase
DV-1-1-Nuc-Lig	Ga0136611_1000086013	RMMBL, DNA_ligase_A_N, DNA_ligase_A_M, DNA_ligase_A_C	105.6		Not tested
DV-1-1-Nuc	Ga0136611_1000086013 (nt 1–1,197; aa 1–399)	RMMBL	44.0	Insoluble	Not tested
DV-1-1-Lig	Ga0136611_1000086013 (nt 1,198–2,874; aa 400–959)	DNA_ligase_A_N, DNA_ligase_A_M, DNA_ligase_A_C	62.0	N-terminal His-tag	DNA ligase
DV-UvdE	Ga0136640_100700761	UvdE	34.1	Insoluble	Not tested
DV-DdrA	Ga0136611_1000290011	Rad52_Rad22	44.9	Insoluble	Not tested

which have implicated the high activity of photolyase proteins in extreme UV resistance of bacteria isolated from low-temperature environments (Albarracín et al., 2014; Marizcurrena et al., 2017).

## Recombinant expression and preliminary characterization of novel nuclease and ligase genes

To gain further insight into potential DNA repair activities of proteins identified by SSN analysis, we attempted recombinant expression of metagenomic proteins with interesting sequence features. Proteins chosen for further investigation, together with the conditions for their production are summarized in Table 4. Briefly, heterologous protein expression using *E. coli* BL21pLysS was achieved for a putative Cluster (2) NucS (hereafter DV-NucS), an Hjc/UPF0102 homolog (hereafter DV-Hjc) and two LigB DNA ligases (hereafter DV-Lig-2 and DV-Lig-5). DV-metagenome proteins were expressed as fusions with the MBP solubilization tag which was successfully removed by digestion with TEV protease in all cases, with the exception of DV-Lig-2 (Supplementary 22, 23). Production of the full-length nuclease-LigB ligase protein was unsuccessful due to low expression levels and poor solubility (data not shown), therefore we attempted expression of the two predicted enzymatic modules, the RMMBL nuclease (hereafter DV-1-1-Nuc) and the LigB ATP-dependent DNA ligase (hereafter DV-1-1-Lig) separately (Supplementary 24A). The new construct for DV-1-1-Lig with the N-terminus at amino acid Asp 400 of the nuclease-ligase interdomain linker had high soluble expression in Origami (DE3) cells and was

readily purified *via* a non-cleavable N-terminal His-tag (Supplementary 24C). In contrast, the nuclease domain, while readily expressed, was insoluble as both His-tagged and MBP-tagged constructs (Supplementary 24B). Attempts at soluble production of the UvdE and Rad52\_Rad22 homologs from the DV-metagenome were unsuccessful, resulting in inclusion bodies (data not shown).

Nucleolytic activity of DV-NucS and DV-Hjc was tested against a range of double-stranded DNA substrates incorporating various DNA damages and mismatches as well as 5'/3' tailed single-stranded/double-stranded duplexes and junctions (listed in tables Supplementary 6, 7, and shown schematically in Supplementary 4, 5). Nuclease activity was detected for DV-NucS with single-stranded DNA as well as site-specific cutting of duplexes containing an abasic site, matched uracil (opposite adenine) and mismatched uracil (opposite thymine) (Figures 7A,B). The previously-described mismatch cleaving activity of *Corynebacterium glutamicum* NucS/EndoMS required the presence of the beta-sliding clamp (DnaN subunit), and it is possible that similar additional factors may promote mismatch cleavage by DV-NucS (Ishino et al., 2018; Wozniak and Simmons, 2022). It is likewise possible that the uracil and abasic site cleavage activity of DV-NucS, which was observed after extended incubations, would be further increased by the inclusion of binding partners. Specific cleavage of uracil and hypoxanthine-containing sites, which are the product of deamination of cytosine and adenine, respectively, were recently described in NucS from the radiation-resistant archaeon *Thermococcus gammatolerans* (Zhang et al. 2020a,b) and DV-NucS may function as part of a similar deamination-removal pathways. The apparent single-stranded degrading activity of DV-NucS is surprising and this observation should be further



