Probing sequence plasticity of Ung inhibitors via synthetic and structural biology

Citation: Muselmani, Wael (2022) Probing sequence plasticity of Ung inhibitors via synthetic and structural biology. [Thesis] (Unpublished)

Birkbeck, University of London

Doctoral Thesis

Probing sequence plasticity of Ung inhibitors via synthetic and structural biology

Author:

Wael Muselmani

Supervisor:

Dr. Renos Savva

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy
to the

University of London

Abstract

Department of Biological Sciences

Institute of Structural and Molecular Biology

Doctor of Philosophy

Probing sequence plasticity of Ung inhibitors via synthetic and structural biology

Wael Muselmani

The uracil-DNA glycosylases constitute a superfamily of DNA repair enzymes. Family 1 enzymes are referred to as Ung or UNG and this branch of the superfamily are also restriction factors, acting against cellular pathogens. UngIn is a global term of reference for anti-restriction proteins inhibiting Ung, which to date are known to be encoded by viruses and the SCC*mec* transposon pathogenicity islands of MRSA bacteria. UngIns belong to discrete protein fold classes (3 are currently known) that nevertheless share a universal mechanism of Ung inhibition. UngIn folds arise from unrelated sequence families, and within any family extreme sequence plasticity is characteristic. Consequently, the effectiveness of conventional sequence-based identification of UngIns is limited.

The study aims were to develop and assess methods for the identification of UngIns in genomes that have a biological need to encode them, but for which no UngIn sequence has yet been identified. We modelled known mutations in UngIns and, via library mutagenesis, generated an expanded repertoire of synthetic UngIns by utilising a novel bacterial conditional lethal assay developed in this study. We also determined newly identified UngIn structures by X-ray crystallography. The insights from these studies permitted us to develop a computational heuristic approach to scan genomes that should encode biologically essential UngIns not yet identified. This approach has enriched our search for incidences of biologically essential UngIns and suggests alternative hypotheses for Ung activity modulation mechanisms.

Acknowledgment

I would like to thank my primary supervisor, Dr. Renos Savva, for all his efforts throughout the journey of my PhD. Thank you for all the continuous support and the valuable advice that you provided to help me accomplish this thesis. Thank you for the motivation, for the career advice, and for all the supportive recommendation letters that you provided to support my applications for valuable courses and financial grants.

I would also like to thank my secondary supervisor, Dr. Mark Williams, for all his insightful guidance and supportive tutorials of sequence and structural bioinformatics tools.

A major thanks for Prof. Nick Keep, my thesis chair, for his generous time in reviewing my structure refinement process. Thank you for your detailed advice and for your contribution to the improvement of the crystal structures that I have deposited to the PDB.

I would like to thank Dr. Claire Bagnéris, the Rosalind Franklin Lab manager, for all the laboratory work support, thanks for helping in crystal mounting, X-ray data collection, and data processing. Thanks to Dr. Nikos Pinotsis for help with X-ray data collection. I am grateful for Dr. Altin Sula for the informative sessions on data processing and structure refinement. A big thank you for the Savva lab, thanks for Rosalia Santangelo and Naail Kashif-Khan for their contribution in the work performed in this thesis.

My biggest thank you is for my wife, Dania, who supported me during the whole journey, thanks for your patience and for lifting up a heavy family load over the last few months. Thanks for my sons: Abdulkarim, Omar and Jad who motivated me to complete this work as soon as I could to be able to enjoy family time with them.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

3meA	N3-methyladenine
3meG	N3-methylguanosine
5-FU	5-Fluorouracil
5-hC	5-hydroxycytosine
5-hmU	5-Hydroxymethyluracil
5-hU	5-hydroxyuracil
7meG	N7-methylguanosine
8-oxoG	8-oxoguanine
A	Adenine
AP	Apurinic/Apyrimidinic
BER	Base Excision Repair
C	Cytosine
ddNTPs	dideoxynucleotide triphosphates
DHU	5,6-Dihydrouracil
DNA	Deoxyribonucleic acid
ds	Double-stranded
DSBs	Double strand DNA breaks
dTMP	Deoxythymidine monophosphate
dUMP	Deoxyuridine monophosphate
dUTP	Deoxyuridine triphosphate
Fapy	Formamidopyrimidines
G	Guanine
Gh	5-guanidinohydantoin
HX	Hypoxanthine
MBD4	Methyl-CpG Binding Domain 4, DNA glycosylase
MjUDG	*Methanocaldococcus jannaschii* uracil-DNA glycosylase
MPG	N-methylpurine-DNA glycosylase
MUG	Mismatch specific uracil-DNA glycosylase

MUTYH	MutY DNA glycosylase
NEIL	Nei-Like DNA glycosylase
NER	Nucleotide Excision Repair
NTHL1	Nth-Like DNA glycosylase 1
OGG1	8-oxoguanine DNA glycosylase
PDB	Protein Data Bank
RNA	Ribonucleic acid
ROS	Reactive oxygen species
SMUG1	Single-strand-selective mono-functional uracil-DNA glycosylase
Sp	Spiroiminodihydantoin
ss	Single-stranded
T	Thymine
TDG	Thymine-DNA glycosylase
Tg	Thymine glycol
U	Uracil
UDG	Uracil-DNA glycosylase
UV	Ultraviolet
εA	1, N6-ethenoadenine
εC	3, N4-ethenocytosine

Chapter 1

1. Introduction

1.1. DNA Damage

DNA integrity and specific pairing of its nucleobases (Adenine with Thymine, and Guanine with Cytosine) are crucial for maintenance of genomic information. DNA, however, is vulnerable to different types of lesions introduced via diverse DNA-damaging agents. These could be endogenous (such as errors introduced by DNA polymerase) and/or exogenous (e.g., UV radiation, X-rays, viruses, and chemotherapy). Mechanisms of DNA damage include diverse reactions such as oxidation, hydrolysis, hydrolytic deamination, and nonenzymatic methylation[1], leading to formation of unusual base lesions[1,2].

1.1.1. Hydrolytic deamination

Deamination can occur in both single-stranded and double-stranded DNA. Four DNA bases are susceptible to hydrolytic deamination; two of these predominate: cytosine (to uracil), and 5-methylcytosine (to thymine). However, adenine (to hypoxanthine) and guanine (to xanthine) can also be subject to a degree of deamination. Between 100 and 500 cytosines per cell per day are deaminated to uracil, requiring a highly efficient repair pathway to avoid CG→TA transition mutation upon transcription or replication[2].

1.2. DNA repair pathways

Structural changes caused by DNA damage can block replication and/or transcription, or may lead to undesirable mutations that may cause cell death or carcinogenesis. Therefore, it is important that such changes are checked and corrected. This is accomplished by means of molecular mechanisms referred to as DNA repair pathways[3]. There are several DNA repair pathways including damage reversal, repair of strand breaks, nucleotide excision repair (NER), and base excision repair (BER).

1.2.1. Base Excision Repair (BER)

DNA base modifications including deamination are corrected by the base excision repair (BER) pathway. In base excision repair, the unusual or mutagenic base is eliminated initially by a DNA glycosylase. Some DNA glycosylases are monofunctional DNA glycosylases, while others are bifunctional DNA glycosylases and Apurinic/apyrimidinic (AP) lyases[4–6] (Table 1.2.1).

Monofunctional DNA glycosylases cut the N-glycosidic bond between the mutagenic base and the deoxyribose, while bifunctional DNA glycosylases/AP lyases also catalyse cleavage of the phosphodiester bond 3' from the AP-site (Figure 1.2.4a).

Table 1.2.1. Examples of monofunctional and bifunctional DNA glycosylases and their major substrates

Enzyme	Mono-/bifunctional	Main Substrates
Ung	M	U, 5-FU
SMUG	M	5-hmU, U:G > U:A > ssU, 5-FU, εC in ss and dsDNA
TDG	M	U:G > T:G
MBD4	M	U:G and T:G, 5-hmU in CpG context
MPG	M	3meA, 7meG, 3meG, Hx, εA
AlkA	M	3meA, 7meG
FPG	B	Me-FapyG, FapyA
OGG1	M/B	8-oxoG:C, Fapy:C
TAG	M	3meA
MUTYH	M	A opposite 8-oxoG/C/G
NTHL1	B	Tg, FapyG, 5-hC, 5-hU in dsDNA
NEIL1	B	Tg, FapyG, FapyA, 8-oxoG, 5-hU, DHU, Sp and Gh in ss and dsDNA
NEIL2	B	Similar to NEIL1
NEIL3	M/B	FapyG, FapyA, Sp and Gh in ssDNA