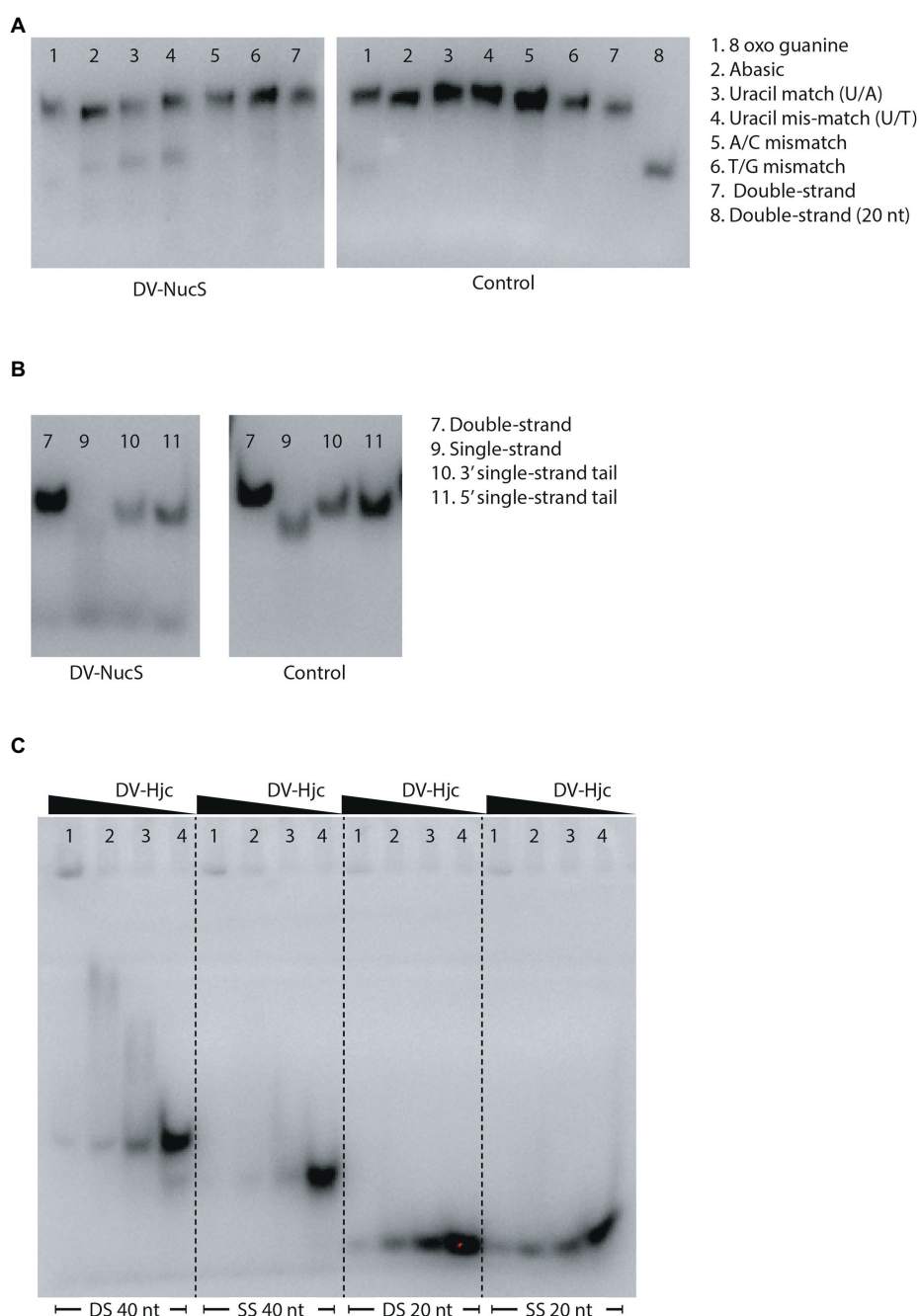


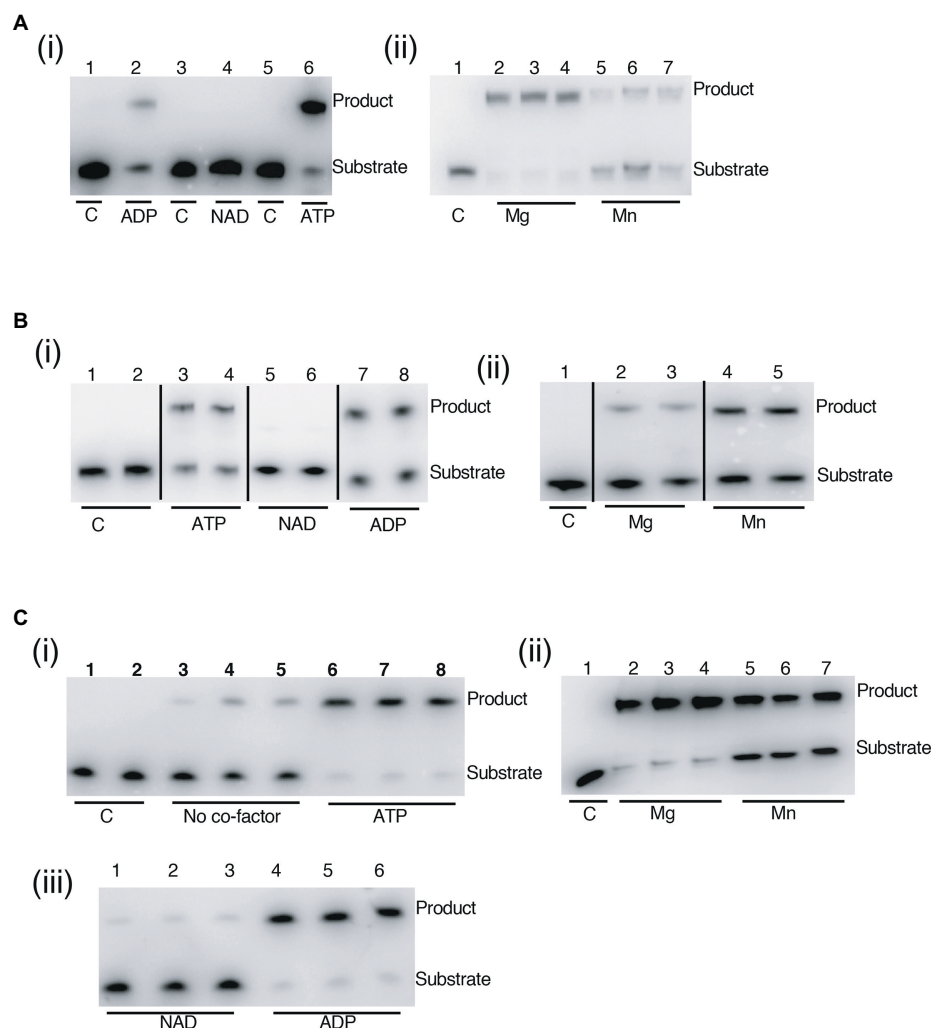
FIGURE 7

(A) Nucleolytic activity of DV-NucS on double-stranded DNA substrates containing damaged or mis-matched bases at a central position as shown in [Supplementary 4](#). Substrates were incubated for 18h with 344nM DV-NucS (DV-NucS gel) or an equivalent volume of buffer (control gel) at 25°C prior to analysis by denaturing urea PAGE. (B) Nucleolytic activity of DV-NucS on DNA substrates with double and single-stranded sections as shown in [Supplementary 5](#). Substrates were incubated for 18h with 344nM DV-NucS (DV-NucS gel) or an equivalent volume of buffer (control gel) at 25°C prior to analysis on 10% native TBE gels. (C) DNA binding activity of DV-Hjc to 40bp double-stranded DNA (DS 40nt), 40 nucleotide single-stranded DNA (SS 40nt), 20bp DNA (DS 20nt) and 20 nucleotide single-stranded DNA (SS 20nt). DV-Hjc concentrations were 3.3mM (lane 1), 657nM (lane 2) and 329nM (lane 3) with and storage buffer as a control (lane 4). DV-Hjc dilutions were pre-incubated with substrates at 15°C for 30min prior to electrophoresis on 10% native TBE gels with a 7% stacking layer.

investigated subsequent to confirming this activity *via* construction of active-site null mutants.

DV-Hjc exhibited no nucleolytic activity on any of the substrates under the conditions tested ([Supplementary 25A–C](#)). Thermal denaturation measured by DSF shows that purified DV-Hjc unfolds with a two-state transition with a  $T_m$  between 41 to 50°C depending

on pH, indicating that this protein possesses a well-folded hydrophobic core and is relatively stable under assay conditions making mis-folding during the production process an unlikely explanation for the lack of activity ([Supplementary 25D](#)). Likewise, EMSA indicates DV-Hjc binds strongly to both single- and double-stranded DNA, suggesting it does function in some DNA-modifying process ([Figure 7C](#)). It is



**FIGURE 8**

DNA ligase activity of LigB-type enzymes on singly nicked substrates. **(A)** DV-Lig-2 activity. Numbers represent single samples in each lane in panel (i), while lanes 2–4 and 5–7 in panel (ii) are triplicate reactions with of Mg and Mn, respectively. **(B)** DV-Lig-5 activity. Lanes 3–4, 5–6 and 7–8 in panel (i) are duplicate reactions with ATP, NAD and ADP, respectively, and lanes 2–3 and 4–5 in panel (ii) are duplicates with Mg and Mn. **(C)** DV-1-1-Lig activity. Lanes 1–2, 3–5, and 6–8 in panel (i) are duplicates of the control or triplicates of the no-cofactor and ATP-controls, respectively. Lanes 2–4 and 5–7 in panel (ii) are triplicates with Mg, and Mn. Lanes 1–3 and 4–6 are triplicates with NAD and ADP respectively. Final enzyme concentrations were 100nM (DV-Lig-5 and DV-1-1-Lig) and 100nM (DV-Lig-2, estimated as a 50:50 ratio cleaved:uncleaved from the SDS-PAGE gel). Concentrations of Mg and Mn were 10mM and nucleotide concentrations were 1 $\mu$ M of ATP, ADP or NAD<sup>+</sup> as indicated. Nucleotide cofactor preference assays [panels (i)] were run with Mg; divalent cation preference assays [panels (ii)] were all run with ATP. All reactions were incubated at 20°C for 2h prior to analysis by denaturing Urea-PAGE gel electrophoresis; the nucleotide cofactor reaction was pre-incubated with unlabeled DNA as described in the methods. C indicates no-protein control in all figures.

possible that DV-Hjc requires assay conditions other than those tested here, for example different metals, cofactors or interaction partners. It is also possible that, despite its homology to resolvases, DV-Hjc does not function as a nuclease but plays some other function in its native host such as DNA protection, recruitment of other nucleic-acid modifying enzymes to DNA or transcription factor activity.

To investigate whether DV-NucS or DV-Hjc are part of operons or gene clusters that include potential interaction partners, we analyzed the synteny of four of the largest contigs to detect common adjacent genes. A putative lysophospholipase was directly upstream of all four DV-NucS genes, however there were no other notably conserved adjacent open reading frames (Supplementary 26A). All five DV-Hjc genes were flanked by a predicted RNaseHIII gene but

again there was relatively low conservation. In other surrounding genes (Supplementary 26B).