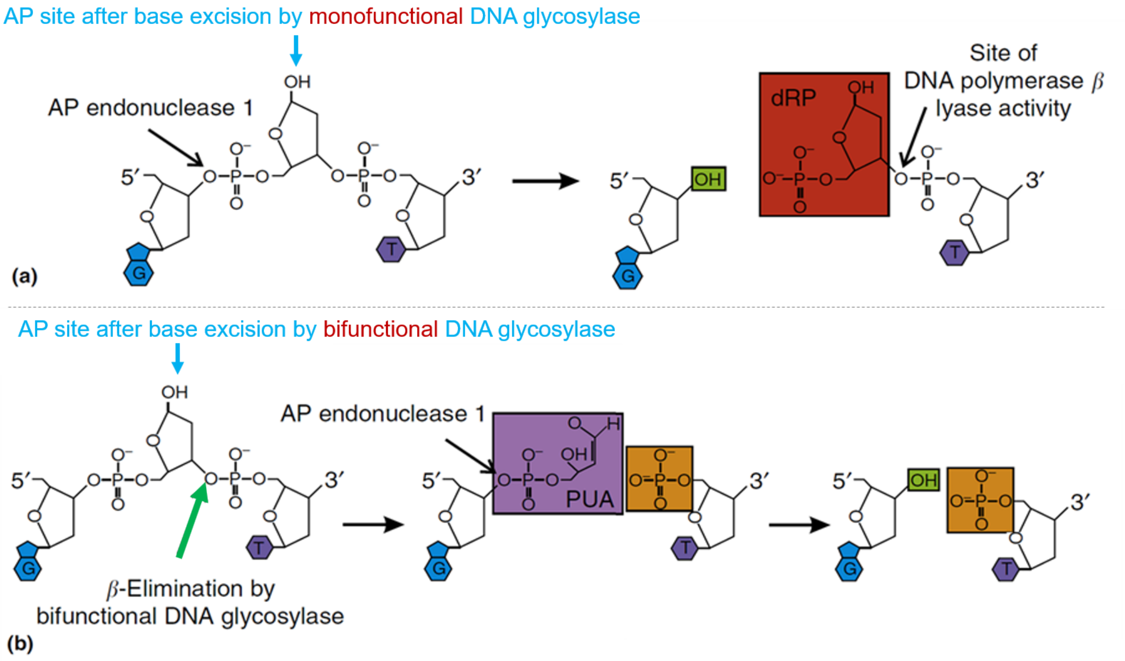

AP site after base excision by monofunctional DNA glycosylase
AP endonuclease 1
OH
5′
3′
G
T
(a)
OH
dRP
Site of
DNA polymerase β
lyase activity
AP site after base excision by bifunctional DNA glycosylase
AP endonuclease 1
PUA
β-Elimination by
bifunctional DNA glycosylase
(b)

Figure 1.2.1a. Different DNA ends generated by monofunctional and bifunctional DNA glycosylases. (a) Monofunctional DNA glycosylases only cut the N-glycosidic bond between the mutagenic base and the deoxyribose using water as a nucleophile, resulting in an apurinic/apyrimidinic (AP) site with an intact phosphodiester backbone in DNA. Apurinic/apyrimidinic endonuclease 1 (APE1) then acts to generate 3’-hydroxyl and 5’-deoxyribose phosphate ends (highlighted red square) to be excised, if no strand displacement synthesis is taking place (long-patch BER; Figure 1.2.1.b), by DNA polymerase β lyase activity generating a 5’-phosphate end. (b) Bifunctional DNA glycosylases/AP lyases use an amino group instead of water as a nucleophile to form a Schiff’s base intermediate. The covalent intermediate goes through cleavage of the phosphodiester bond 3’ from the AP-site by an enzyme-catalysed β-elimination step (green arrow), leaving a 3’-phospho-α,β-unsaturated aldehyde (PUA, highlighted purple square) and a 5’-phosphate end (highlighted brownish-yellow square). APE1 then acts to generate 3’-hydroxyl and 5’-phosphate ends (Modified from Parsons and Edmonds, 2016).[7]

AP-sites must be further treated, otherwise, they can lead to mutation during semiconservative replication. Further processing may take place by “short-patch” or “long-patch” BER[8], with slight differences in the set of enzymes used in prokaryotes and eukaryotes (Figure 1.2.1b). The factors that determine whether short-patch or long-patch BER takes place are still poorly understood. Several hypotheses suggested the switch between the two patches depending on ATP concentrations near the AP site (short-patch BER is preferred with higher concentrations of ATP) or the polymerase lyase activity (long-patch is likely to take place with low lyase activity)[9].

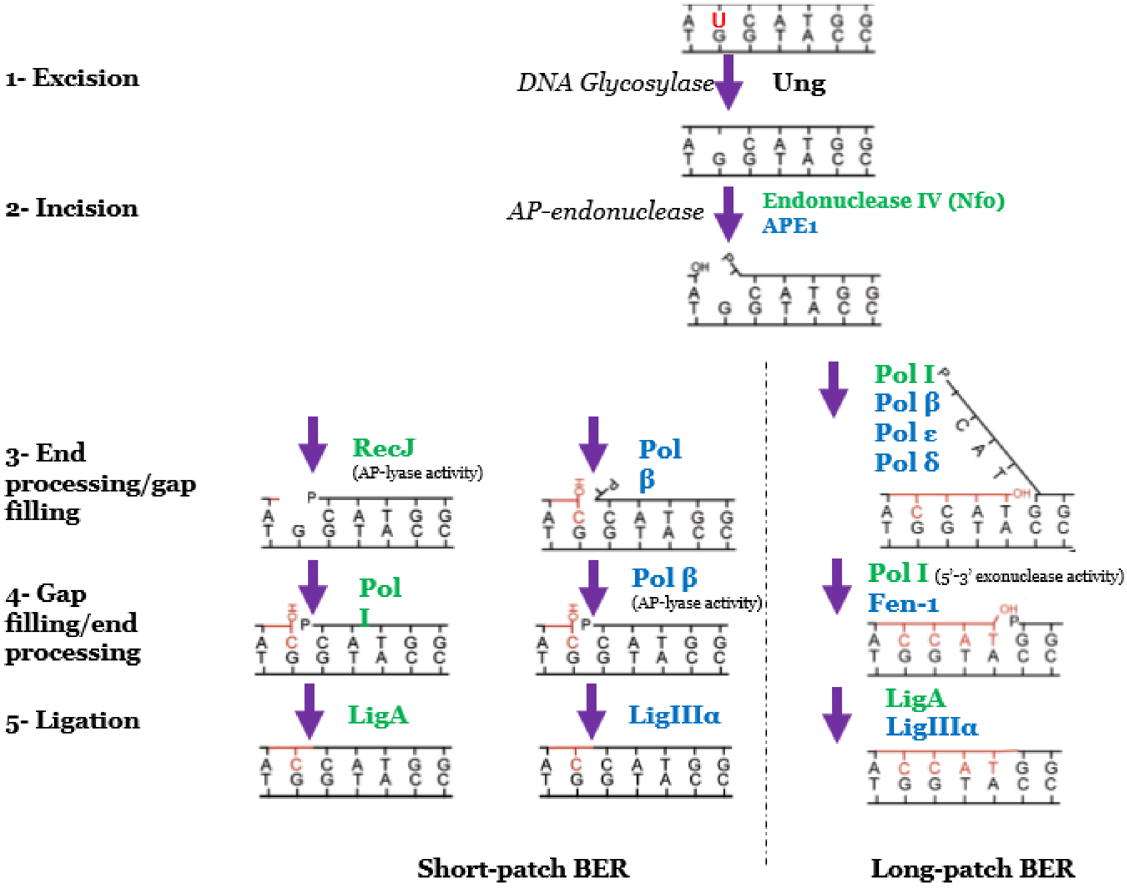


Figure 1.2.1b. BER short-patch and long-patch pathways. The five core steps of BER are: (1) Excision, (2) incision, (3) end processing, (4) gap filling, and (5) ligation. Prokaryotes and eukaryotes use similar BER mechanisms utilising slightly different sets of enzymes (green font is used for prokaryotic enzymes while blue font is used for eukaryotic ones). Ung enzymatically removes uracil from DNA leaving an apurinic/apyrimidinic (AP) site, following which, the action of AP-endonuclease generates a single-strand break 5' to the resulting AP-site. BER is completed by either of two sub-pathways: short-patch or long-patch. The short-patch pathway, wherein only 1 nucleotide is replaced, utilises an AP-lyase, a DNA polymerase, and a DNA ligase activity for the completion of the repair process. In the long-patch pathway, wherein 2-13 nucleotides are replaced, a flap structure is formed and then cleaved before sealing the nick and completing the repair process.

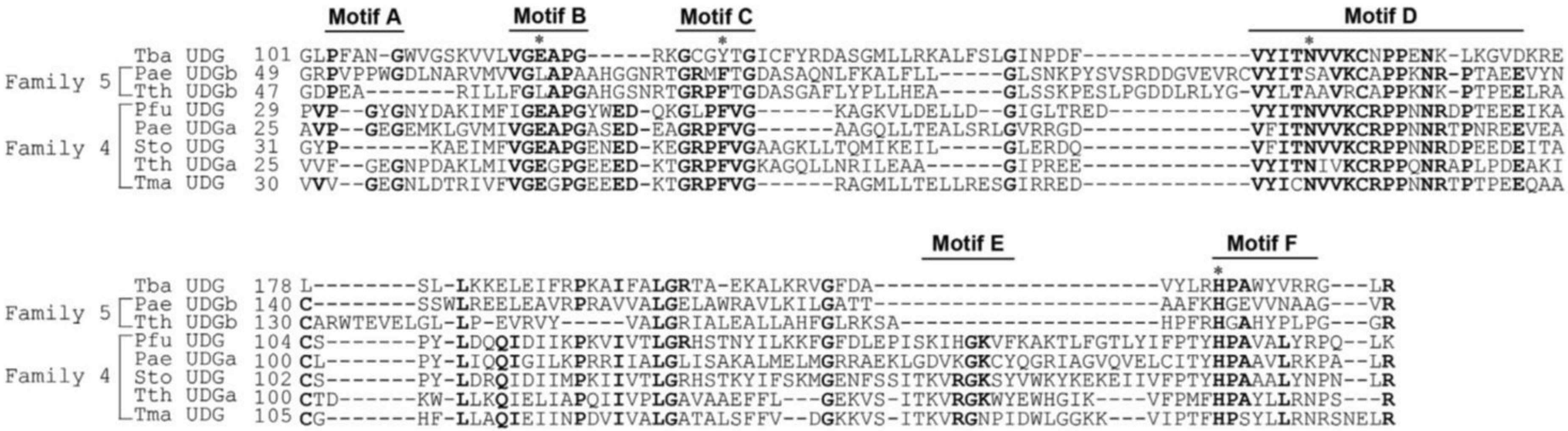
1.3. Uracil-DNA glycosylases (UDGs)

Uracil-DNA glycosylases (UDGs) are generally monofunctional DNA glycosylases. They excise uracil bases from DNA. Uracil bases in DNA may result from spontaneous deamination of cytosine bases or insertion of deoxyuridine instead of thymidine during DNA synthesis. UDG enzymes have been identified in archaea (e.g., *Methanococcus jannaschii*, *Pyrococcus* *furiosis*), viruses of prokaryotes (e.g., crAss-like phages), viruses of eukaryotes (e.g., herpesviruses and poxviruses), and all known eubacteria and eukaryotes. UDGs can be classified into seven families (although note the dichotomy in the definition of what constitutes family VI, as presented below) according to substrate specificity and substrate recognition mechanism. Family I UDGs (also called Ungs) includes *E. coli* Ung, which was the first identified DNA repair enzyme[10]. Ungs have high specificity for uracil (U), they also cleave 5-fluorouracil (5-fU) at a reduced rate[11]. Their major biological function is to remove the uracil produced by the spontaneous deamination of cytosine or via dUTP misincorporation. The Ung enzymes remove uracil (U) from both single-stranded DNA (ssDNA) and double-stranded DNA (dsDNA). Excision preference is uracil in ssDNA, uracil mispaired with guanine, then uracil mispaired with adenine[8].