Assay of DV-Lig-2, DV-Lig-5 and the DV-1-1-Lig domain with nicked DNA showed that all three enzymes function as DNA ligases in the presence of ATP (Figure 8). Both DV-Lig-5 and DV-1-1-Lig also exhibit considerable activity with ADP as an adenylate donor while DV-Lig-2 has some ligation activity with ADP, although it is considerably lower than that observed with ATP. Similar dual usage of these cofactors has been observed for many ATP-dependent DNA ligases, typically with a preference for the nucleotide triphosphate form over the diphosphate (Jeon and Ishikawa, 2003; Williamson and Pedersen, 2014). Neither DV-Lig-2 nor DV-Lig-5 ligate DNA in the presence of NAD<sup>+</sup>, while the small amount of ligation observed from

DV-1-1-Lig. This activity with NAD was similar to the no-cofactor control background, indicating that some purified enzyme-adenylate remained despite the pre/incubation step with unlabeled DNA rather than indicating the use of NAD by this enzyme. A comparison of magnesium and manganese show both are used as cofactors by all three enzymes, however DV-Lig-2 and DV-1-1-Lig have considerably higher activity with magnesium while DV-Lig-5 has a preference for manganese.

## Conclusion

In summary, our sequence homology approach using SSN analysis identified a large number of known genes known to be associated with repair functions, for example from *D. radiodurans* and mycobacteria. In addition, we have identified a several new genes which appear to be unique to this environment, or have very few representatives in current databases. Activity assays on recombinantly-produced enzymes encoded by these genes have demonstrated their DNA-processing abilities, validating this approach as a method to identify novel DNA repair or replication enzymes in metagenomic datasets. As the volume of DNA sequence data continues to increase, so too do the opportunities to discover new and fascinating aspects of the microbial diversity that inhabit our planet, making approaches necessary that can evaluate the true functional extent of this diversity (Temperton and Giovannoni, 2012). The integrated application of biochemical and *in silico* approaches to unique samples, such as those in this study, has significant potential impacts ranging from experimental validation of hypothetical protein functions, through to the description of entirely new DNA repair mechanisms underpinning survival in these highly adapted communities. The present work serves to enhance our general understanding of microbial adaptation to DNA-damaging environments. Further, as many molecular biological protocols utilize DNA modifying enzymes, there is enormous potential for the development of biotechnological tools from newly discovered activities, which may be facilitated by this approach.

## Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary material.

## Author contributions

AW and SC conceived, designed, and coordinated the study. MM and SC selected sample sites for sequencing, purified, and sequenced

the metagenomic DNA. ER-S, RS, and AW recombinantly-produced and assayed enzymes. ER-S and AW drafted the manuscript. All authors contributed to the article and approved the submitted version.

## Funding

This work was supported by the The Marsden Fund of New Zealand (18-UOW-034) to AW, ER-S, and MM are supported by University of Waikato Doctoral Scholarships, RS was supported by a University of Waikato Summer Studentship. ER-S received a University of Waikato School of Science Student Trust Research Grant toward consumables. The metagenomic sequencing of the Dry Valley samples was supported through awards to SC from The Community Sequencing Program (The Joint Genome Institute, DOE, USA) and The New Zealand Antarctic Research Institute (NZARI). Logistics support for the field seasons was provided by Antarctica New Zealand.

## Acknowledgments

We gratefully acknowledge the many researchers who were involved in the New Zealand Terrestrial Antarctic Biocomplexity Survey (nzTABS) for their time and commitment with this work. We thank Roanna Richards-Babbage and Amanda Maguren for their technical assistance with sample purification.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmicb.2023.1156817/full#supplementary-material>

## References

- Albarracin, V. H., Simon, J., Pathak, G. P., Valle, L., Douki, T., Cadet, J., et al. (2014). First characterisation of a CPD-class I photolyase from a UV-resistant extremophile isolated from high-altitude Andean lakes. *Photochem. Photobiol. Sci.* 13, 739–751. doi: 10.1039/c3pp50399b
- Aravind, L., Walker, D. R., and Koonin, E. V. (1999). Conserved domains in DNA repair proteins and evolution of repair systems. *Nucleic Acids Res.* 27, 1223–1242. doi: 10.1093/nar/27.5.1223
- Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., et al. (2016). PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* 44, W16–W21. doi: 10.1093/nar/gkw387
- Blasius, M., Buob, R., Shevelev, I. V., and Hubscher, U. (2007). Enzymes involved in DNA ligation and end-healing in the radioresistant bacterium *Deinococcus radiodurans*. *BMC Mol. Biol.* 8:69. doi: 10.1186/1471-2199-8-69