Family II UDGs include *E. coli* mispair-specific uracil glycosylase (MUG) and human thymine-DNA glycosylase (TDG). Family II UDGs act as mismatch-specific enzymes on dsDNA. They are active against the U:G mispair but show low activity against the U:A base pair. TDG and high concentrations of MUG can excise thymine (T) from T:G mispairs[12]. Family III UDGs (sMUGs) act on both ssDNA and dsDNA substrates. Additional to eliminating U from U:G mispairs and U:A base pairs, the sMUG enzymes can also remove 5-hydroxymethyluracil (5-HmU)[13]. Family IV and family V UDG enzymes can remove U from uracil-mismatched dsDNA. Family IV but not family V can also excise U from ssDNA substrates. UDG families I through V, in spite of their <10% identity in amino acid sequence, share a common fold of the core domains and utilize a common pyrimidine binding motif at the N-terminus and glycosidic bond hydrolysis motif at the C-terminus[14]. There is a dichotomy in the definition of family VI UDG. Chung *et al.* (2003) suggested that a reported uracil-DNA glycosylase from *Methanococcus jannaschii* (MjUDG) could be classified into a novel sixth UDG family. MjUDG contains, in common with families IV and V, an iron-sulfur (4Fe-4S) cluster which seems to have only a structural role as it is distant from the DNA-binding surface[8]. In addition, MjUDG utilizes a helix-hairpin-helix (HhH) motif that is important for catalytic activity. MjUDG removes uracil form both dsDNA and ssDNA. Moreover, unlike any other UDGs, MjUDG can also excise 8-oxoguanine[15].

Lee *et al.* (2011) conducted a PSI-BLAST search for new uracil-DNA repair enzymes in archaea. The search led, via PSI-BLAST, to the identification of different genes in archaea, eubacteria, and eukaryotes with homology to uracil-DNA glycosylase superfamily enzymes. Experimental cloning, expression, purification and DNA glycosylase activity assay demonstrated that this was indeed a new class of DNA glycosylases, however lacking activity on uracil-DNA. Instead, these enzymes have shown hypoxanthine-DNA glycosylase activity. Based on unique biochemical properties, it was proposed that the new hypoxanthine-DNA glycosylases (HDG) be classified as family VI UDG[16]. UDG families I-VI at the present time represent examples of monofunctional DNA glycosylases only.

Recently, Zhang *et al*. (2021) has identified a novel bifunctional UDG encoded by *Thermococcus barophilus* Ch5. The sequence of that novel UDG is most closely related to sequences of family IV and family V UDG (Figure 1.3). However, unlike other members of family IV and family V UDG, this identified UDG is the first reported UDG with a glycosylase/AP-lyase activity, hence was suggested to be classified as family VII UDG[17,18]. A phylogenetic tree analysis of UDG families was performed in this thesis. A new suggested classification of UDG families is articulated in section 6.6.



```
                          Motif A       Motif B      Motif C                                           Motif D
                                          *             *                                            *
          Tba UDG   101 GLPFAN-GWVGSKVVLVGEAPG-----RKGCGYTGICFYRDASGMLLRKALFSLGINPDF-------------VYITNVVKCNPPENK-LKGVDKRE
Family 5 ┌Pae UDGb   49 GRPVPPWGDLNARVMVVGLAPAAHGGNRTGRMFTGDASAQNLFKALFLL-----GLSNKPYSVSRDDGVEVRCVYITSAVKCAPPKNR-PTAEEVYN
         └Tth UDGb   47 GDPEA-------RILLFGLAPGAHGSNRTGRPFTGDASGAFLYPLLHEA-----GLSSKPESLPGDDLRLYG-VYLTAAVRCAPPKNK-PTPEELRA
         ┌Pfu UDG    29 PVP--GYGNYDAKIMFIGEAPGYWED-QKGLPFVG------KAGKVLDELLD--GIGLTRED-----------VYITNVVKCRPPNNRDPTEEEIKA
         │Pae UDGa   25 AVP--GEGEMKLGVMIVGEAPGASED-EAGRPFVG------AAGQLLTEALSRLGVRRGD-------------VFITNVVKCRPPNNRTPNREEVEA
Family 4 │Sto UDG    31 GYP-------KAEIMFVGEAPGENED-KEGRPFVGAAGKLLTQMIKEIL-----GLERDQ-------------VFITNVVKCRPPNNRDPEEDEITA
         │Tth UDGa   25 VVF--GEGNPDAKLMIVGEGPGEEED-KTGRPFVGKAGQLLNRILEAA------GIPREE-------------VYITNIVKCRPPQNRAPLPDEAKI
         └Tma UDG    30 VVV--GEGNLDTRIVFVGEGPGEEED-KTGRPFVG------RAGMLLTELLRESGIRRED-------------VYICNVVKCRPPNNRTPTPEEQAA

                                                                         Motif E              Motif F
                                                                                              *
          Tba UDG   178 L--------SL-LKKELEIFRPKAIFALGRTA-EKALKRVGFDA----------------------VYLRHPAWYVRRG---LR
Family 5 ┌Pae UDGb  140 C--------SSWLREELEAVRPRAVVALGELAWRAVLKILGATT----------------------AAFKHGEVVNAAG---VR
         └Tth UDGb  130 CARWTEVELGL-LP-EVRVY-----VALGRIALEALLAHFGLRKSA--------------------HPFRHGAHYPLPG---GR
         ┌Pfu UDG   104 CS-------PY-LDQQIDIIKPKVIVTLGRHSTNYILKKFGFDLEPISKIHGKVFKAKTLFGTLYIFPTYHPAVALYRPQ--LK
         │Pae UDGa  100 CL-------PY-LIQQIGILKPRRIIALGLISAKALMELMGRRAEKLGDVKGKCYQGRIAGVQVELCITYHPAAVLRKPA--LR
Family 4 │Sto UDG   102 CS-------PY-LDRQIDIIMPKIIVTLGRHSTKYIFSKMGENFSSITKVRGKSYVWKYKEKEIIVFPTYHPAAALYNPN--LR
         │Tth UDGa  100 CTD------KW-LLKQIELIAPQIIVPLGAVAAEFFL---GEKVS-ITKVRGKWYEWHGIK----VFPMFHPAYLLRNPS---R
         └Tma UDG   105 CG-------HF-LLAQIEIINPDVIVALGATALSFFV--DGKKVS-ITKVRGNPIDWLGGKK---VIPTFHPSYLLRNRSNELR
```

Pae: *Pyrobaculum aerophilum* (UDGb: AAL63408; UDGa: AAL62921); **Tth**: *Thermus thermophilus* HB8 (UDGb: YP_144415; UDGa: BAC79245); **Pfu**: *Pyrococcus furiosus* (WP_011012532.1); **Sto**: *Sulfolobus tokodaii* (PDB: 4ZBY); **Tma**: *Thermotoga maritima* (PDB: 1L9G_A).

Figure 1.3. Family VII UDG sequence alignment with other UDG variants. Tba UDG, a novel family VII UDG sequence (accession: WP_056934618), is one of two UDG enzymes encoded by *Thermococcus barophilus* Ch5. Sequence alignment of Tba UDG with variant family IV and family V UDG sequences shows similarities to family IV UDG sequences (Motifs B, D, and F; and a gap preceding Motif D) and to family V UDG sequences (Motifs B, C, D, and F, and a lack of Motif E). Tba UDG is the 1st reported bifunctional UDG, hence it was suggested to be classified as family VII UDG (Modified from Shi *et al*., 2019).[17]

1.4. Family I Uracil-DNA Glycosylases (Ungs)

1.4.1. Ung Properties (folds, motifs, sequences)

Ungs are ubiquitous to all kingdoms of life except archaea, and they are the UDGs carried by viruses of eukaryotes. The conserved catalytic domain of these proteins comprises the C-terminal ~200 amino acid residues. The extra-catalytic N-terminal region of eukaryotic and viral Ungs is diverse. Some extensions of the N-terminus lead to different subcellular localization and protein-protein interactions[19]. For example, there are 2 mammalian isoforms of UNG, mitochondrial UNG1 and nuclear UNG2, that are generated by alternative splicing and transcription from different positions in the UNG gene[20]. UNG2 includes extra motifs in the N-terminal region such as a proliferating cell nuclear antigen (PCNA) binding motif.

Ung catalytic domain structures, motifs, and amino acid sequences are well conserved. The catalytic domain consists commonly of 4 β-strands and 4 α-helices[21] (Figure 1.4.1a). There are five conserved motifs: (1) the water-activating loop (the catalytic motif), (2) the proline-rich loop, which compresses the DNA backbone 5’ to the sampled base, (3) the uracil-binding motif, (4) the glycine-serine loop that compresses the DNA backbone 3’ to the sampled base, and (5) the DNA minor groove intercalation loop. These motifs are shown in Figure 1.4.1b, on a sequence alignment of *E. coli* Ung (eUng, PDB ID: 1UDG) and human UNG2 (hUNG, PDB ID: 1AKZ).

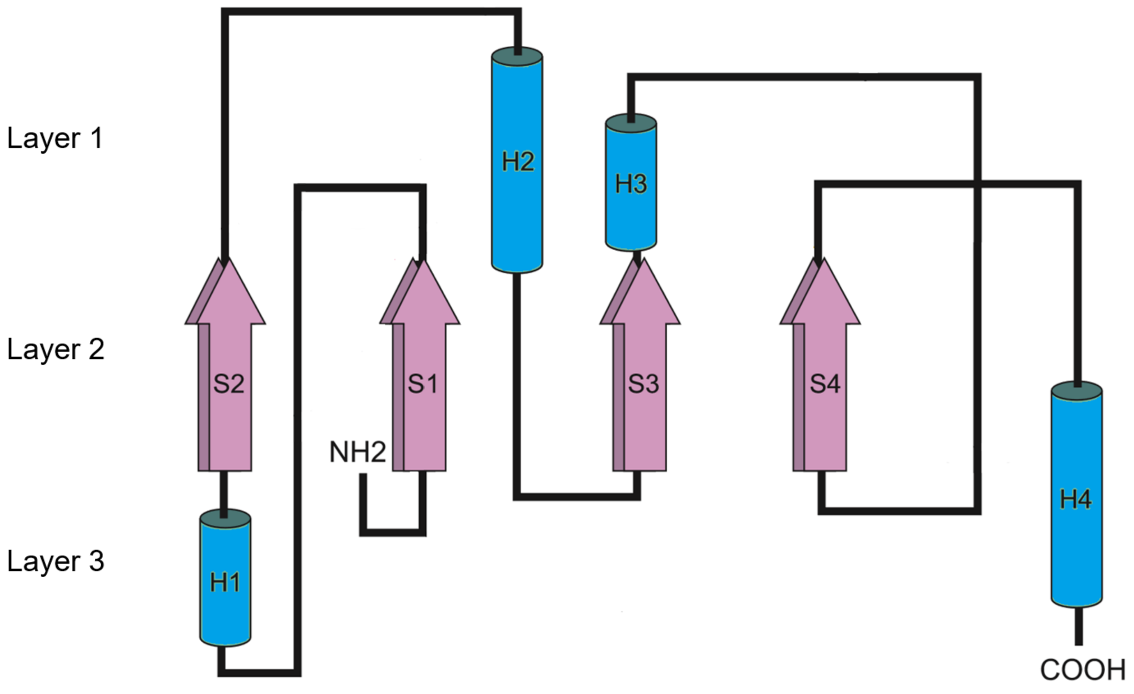
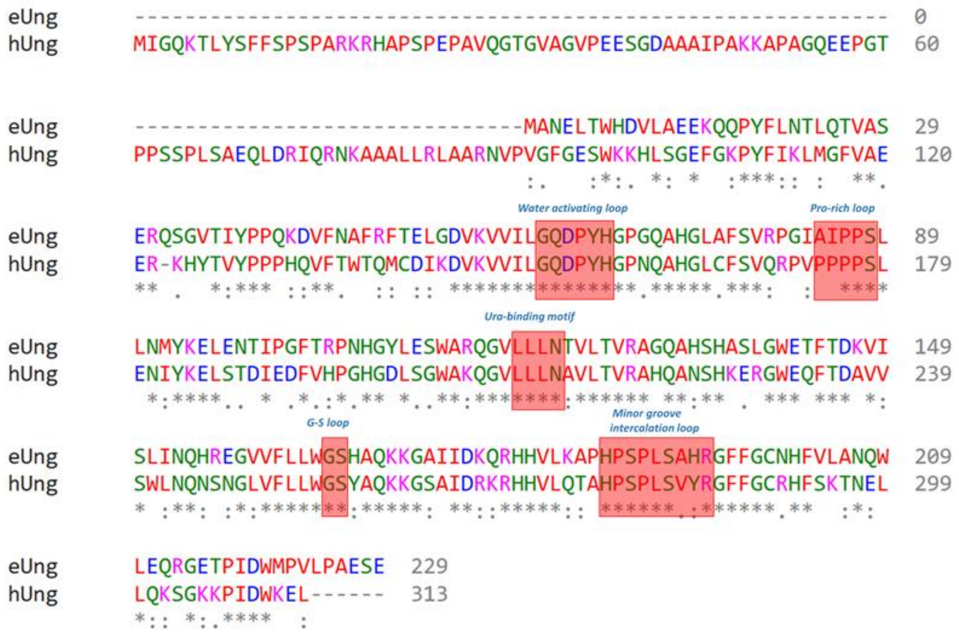


Figure 1.4.1a. Topology of Ung. The topology diagram highlights the conserved fold: three layers α/β/α, two α-helices (labelled H and coloured blue) on each side of the core of four parallel β-strands (labelled S and coloured pink) in the order 2134.



```
eUng      ------------------------------------------------------------  0
hUng      MIGQKTLYSFFSPSPARKRHAPSPEPAVQGTGVAGVPEESGDAAAIPAKKAPAGQEEPGT  60

eUng      -------------------------------MANELTWHDVLAEEKQQPYFLNTLQTVAS  29
hUng      PPSSPLSAEQLDRIQRNKAAALLRLAARNVPVGFGESWKKHLSGEFGKPYFIKLMGFVAE  120
                                         :.   :*:. *: *  :***:: :  **.

eUng      ERQSGVTIYPPQKDVFNAFRFTELGDVKVVILGQDPYHGPGQAHGLAFSVRPGIAIPPSL  89
hUng      ER-KHYTVYPPPHQVFTWTQMCDIKDVKVVILGQDPYHGPNQAHGLCFSVQRPVPPPPSL  179
          ** .  *:*** ::**.  :: :: ***************.*****.***:  :  ****

eUng      LNMYKELENTIPGFTRPNHGYLESWARQGVLLLNTVLTVRAGQAHSHASLGWETFTDKVI  149
hUng      ENIYKELSTDIEDFVHPGHGDLSGWAKQGVLLLNAVLTVRAHQANSHKERGWEQFTDAVV  239
           *:****.. * .*.:*.** *..**:*******.****** **:** . *** *** *:

eUng      SLINQHREGVVFLLWGSHAQKKGAIIDKQRHHVLKAPHPSPLSAHRGFFGCNHFVLANQW  209
hUng      SWLNQNSNGLVFLLWGSYAQKKGSAIDRKRHHVLQTAHPSPLSVYRGFFGCRHFSKTNEL  299
          * :**: :*:*******:*****: **::*****:: ******.:******.**  :*:

eUng      LEQRGETPIDWMPVLPAESE  229
hUng      LQKSGKKPIDWKEL------  313
          *:: *:.****  :
```

Figure 1.4.1b. Sequence alignment for *E. coli* Ung [1UDG] and Human UNG2 [1AKZ]. The conserved motifs are highlighted in red. Sequence alignment was performed with default settings at Clustal Omega web server: https://www.ebi.ac.uk/Tools/msa/clustalo/.

The symbols below each aligned residue represent conserved attributes, as follows:

*	Indicates positions which have a single, fully conserved residue.
:	Indicates conservation of strongly similar properties - scoring > 0.5 in the Gonnet PAM 250 matrix.
.	Indicates conservation of weakly similar properties - scoring =< 0.5 in the Gonnet PAM 250 matrix.

1.4.2. Ung Substrate specificity

Ungs form a uracil-binding pocket by co-localising amino acids from several of the conserved motifs (in *E. coli* Ung: Q63, D64, and Y66 of the catalytic water-activating Loop; F77 of the active site; S88 of the pro-rich loop; N123 of the uracil binding motif; and H187 of DNA minor groove intercalation loop; Figure 1.4.2a). Conserved aromatic residues, phenylalanine and tyrosine, near the active site stack against the DNA base. The size of uracil-binding pocket contributes to specificity by excluding larger purine bases.

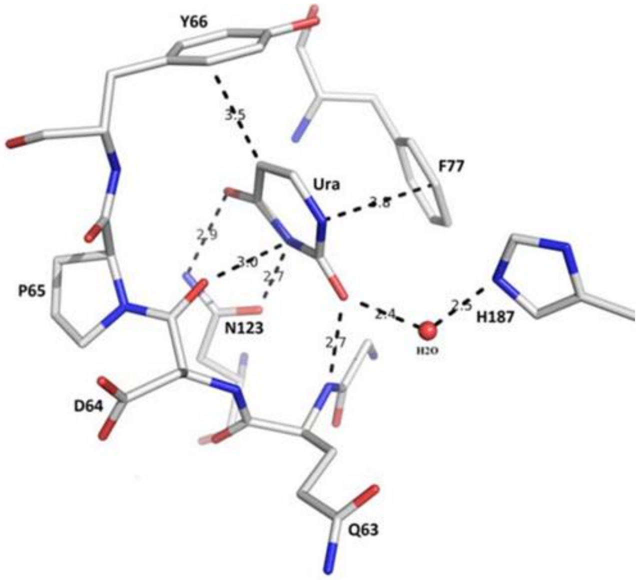


Figure 1.4.2a. Uracil-binding pocket in *E. coli* Ung. The active site in eUng [2EUG] highlighting the hydrogen bonds and van der Waals bonds (dashed lines, distances in Å) of uracil with active site residues. (Modified from Schormann *et al*., 2014)[8]

A conserved tyrosine residue (Y66 in *E. coli* Ung) achieves discrimination against thymine and other 5-substituted pyrimidines by presenting a steric barrier against the C5 of the pyrimidine nucleotide[8,22]. Additionally, thymine exclusion is assured by a second mechanism of trapping the thymine base at the mouth of the pocket via packing against the side chain ring of the conserved tyrosine, with the 5-methyl group in contacts with the top edge of the ring of the conserved phenylalanine[23]. However, in some synthetic mutants like UNG2 Y147A, the conserved tyrosine residue is altered to alanine which is smaller and lacks the aromatic ring, and such mutants therefore can excise thymine due to removal of the pre-catalytic thymine filter and removal of steric hindrance at the active site[24]. Ungs form specific hydrogen bonds with the O2, N3 and O4 of uracil via a conserved asparagine residue (UNG2 N204, *E. coli* Ung N123). The conformation of that asparagine residue is held by a triad of water molecules that make it specific for uracil recognition (Figure 1.4.2b). These specific bonds cannot occur with cytosine, which precludes its excision. Nevertheless, another synthetic mutant version of UNG2 (N204D) can also excise cytosine[25]. Aspartate in the mutant version does not form these bonds with the water molecules and it can freely rotate around its side chain, enabling it to retain activity against both uracil and cytosine[26] (Figure 1.2.4b).