- Blasius, M., Hübscher, U., and Sommer, S. (2008). *Deinococcus radiodurans*: what belongs to the survival kit? *Crit. Rev. Biochem. Mol. Biol.* 43, 221–238. doi: 10.1080/10409230802122274
- Bottos, E. M., Laughlin, D. C., Herbold, C. W., Lee, C. K., McDonald, I. R., and Cary, S. C. (2020). Abiotic factors influence patterns of bacterial diversity and community composition in the dry valleys of Antarctica. *FEMS Microbiol. Ecol.* 96. doi: 10.1093/femsec/iaa042
- Bowater, R., and Doherty, A. J. (2006). Making ends meet: repairing breaks in bacterial DNA by non-homologous end-joining. *PLoS Genet.* 2:e8. doi: 10.1371/journal.pgen.0020008
- Brissett, N. C., Zabradý, K., Płociński, P., Bianchi, J., Korycka-Machala, M., Brzostek, A., et al. (2020). Molecular basis for DNA repair synthesis on short gaps by mycobacterial primase-polymerase C. *Nat. Commun.* 11:4196. doi: 10.1038/s41467-020-18012-8
- Bruner, S. D., Norman, D. P., and Verdine, G. L. (2000). Structural basis for recognition and repair of the endogenous mutagen 8-oxoguanine in DNA. *Nature* 403, 859–866. doi: 10.1038/35002510
- Cary, S. C., McDonald, I. R., Barrett, J. E., and Cowan, D. A. (2010). On the rocks: the microbiology of Antarctic Dry Valley soils. *Nat. Rev. Microbiol.* 8, 129–138. doi: 10.1038/nrmicro2281
- Castañeda-García, A., Prieto, A. I., Rodríguez-Beltrán, J., Alonso, N., Cantillon, D., Costas, C., et al. (2017). A non-canonical mismatch repair pathway in prokaryotes. *Nat. Commun.* 8:14246. doi: 10.1038/ncomms14246
- Chown, S. L., Clarke, A., Fraser, C. I., Cary, S. C., Moon, K. L., and McGeoch, M. A. (2015). The changing form of Antarctic biodiversity. *Nature* 522, 431–438. doi: 10.1038/nature14505
- Christner, B. C., Priscu, J. C., Achberger, A. M., Barbante, C., Carter, S. P., Christianson, K., et al. (2014). A microbial ecosystem beneath the West Antarctic ice sheet. *Nature* 512, 310–313. doi: 10.1038/nature13667
- Creze, C., Lestini, R., Kühn, J., Ligabue, A., Becker, H. F., Czjzek, M., et al. (2011). Structure and function of a novel endonuclease acting on branched DNA substrates. *Biochem. Soc. Trans.* 39, 145–149. doi: 10.1042/BST0390145
- De Maayer, P., Anderson, D., Cary, C., and Cowan, D. A. (2014). Some like it cold: understanding the survival strategies of psychrophiles. *EMBO Rep.* 15, 508–517. doi: 10.1002/embr.201338170
- Earl, A. M., Rankin, S. K., Kim, K. P., Lamendola, O. N., and Battista, J. R. (2002). Genetic evidence that the *uvrE* gene product of *Deinococcus radiodurans* R1 is a UV damage endonuclease. *J. Bacteriol.* 184, 1003–1009. doi: 10.1128/jb.184.4.1003-1009.2002
- Ejaz, A., and Shuman, S. (2018). Characterization of Lhr-Core DNA helicase and manganese-dependent DNA nuclease components of a bacterial gene cluster encoding nucleic acid repair enzymes. *J. Biol. Chem.* 293, 17491–17504. doi: 10.1074/jbc.RA118.005296
- Ericsson, U. B., Hallberg, B. M., DeTitta, G. T., Dekker, N., and Nordlund, P. (2006). Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Anal. Biochem.* 357, 289–298. doi: 10.1016/j.ab.2006.07.027
- Gong, C. L., Martins, A., Bongiorno, P., Glickman, M., and Shuman, S. (2004). Biochemical and genetic analysis of the four DNA ligases of *Mycobacterium*. *J. Biol. Chem.* 279, 20594–20606. doi: 10.1074/jbc.M401841200
- Goosen, N., and Moolenaar, G. F. (2008). Repair of UV damage in bacteria. *DNA Repair (Amst)* 7, 353–379. doi: 10.1016/j.dnarep.2007.09.002
- Gourse, R. L., and Keck, J. L. (2007). Magic spots cast a spell on DNA primase. *Cells* 128, 823–824. doi: 10.1016/j.cell.2007.02.020