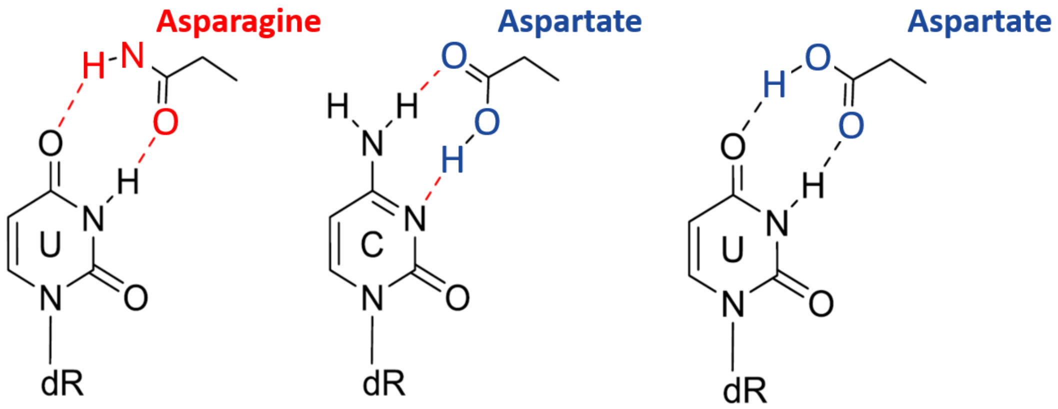


Figure 1.4.2b. Interaction of uracil with key asparagine residue [UNG2 N204, *E. coli* Ung N123; left] and proposed recognition of both cytosine (centre) and uracil (right) by N123D and N204D mutants of *E. Coli* Ung and UNG2, respectively.

1.4.3. Ung Base Recognition

Previous studies suggest that Ungs bind any DNA whether damaged or not. However, the affinity for uracil-DNA is 10–30 fold higher than for canonical DNA. Local separation of the double helix into two strands is reported to be a first step of Ung action, thus dsDNA is catalysed relatively more slowly than ssDNA[11,19,23].

Although DNA bases define substrate recognition, they are not major contributors to recognition of DNA by Ung, per se. However, interaction of conserved motif residues of Ung with some internucleotide phosphate groups is essential for DNA recognition (Figure 1.4.3). Ung demonstrated insignificant binding to a non-nucleotide polymer including uracil in an uncharged peptide chain instead of sugar-phosphate backbone, indicating the importance of DNA backbone interaction with Ung for significant binding[27]. The sugar conformation plays a role in the binding of Ung since the 3'-endo conformation of the sugar moiety (as found in A-form RNA) prevents interaction with Ung. Presence of an NH_2 group in the 2'-position of the sugar does not affect Ung binding but leads to inhibition of uracil excision activity due to steric hindrance caused by the NH2 group in the 2'-position. These results suggest an important role of C5 and C2' positions of deoxyuridine to remove any steric hindrance that might prevent the emergence of a functional enzyme-substrate complex[8,28].

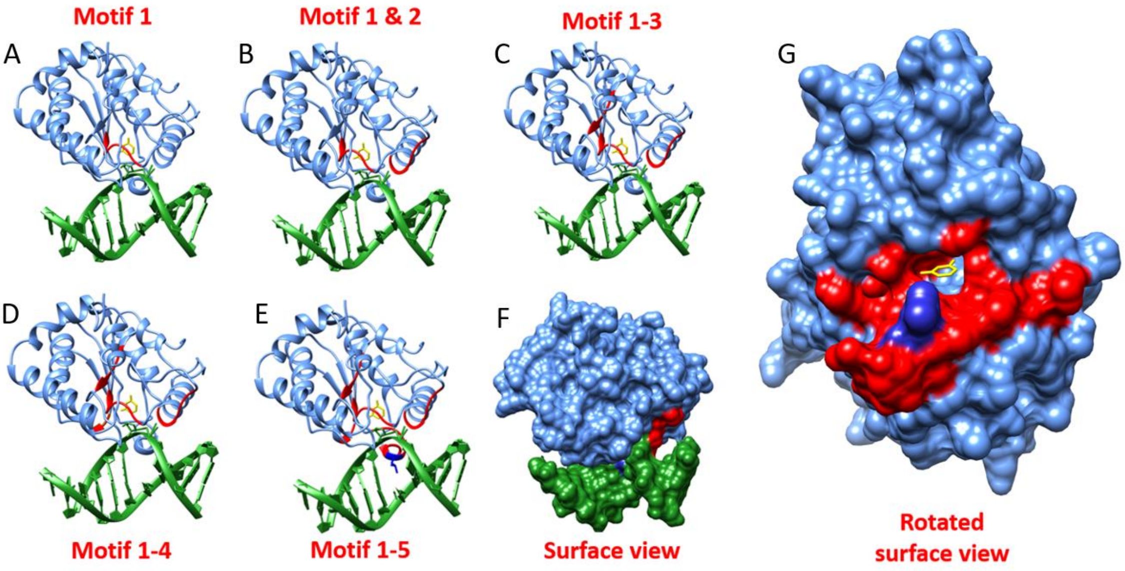


Figure 1.4.3. hUNG in complex with dsDNA (PDB ID: 1SSP). Ung is coloured cornflower blue with the 5 conserved motifs from sequence alignment coloured red and shown sequentially in figure panels A-E. dsDNA is coloured forest green. Uracil base is coloured yellow. Apical hydrophobic residue (leucine) in motif 5 (DNA minor groove intercalation loop) is coloured medium blue in figure panels E, F, and G. In the right cartoon (G), dsDNA is removed for clarity. Rotated surface view shows the DNA-binding cleft formed by the conserved motifs, this cleft binds specifically to phosphate-sugar backbone of DNA. This site therefore has no affinity for uracil containing polymers (such as uncharged peptide chain covalently attached to uracil). The uracil binding pocket can also be observed in the rotated surface view (uracil base is seen as a yellow stick). Panel graphics were rendered in Chimera[29].

1.4.4. Ung mode of action

Ung binds uracil-DNA and catalyses the cleavage of N-glycosidic bond between the uracil and the deoxyribose. This hydrolytic cleavage take place at the uracil-binding pocket of Ung. Upon binding DNA, a substrate-induced Ung conformational change from open to closed conformation takes place[30]. This conformation change involves mainly the Ung-DNA binding cleft forming residues with distance changes of more than 5 Å between some atoms in the 2 different conformations (Figure 1.4.4a).

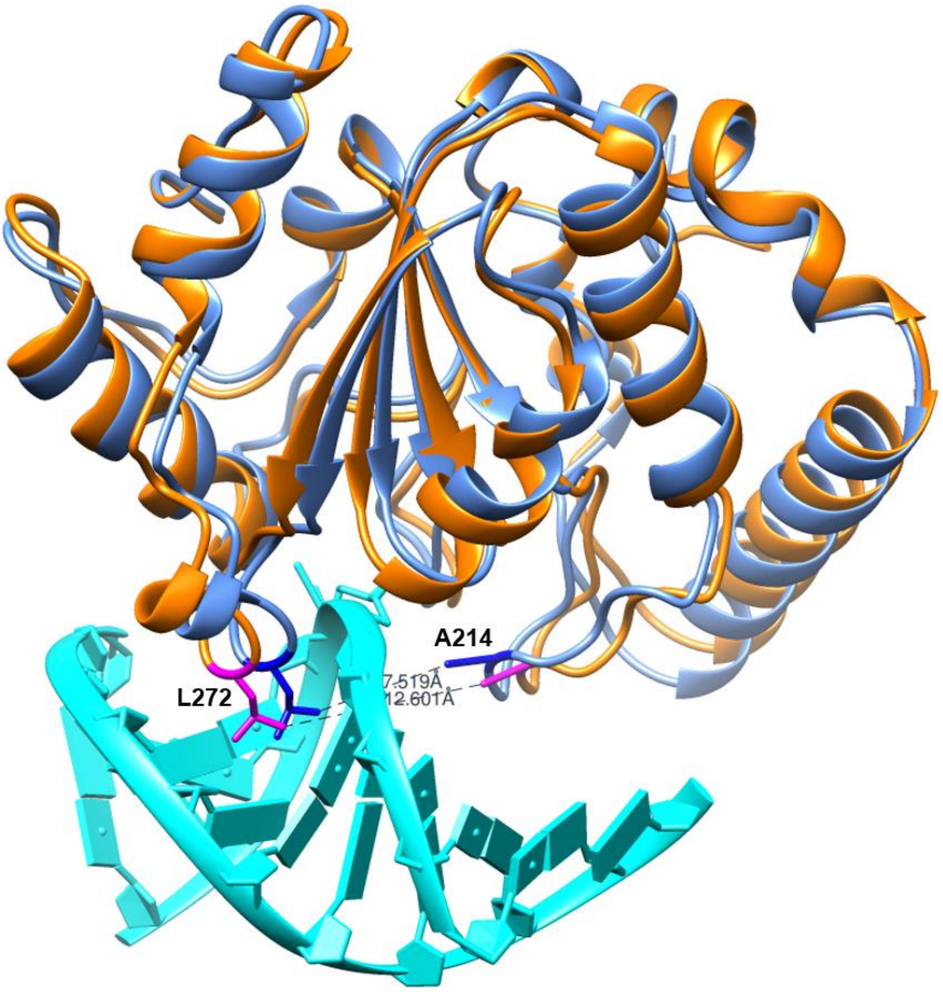


Figure 1.4.4a. Open and closed conformational changes of hUNG upon binding a double-stranded U-DNA substrate. 1AKZ (orange) = apo form of hUNG; 1SSP = hUNG (cornflower blue) complexed with dsDNA (cyan). The distance between atom CD1 of residue L272 and atom CB of residue A214 decreases from 12.601 Å in the open conformation (apo form of hUNG, magenta residues) to 7.519 Å in closed conformation (dsDNA-bound hUNG, medium blue residues). Cartoon structures were rendered in Chimera using matchmaker tool of 1AKZ and 1SSP. The matching was performed using 1AKZ as a reference structure.

The stereochemical landscape of the Ung DNA-binding cleft confers a local distortion to the DNA double helix. This distortion exaggerates the natural breathing motion of DNA bases. In the breathing motion, in which natural compression/torsional motions of the DNA induce distance and geometry fluctuations, the base pairs break spontaneously and re-form quickly[31].

The Ung reaction can be divided into four steps: (1) pinch, (2) push, (3) plug, and (4) pull (Figure 1.4.4b). These steps describe the nucleotide flipping mechanism.

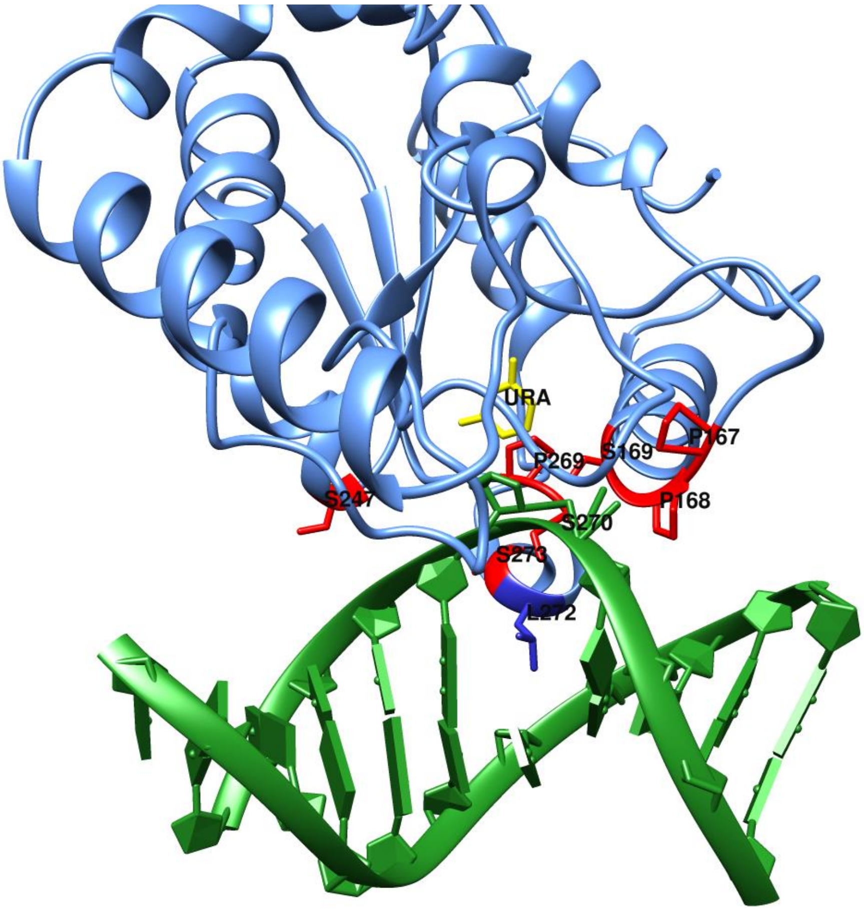


Figure 1.4.4b. Ung reaction steps shown on UNG2-dsDNA complex (PDB: 1SSP). The Ung reaction can be divided into four steps: (1) pinch, (2) push, (3) plug, and (4) pull. In the pinch step (coloured red residues), 4 conserved serine residues (In UNG2: S169 from the Pro-rich loop, S247 from the Gly-Ser loop, S270 and S273 from the DNA minor groove intercalation loop) and 3 conserved proline residues (in UNG2: P167 and P168 from the Pro-rich loop and P269 from the DNA minor groove intercalation loop) combine their effects to bend DNA by compressing the phosphate backbone (pinch) after the enzyme assisted (active) or spontaneous (passive) breathing motions that lead to flipping of deoxyuridine (dU). In the push step, the apical hydrophobic residue (coloured medium blue) of the DNA minor groove intercalation loop motif enters the DNA minor groove to displace deoxyuridine and form a pseudo base-pair with the partner base. This results in productive uracil-binding and an increased lifetime in the active site of the flipped-out uracil (coloured yellow, plug step). Finally, after cleavage of the N-glycosidic bond of the target base, the pull step takes place by retraction of the Ung minor groove intercalation loop apical hydrophobic residue[8].

Abasic sites might cause transcriptional mutagenesis as they can be bypassed by RNA polymerase[32]. Additionally, base insertion propensities opposite AP sites heavily skew towards insertion of adenine, which could lead to transition mutations[33]. Ung plays a protective role against that mutagenesis, it remains bound to (or binds back) cleaved products until the ensuing action of AP-endonuclease takes place in the BER pathway[34].

1.4.5. Ung roles beyond DNA-repair

Ung is known to play essential roles in innate and humoral immunity (Figure 1.4.5). The presence of uracil residues proximal to each other, at oligonucleotide distance, on opposite strands of DNA leads to the creation of double-strand breaks and the fragmentation of DNA upon the action of Ung in concert with AP-endonuclease. The forces of hydrogen bonds between complementary bases of DNA would not be sufficient to resist thermal disintegration of double stranded DNA at these short oligonucleotide lengths. Ung in this case would act as a restriction enzyme in the innate immunity of prokaryotes and eukaryotes[35,36].

Heavily uracilated DNA can occur naturally in certain bacteriophages, where thymine is entirely replaced by uracil in their genomes[37,38]. Additionally, human immunodeficiency viruses are known to involve heavily uracilated DNA in their reverse transcribed genome[39].

In addition, heavily uracilated pathogen DNA can be stimulated by the pathogen DNA detection in the cytoplasm through triggering the APOBEC (Apolipoprotein B mRNA Editing Catalytic polypeptide-like family) enzymes, which cause enzymatic cytosine deamination, and subsequently, the resulting accumulation of DNA uracil leads to BER-induced double strand DNA breaks (DSBs) due to proximity of Ung-generated abasic sites on opposite DNA strands[40].

Ung also plays an important role in humoral immunity in mammals, specifically in antibody diversification which relies on processes including somatic hypermutation (SHM) and class switch recombination (CSR). These processes rely on enzymatic cytidine deamination by activation-induced deaminase (AID) to produce uracil bases in DNA. In SHM, the action of Ung on the U/G mismatches created by AID creates abasic sites. The outcomes of these AP sites vary depending on a competition between BER and DNA replication. High-fidelity polymerases would be replaced by error-prone DNA polymerases if a replication fork arrives at the AP site. These error-prone polymerases would add any nucleotide opposite to the AP site and subsequently lead to a hypermutation in the variable region of immunoglobulin locus. In CSR, AID deaminate cytosine at specific G-rich tandem repeated DNA sequences called switch regions (S regions). Several cytosine residues are deaminated to uracil residues at both donor and acceptor S regions. Ung action on dsDNA produces abasic sites that are proximal to each other on opposite strands, leading subsequently upon action of AP-endonuclease to double strand DNA breaks (DSBs) and to detachment of the heavy chain variable region from the constant region μ of the immunoglobulin. Subsequent non-homologous end joining recombination events between a downstream constant region and the detached heavy chain (i.e., joining both donor and acceptor S regions or S-S recombination), involving several proteins including DNA-PK, ATM, Mre11-Rad50-Nbs1, γH2AX, 53BP1, Mdc1, and XRCC4-ligase IV, leads to switching of the immunoglobulin isotype[41,42].