- Gutsche, I., Vujčić-Zagar, A., Siebert, X., Servant, P., Vannier, F., Castaing, B., et al. (2008). Complex oligomeric structure of a truncated form of DdrA: a protein required for the extreme radiotolerance of *Deinococcus*. *Biochim. Biophys. Acta* 1784, 1050–1058. doi: 10.1016/j.bbapap.2008.03.009
- Harris, D. R., Tanaka, M., Saveliev, S. V., Jolivet, E., Earl, A. M., Cox, M. M., et al. (2004). Preserving genome integrity: the DdrA protein of *Deinococcus radiodurans* R1. *PLoS Biol.* 2:e304. doi: 10.1371/journal.pbio.0020304
- Huang, Q., Li, Y., Zeng, C., Song, T., Yan, Z., Ni, J., et al. (2015). Genetic analysis of the Holliday junction resolvases Hje and Hjc in *Sulfolobus islandicus*. *Extremophiles* 19, 505–514. doi: 10.1007/s00792-015-0734-5
- Ishino, Y., and Narumi, I. (2015). DNA repair in hyperthermophilic and hyperthermoresistant microorganisms. *Curr. Opin. Microbiol.* 25, 103–112. doi: 10.1016/j.mib.2015.05.010
- Ishino, S., Nishi, Y., Oda, S., Uemori, T., Sagara, T., Takatsu, N., et al. (2016). Identification of a mismatch-specific endonuclease in hyperthermophilic archaea. *Nucleic Acids Res.* 44, 2977–2986. doi: 10.1093/nar/gkv153
- Ishino, S., Skouloubris, S., Kudo, H., C. l'Hermitte-SteadEs-Sadiq, A., Lambry, J. C., et al. (2018). Activation of the mismatch-specific endonuclease EndoMS/NucS by the replication clamp is required for high fidelity DNA replication. *Nucleic Acids Res.* 46, 6206–6217. doi: 10.1093/nar/gky460
- Jeon, S. J., and Ishikawa, K. (2003). A novel ADP-dependent DNA ligase from *Aeropyrum pernix* K1. *FEBS Lett.* 550, 69–73. doi: 10.1016/S0014-5793(03)00821-4
- Komori, K., Sakae, S., Daiyasu, H., Toh, H., Morikawa, K., Shinagawa, H., et al. (2000). Mutational analysis of the *Pyrococcus furiosus* Holliday junction resolvase *hjc* revealed functionally important residues for dimer formation, junction DNA binding, and cleavage activities. *J. Biol. Chem.* 275, 40385–40391. doi: 10.1074/jbc.M006294200
- Kowalczykowski, S. C. (2015). An overview of the molecular mechanisms of Recombinational DNA repair. *Cold Spring Harb. Perspect. Biol.* 7. doi: 10.1101/cshperspect.a016410
- Kreuzer, K. N. (2013). DNA damage responses in prokaryotes: regulating gene expression, modulating growth patterns, and manipulating replication forks. *Cold Spring Harb. Perspect. Biol.* 5:a012674. doi: 10.1101/cshperspect.a012674
- Kuzminov, A. (1999). Recombinational repair of DNA damage in *Escherichia coli* and bacteriophage lambda. *Microbiol. Mol. Biol. Rev.* 63, 751–813, table of contents. doi: 10.1128/MMBR.63.4.751-813.1999
- Le, P. T., Makhalyane, T. P., Guerrero, L. D., Vikram, S., Van de Peer, Y., and Cowan, D. A. (2016). Comparative metagenomic analysis reveals mechanisms for stress response in *Hypoliths* from extreme Hyperarid deserts. *Genome Biol. Evol.* 8, 2737–2747. doi: 10.1093/gbe/evw189
- Lee, K. C., Caruso, T., Archer, S. D. J., Gillman, L. N., Lau, M. C. Y., Cary, S. C., et al. (2018). Stochastic and deterministic effects of a moisture gradient on soil microbial communities in the McMurdo dry valleys of Antarctica. *Front. Microbiol.* 9:2619. doi: 10.3389/fmicb.2018.02619
- Lee, C. K., Laughlin, D. C., Bottos, E. M., Caruso, T., Joy, K., Barrett, J. E., et al. (2019). Biotic interactions are an unexpected yet critical control on the complexity of an abiotically driven polar ecosystem. *Commun. Biol.* 2:62. doi: 10.1038/s42003-018-0274-5
- Lestini, R., Duan, Z., and Allers, T. (2010). The archaeal Xpf/Mus81/FANCM homolog Hef and the Holliday junction resolvase Hjc define alternative pathways that are essential for cell viability in *Haloferax volcanii*. *DNA Repair (Amst)* 9, 994–1002. doi: 10.1016/j.dnarep.2010.06.012
- Lilley, D. M. J. (2017). Holliday junction-resolving enzymes-structures and mechanisms. *FEBS Lett.* 591, 1073–1082. doi: 10.1002/1873-3468.12529
- Lim, S., Jung, J. H., Blanchard, L., and de Groot, A. (2019). Conservation and diversity of radiation and oxidative stress resistance mechanisms in *Deinococcus* species. *FEMS Microbiol. Rev.* 43, 19–52. doi: 10.1093/femsre/fuy037
- Marizcurrena, J. J., Morel, M. A., Brana, V., Morales, D., Martínez-López, W., and Castro-Sowinski, S. (2017). Searching for novel photolyases in UVC-resistant Antarctic bacteria. *Extremophiles* 21, 409–418. doi: 10.1007/s00792-016-0914-y
- Markowitz, V. M., Chen, I. M., Palaniappan, K., Chu, K., Szeto, E., Pillay, M., et al. (2014). IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res.* 42, D560–D567. doi: 10.1093/nar/gkt963
- Matsuoka, K., Skoglund, A., Roth, G., de Pomereu, J., Griffiths, H., Headland, R., et al. (2021). Quantarctica, an integrated mapping environment for Antarctica, the Southern Ocean, and sub-Antarctic islands. *Environ. Model. Softw.* 140:105015. doi: 10.1016/j.envsoft.2021.105015
- Niederberger, T. D., Bottos, E. M., Sohm, J. A., Gunderson, T., Parker, A., Coyne, K. J., et al. (2019). Rapid microbial dynamics in response to an induced wetting event in Antarctic Dry Valley soils. *Front. Microbiol.* 10:621. doi: 10.3389/fmicb.2019.00621
- Nishino, T., Komori, K., Tsuchiya, D., Ishino, Y., and Morikawa, K. (2001). Crystal structure of the archaeal Holliday junction resolvase Hjc and implications for DNA recognition. *Structure* 9, 197–204. doi: 10.1016/S0969-2126(01)00576-7
- Norais, C. A., Chitteni-Pattu, S., Wood, E. A., Inman, R. B., and Cox, M. M. (2009). DdrB protein, an alternative *Deinococcus radiodurans* SSB induced by ionizing radiation. *J. Biol. Chem.* 284, 21402–21411. doi: 10.1074/jbc.M109.010454
- Nunn, B. L., Slattery, K. V., Cameron, K. A., Timmins-Schiffman, E., and Junge, K. (2014). Proteomics of *Colwellia psychrerythraea* at subzero temperatures—a life with limited movement, flexible membranes and vital DNA repair. *Environ. Microbiol.* doi: 10.1111/1462-2920.12691
- Otani, J., Arita, K., Kato, T., Kinoshita, M., Kimura, H., Suetake, I., et al. (2013). Structural basis of the versatile DNA recognition ability of the methyl-CpG binding domain of methyl-CpG binding domain protein 4. *J. Biol. Chem.* 288, 6351–6362. doi: 10.1074/jbc.M112.431098
- Pavlopoulou, A., Savva, G. D., Louka, M., Bagos, P. G., Vorgias, C. E., Michalopoulos, I., et al. (2016). Unraveling the mechanisms of extreme radioresistance in prokaryotes: lessons from nature. *Mutat. Res. Rev. Mutat. Res.* 767, 92–107. doi: 10.1016/j.mrrev.2015.10.001
- Pergolizzi, G., Wagner, G. K., and Bowater, R. P. (2016). Biochemical and structural characterisation of DNA ligases from bacteria and archaea. *Biosci. Rep.* 36:00391. doi: 10.1042/BSR20160003
- Perugini, G., Miggiano, R., Serpe, M., Vettone, A., Valenti, A., Lahiri, S., et al. (2015). Structure-function relationships governing activity and stability of a DNA alkylation damage repair thermostable protein. *Nucleic Acids Res.* 43, 8801–8816. doi: 10.1093/nar/gkv774
- Perugini, G., Valenti, A., D'Amato, A., Rossi, M., and Ciarrella, M. (2009). Reverse gyrase and genome stability in hyperthermophilic organisms. *Biochem. Soc. Trans.* 37, 69–73. doi: 10.1042/BST0370069
- Pidugu, L. S., Bright, H., Lin, W. J., Majumdar, C., Van Ostrand, R. P., David, S. S., et al. (2021). Structural insights into the mechanism of base excision by MBD4. *J. Mol. Biol.* 433:167097. doi: 10.1016/j.jmb.2021.167097