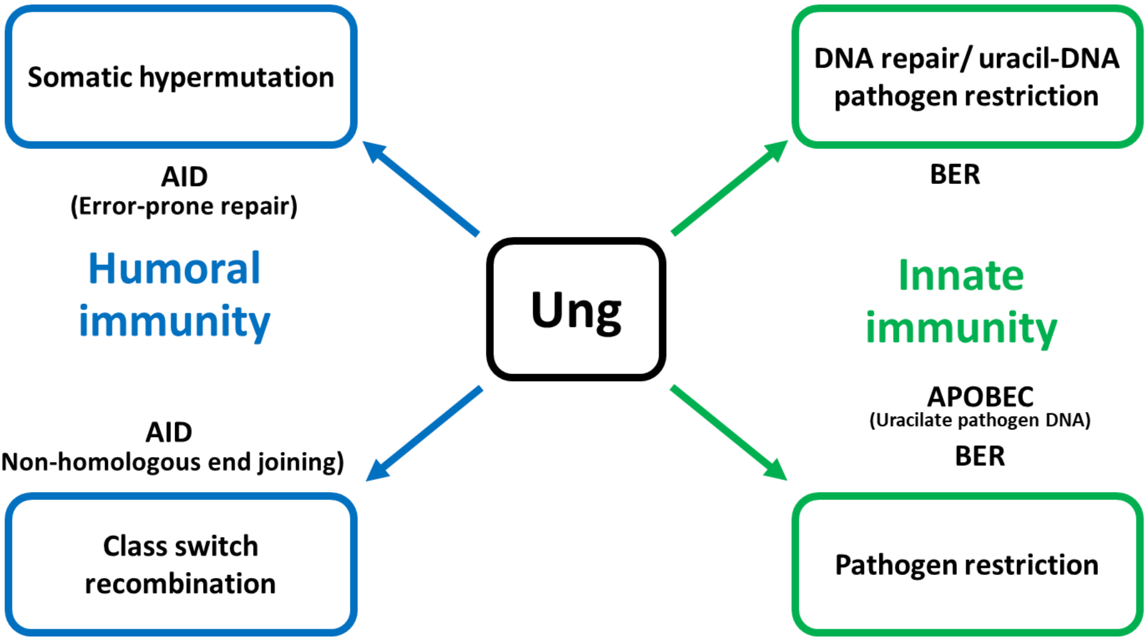


Figure 1.4.5. Ung roles in cellular programs. In the Innate immunity, Ung-initiated BER leads to uracil-DNA pathogen restriction; additionally, cytosine deamination of pathogen DNA by APOBEC enzymes leads to hyperuracilation and subsequent pathogen restriction by Ung. In the humoral immunity, Activation-Induced Cytidine Deaminase (AID) generates uracil residues in immunoglobulins; a downstream action of Ung to create abasic sites, followed by Error-prone repair complete the somatic hypermutation molecular process, while a downstream action of Ung followed by non-homologous end joining complete the class switch recombination process (Modified from Savva R, 2020)[40].

1.5. DNA mimic proteins

Ung provides innate cellular immunity by acting as a restriction enzyme of pathogens (section 1.4.5). There are organisms that utilise uracil-DNA or utilise Ung-sensitive DNA replication strategies. Some of these organisms are known to encode proteins that mimic physicochemical signatures of DNA and antagonise Ung-restriction.

Proteins that mimic DNA can be employed by viruses and other pathogens to strategically interfere with DNA regulation processes and innate cellular immunity defences. Several examples of proteins that specifically mimic the substrates of DNA binding proteins are known. These DNA mimic proteins control the DNA binding activity by occupying the DNA binding sites of those DNA binding proteins[43,44]. DNA mimic proteins are known to be encoded by some prokaryotes, eukaryotes, bacteriophages and eukaryotic viruses[45,46]. Several DNA regulatory mechanisms are known to be affected by DNA mimic proteins, including DNA packing[47–49], transcription[50–53], restriction-modification[54,55], recombination[56,57], and uracil-DNA repair[58–63]. DNA mimicry is also involved in p53 activity control[64], and in antibiotic resistance through protecting DNA gyrase from fluoroquinolones activity[65].

In order to mimic the DNA molecule, some of its characteristics must be imitated, especially the regular negative charge distribution on its two helical strands. Although the physicochemical properties of amino acids differ from those of nucleic acids, DNA mimic proteins achieve DNA negative charge distribution mimicry with the two acidic amino acids: glutamic acid (Glu or E) and aspartic acid (Asp or D). A complementary positive charge is usually present on the surface of proteins targeted by DNA mimic proteins, indicating the importance of charge-charge interaction to achieve the tight binding between the DNA mimic protein and its target. At least 24 DNA mimic proteins with different biological roles have been identified (Table 1.5); of which, 8 have been identified in the last 5 years, increasing the knowledge of their biological importance and function. Knowledge of these DNA mimic proteins and how they interact with their target proteins has stimulated the development of artificial systems that mimic DNA and specifically bind to certain targeted DNA-binding proteins. DNA mimicry-based therapeutic approaches open new avenues in molecular mimicry for protein surface recognition for challenging DNA binding therapeutic targets[66,67].

Table 1.5. Identified DNA mimic proteins

DNA mimic protein	Targeted DNA binding protein	Related function
UGI[22] (*Bacillus* phage PBS2)	Uracil-DNA glycosylase	DNA repair
p56[59] (*Bacillus* phage ϕ29)	Uracil-DNA glycosylase	DNA repair
SAUGI[62] (*Staphylococcus aureus*)	Uracil-DNA glycosylase	DNA repair
Vpr[63] (human immunodeficiency virus)	Uracil-DNA glycosylase	Virus replication; DNA repair
Ocr[54] (enterobacteria phage T7)	Type I restriction enzymes	Restriction
ArdA[55] (*Enterococcus faecalis*)	Type I and II restriction enzymes	Restriction
AcrF2[68] (*Pseudomonas* phage D3112)	CRISPR-Csy	Bacterial defence; gene editing
AcrF10[69] (*Shewanella xiamenensis* prophage)	CRISPR-Csy	Bacterial defence; gene editing
AcrIIA4[70,71] (*Listeria monocytogenes* prophage)	CRISPR-Cas9	Bacterial defence; gene editing
CarS[51] (*Myxococcus xanthus*)	CarA transcription repressor	Transcription
Gp44[53] (*Bacillus* phage SPO1)	RNA polymerase	Transcription
TAFII230[50] (residues 11−77) (*Drosophila*)	RNA polymerase	Transcription
DMP19[52,72] (*Neisseria meningitidis*)	NHTF transcription repressor; Nucleoid-associated protein HU	Transcription; DNA packaging
HI1450[48] (*Haemophilus influenzae* PittGG)	Nucleoid-associated protein HU	DNA packaging
DMP12[49] (*N. meningitidis*)	Nucleoid-associated protein HU	DNA packaging
Arn[51] (Enterobacteria phage T4)	Histone-like protein H-NS	DNA packaging
ICP11[73] (white spot syndrome virus)	Histone proteins	Nucleosome assembly
Gam[56] (bacteriophage λ)	RecBCD	Recombination
DinI[57] (*Escherichia coli* BL21)	RecA	Recombination
Mfpa[65] (*Mycobacterium tuberculosis*)	DNA gyrase	Topology
NuiA[74] (Nostoc sp.)	Nonspecific nuclease	Phosphodiester bond Digestion
AbbA[75] (*Bacillus subtilis*)	AbrB gene regulator	Gene expression
P53[46] transactivation domain (residues 33−60) (*Homo sapiens*)	Replication protein A	Single-strand binding
MBD3[76] (residue 263–286) (*Homo sapiens*)	ADAR1 (Zα domain; residue 133–209)	B-Z DNA transition

1.6. Uracil-DNA Glycosylase inhibitors

DNA mimicry appears to have independently arisen as a counter to Ungs as restriction enzymes in innate cellular immunity. There are four known different types of Ung inhibitors (UngIns) belonging to three architecturally discrete families: *Bacillus* phage Ugi and its structural homolog *Staphylococcus aureus* Uracil-DNA glycosylase inhibitor (SAUGI), *Bacillus* phage Phi29 protein p56, and human immunodeficiency virus (HIV) viral protein R (Vpr)[22,62,63,77]. All these UngIns use a charge-based alignment to dock to the Ung-DNA binding cleft, they also converge upon a universal mechanism of Ung inhibition via hydrophobic sequestration of an essential residue for Ung catalytic activity on the apex of DNA minor groove intercalation loop (section 1.4.4).

1.6.1. Uracil-DNA glycosylase inhibitory protein (Ugi)

Bacillus subtilis phages PBS1 and AR9 use uracil instead of thymine in their DNA[38,78]. This strategy grants potential advantages to these phages by making their DNA resistant to many endonucleases[79]. Ung acts as a restriction factor if the uracil bases are closely spaced on the opposite strands of dsDNA[35,36]; therefore, phages that have uracil-DNA must subvert the host cell uracil-DNA glycosylase activity to survive. The inhibition of Ung is achieved by these phages through encoding a small inhibitory protein called Ugi, which forms a tight complex with Ung and inhibits it completely. Ugi is an 84 amino acid protein that has a significant resistance to thermal denaturation[80]. The gene that encodes Ugi has an identical DNA sequence in PBS1 and AR9 phages[78]. In addition to *B. subtilis* Ung inhibition, Ugi has been shown to inactivate Ung proteins from bacteria, mammals, as well as mammalian viruses. Identification of the structure of Ugi-Ung complex showed that Ugi inactivates Ung by forming a 1:1 stoichiometric complex[23].

1.6.1.1. Ugi structure

The structure of Ugi includes 5 antiparallel β-strands that form a highly twisted β-sheet sandwiched between 2 short α-helices (Figure 1.6.1.1). This structure was consistent whether determined in an Ung-Ugi complex or by the study of a free inhibitor, showing that inhibition is achieved without gross structural changes of Ugi upon binding to Ung[81].

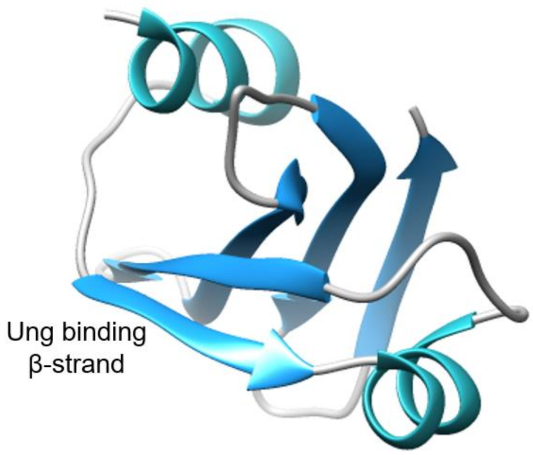


Figure 1.6.1.1. Secondary structure of Ugi in the Ugi-Ung complex (PDB ID: 1UUG, chain B). A highly twisted β-sheet composed of 5 anti-parallel strands lies between 2 α-helices. Ribbon cartoon representation was rendered in Chimera.