- Pitcher, R. S., Brissett, N. C., and Doherty, A. J. (2007b). Nonhomologous end-joining in bacteria: A microbial perspective. *Annu. Rev. Microbiol.* 61, 259–282. doi: 10.1146/annurev.micro.61.080706.093354
- Pitcher, R. S., Green, A. J., Brzostek, A., Korycka-Machala, M., Dziadek, J., and Doherty, A. J. (2007a). NHEJ protects *Mycobacteria* in stationary phase against the harmful effects of desiccation. *DNA Repair* 6, 1271–1276. doi: 10.1016/j.dnarep.2007.02.009
- Płociński, P., Brissett, N. C., Bianchi, J., Brzostek, A., Korycka-Machala, M., Dziembowski, A., et al. (2017). DNA ligase C and prim-PolC participate in base excision repair in mycobacteria. *Nat. Commun.* 8:1251. doi: 10.1038/s41467-017-01365-y
- Selvam, K., Duncan, J. R., Tanaka, M., and Battista, J. R. (2013). DdrA, DdrD, and PprA: components of UV and mitomycin C resistance in *Deinococcus radiodurans* R1. *PLoS One* 8:e69007. doi: 10.1371/journal.pone.0069007
- Sharma, J. K., Stevenson, L. J., Robins, K. J., Williamson, A. K., Patrick, W. M., and Ackerley, D. F. (2020). Methods for competitive enrichment and evaluation of superior DNA ligases. *Methods Enzymol.* 644, 209–225. doi: 10.1016/bs.mie.2020.04.061
- Shin, D. S., Pratt, A. J., and Tainer, J. A. (2014). Archaeal genome guardians give insights into eukaryotic DNA replication and damage response proteins. *Archaea* 2014:206735. doi: 10.1155/2014/206735
- Shuman, S., and Glickman, M. S. (2007). Bacterial DNA repair by non-homologous end joining. *Nat. Rev. Microbiol.* 5, 852–861. doi: 10.1038/nrmicro1768
- Stomeo, F., Makhallanyane, T. P., Valverde, A., Pointing, S. B., Stevens, M. I., Cary, C. S., et al. (2012). Abiotic factors influence microbial diversity in permanently cold soil horizons of a maritime-associated Antarctic Dry Valley. *FEMS Microbiol. Ecol.* 82, 326–340. doi: 10.1111/j.1574-6941.2012.01360.x
- Temperton, B., and Giovannoni, S. J. (2012). Metagenomics: microbial diversity through a scratched lens. *Curr. Opin. Microbiol.* 15, 605–612. doi: 10.1016/j.mib.2012.07.001
- Tiao, G., Lee, C. K., McDonald, I. R., Cowan, D. A., and Cary, S. C. (2012). Rapid microbial response to the presence of an ancient relic in the Antarctic dry valleys. *Nat. Commun.* 3:660. doi: 10.1038/ncomms1645
- Timmins, J., and Moe, E. (2016). A decade of biochemical and structural studies of the DNA repair machinery of *Deinococcus radiodurans*: major findings, functional and mechanistic insight and challenges. *Comput. Struct. Biotechnol. J.* 14, 168–176. doi: 10.1016/j.csbj.2016.04.001
- Tropea, J. E., Cherry, S., and Waugh, D. S. (2009). Expression and purification of soluble his(6)-tagged TEV protease. *Methods Mol. Biol.* 498, 297–307. doi: 10.1007/978-1-59745-196-3\_19
- Wang, B. B., Xu, J. Z., Zhang, F., Liu, S., Liu, J., and Zhang, W. G. (2022). Review of DNA repair enzymes in bacteria: with a major focus on AddAB and RecBCD. *DNA Repair (Amst)* 118:103389. doi: 10.1016/j.dnarep.2022.103389
- Wei, S. T., Lacap-Bugler, D. C., Lau, M. C., Caruso, T., Rao, S., de Los Rios, A., et al. (2016). Taxonomic and functional diversity of soil and Hypolithic microbial communities in Miers Valley, McMurdo dry valleys, Antarctica. *Front. Microbiol.* 7:1642. doi: 10.3389/fmicb.2016.01642
- Weller, G. R., Kysela, B., Roy, R., Tonkin, L. M., Scanlan, E., Della, M., et al. (2002). Identification of a DNA nonhomologous end-joining complex in bacteria. *Science* 297, 1686–1689. doi: 10.1126/science.1074584
- Wilkinson, A., Day, J., and Bowater, R. (2001). Bacterial DNA ligases. *Mol. Microbiol.* 40, 1241–1248. doi: 10.1046/j.1365-2958.2001.02479.x
- Williamson, A., Hjerde, E., and Kahlke, T. (2016). Analysis of the distribution and evolution of the ATP-dependent DNA ligases of bacteria delineates a distinct phylogenetic group 'Lig E'. *Mol. Microbiol.* 99, 274–290. doi: 10.1111/mmi.13229
- Williamson, A., and Leiros, H. S. (2019). Structural intermediates of a DNA-ligase complex illuminate the role of the catalytic metal ion and mechanism of phosphodiester bond formation. *Nucleic Acids Res.* 47, 7147–7162. doi: 10.1093/nar/gkz596
- Williamson, A., and Leiros, H. S. (2020). Structural insight into DNA joining: from conserved mechanisms to diverse scaffolds. *Nucleic Acids Res.* 48, 8225–8242. doi: 10.1093/nar/gkaa307
- Williamson, A., and Pedersen, H. (2014). Recombinant expression and purification of an ATP-dependent DNA ligase from *Aliivibrio salmonicida*. *Protein Expr. Purif.* 97, 29–36. doi: 10.1016/j.pep.2014.02.008
- Wozniak, K. J., and Simmons, L. A. (2022). Bacterial DNA excision repair pathways. *Nat. Rev. Microbiol.* 20, 465–477. doi: 10.1038/s41579-022-00694-0
- Wright, D., DeBeaux, A., Shi, R. H., Doherty, A. J., and Harrison, L. (2010). Characterization of the roles of the catalytic domains of *Mycobacterium tuberculosis* ligase D in Ku-dependent error-prone DNA end joining. *Mutagenesis* 25, 473–481. doi: 10.1093/mutage/geq029
- Wyatt, H. D., and West, S. C. (2014). Holliday junction resolvases. *Cold Spring Harb. Perspect. Biol.* 6:a023192. doi: 10.1101/cshperspect.a023192
- Yi, C., and He, C. (2013). DNA repair by reversal of DNA damage. *Cold Spring Harb. Perspect. Biol.* 5:a012575. doi: 10.1101/cshperspect.a012575
- Yung, C. C., Chan, Y., Lacap, D. C., Perez-Ortega, S., de Los Rios-Murillo, A., Lee, C. K., et al. (2014). Characterization of chasmoendolithic community in Miers Valley, McMurdo dry valleys, Antarctica. *Microb. Ecol.* 68, 351–359. doi: 10.1007/s00248-014-0412-7
- Zallot, R., Oberg, N., and Gerlt, J. A. (2019). The EFI web resource for genomic enzymology tools: leveraging protein, genome, and metagenome databases to discover novel enzymes and metabolic pathways. *Biochemistry* 58, 4169–4182. doi: 10.1021/acs.biochem.9b00735
- Zhang, L., Jiang, D., Wu, M., Yang, Z., and Oger, P. M. (2020a). New insights into DNA repair revealed by NucS endonucleases from Hyperthermophilic archaea. *Front. Microbiol.* 11:1263. doi: 10.3389/fmicb.2020.01263
- Zhang, L., Shi, H., Gan, Q., Wang, Y., Wu, M., Yang, Z., et al. (2020b). An alternative pathway for repair of deaminated bases in DNA triggered by archaeal NucS endonuclease. *DNA Repair (Amst)* 85:102734. doi: 10.1016/j.dnarep.2019.102734
- Zhu, H., and Shuman, S. (2007). Characterization of *agrobacterium tumefaciens* DNA ligases C and D. *Nucleic Acids Res.* 35, 3631–3645. doi: 10.1093/nar/gkm145