1.6.1.2. Ugi-Ung interaction

Crystal structures of Ugi-Ung complex have been identified with different variants of Ung including hUNG, *E. coli* Ung, and HSV-1 UNG. Ugi mimics DNA backbone interactions with Ung. It forms a tight complex by targeting the DNA-binding cleft of Ung and prevents Ung-DNA binding. Moreover, Ugi can dissociate Ung from a complex of Ung-DNA[8] due to higher binding affinity and lower energetic costs. Ugi-Ung complex formation does not require the energetic costs paid in the conformational changes that take place upon Ung-DNA complex formation. Hence, Ugi binds Ung with high affinity, while the Ung-DNA complex has a low affinity. This low affinity is useful for releasing the abasic site to later enzymes in the BER pathway[81].

Several contacts have been identified in the Ugi-Ung interface. The Ugi-Ung interaction does not involve an induced fit mechanism. The shape and electrostatic complementarity, and crucially, the hydrophobic binding of Ung by Ugi (Figure 1.6.1.2a), are the hallmarks of this interaction.

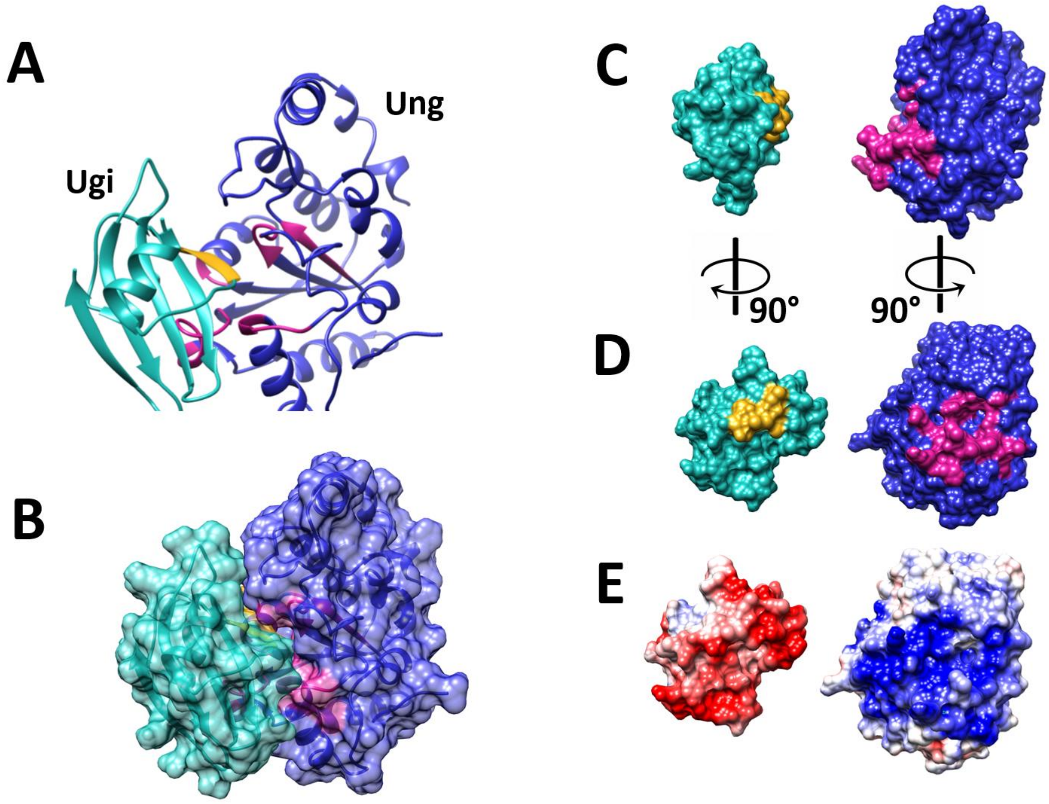


Figure 1.6.1.2a. Ugi-Ung interaction (PDB ID: 1UUG, chains A-B). (A) Ribbon view of Ugi-Ung complex: Ugi (coloured light sea green) occupies the Ung DNA-binding cleft by its Ung binding strand (coloured golden), Ung (coloured medium blue) conserved motifs are coloured violet. (B) a 50% transparency surface view of part A. (C) Ugi and Ung are separated and considered as 2 separate models to show the shape of Ung binding strand of Ugi (golden) and the DNA-binding cleft of Ung (violet). (D) the same components of part C with rotation of both Ugi and Ung for more clarity. (E) Coulombic surface colouring showing the electrostatic complementarity of the Ugi negative charge (red) and Ung positive charge (blue). Depictions were rendered in Chimera.

Eight hydrophobic residues of Ugi (Met24, Val29, Val32, Ile33, Val43, Met56, Leu58, and Val71) form Van der Waals bonds with the DNA minor groove intercalation loop apical residue (Figure 1.6.1.2b). Six acidic residues of Ugi (Glu20, Glu27, Glu30, Glu31, Asp61 and Glu78) form electrostatic interactions with key active site residues of Ung (in *E. coli* Ung: Gln63, Asp64, Tyr66, His67, and His187). A two-step model was suggested for Ugi association[81]. The first step is a charge-based alignment, from which docking may proceed. Second, nanomolar complex formation proceeds via tight docking mediated by hydrophobic sequestration of the Ung minor groove intercalating loop apical hydrophobic residue and a concomitant conformational change of the invariant Ugi E20 residue[62,81]. E20I and E28L mutants of Ugi apparently form a lower affinity complex with *E. coli* Ung[82] (wild-type Ugi could displace these mutants from their complexes with Ung), suggesting a role of these 2 residues in the locking mechanism. However, other mutants including E27A, E30L, E31L, D61G, and E78V formed high affinity Ugi-Ung complexes (wild-type Ugi was unable to displace these mutants from their complexes with Ung)[82]. On the other hand, mutational studies of the *E. coli* Ung leucine residue in the DNA minor groove intercalation loop including L191G, L191A, L191V, and L191F showed, with the exception of the L191F mutant, reduced stability of the complex with Ugi. Nonetheless, both L191V and L191F retained the enzymatic activity of wild type Ung[81]. Unlike DNA binding to Ung, Ugi binding to Ung causes only minimal conformational changes; i.e., it does not induce the Ung closed conformation.

The interface generated by all interactions in the Ugi-Ung complex has surface area of ~1100 Å^2, this interface represents approximately a quarter of the accessible Ugi surface area and an eighth of that for *E. coli* Ung. The total buried surface area of the *E. coli* Ung-Ugi complex and hUNG-Ugi complex is ~2200 Å^2. While the total buried surface area of the HSV UNG-Ugi complex is ~2000 Å^2. [22,81,83]

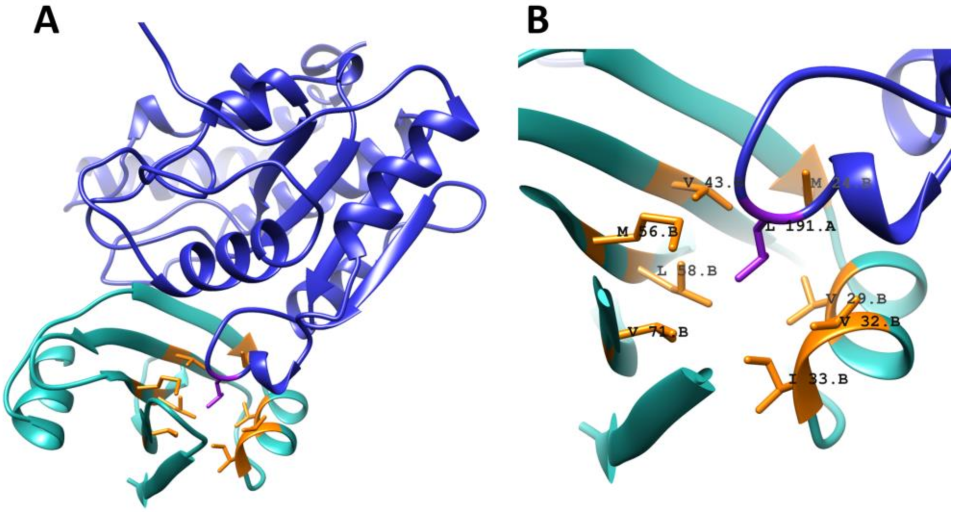


Figure 1.6.1.2b. Hydrophobic sequestration of Ung minor groove intercalation loop apical hydrophobic residue by eight hydrophobic residues of Ugi (PDB ID: 1UUG, chains A-B). (A) Ribbon view of Ugi-Ung complex: Ugi (coloured light sea green) eight hydrophobic residues (coloured orange) bind Ung (coloured medium blue) minor groove intercalation loop apical hydrophobic residue (coloured purple). (B) A focused view of part A showing the labels of interacting hydrophobic residues. Cartoon depictions were rendered in Chimera.

1.6.1.3. Ugi DNA mimicry

Ung inhibition by Ugi depends upon Ugi's mimicry of Ung-bound DNA. A structural comparison between the *E. coli* Ung-bound Ugi and hUNG-bound DNA demonstrates that Ugi possesses astonishing overall similarities with DNA deformed by Ung binding. The twisted β-sheet of Ugi mimics the bent shape of the Ung-bound DNA. Ugi 1st β-strand has distortions along its edge that enable it to follow the path of the compressed DNA first strand backbone more closely. The second strand of DNA can also be traced within Ugi (Figure 1.6.1.3). The Ung-Ugi complex possesses more affinity than the Ung-DNA complex due to stronger hydrophobic interactions. The first sequesters a total accessible surface area of ~2000 Å^2 as compared with ~1690 Å^2, ~1640 Å^2, and ~1550 Å^2 for U:A, U:G, and abasic DNA substrates with hUNG, respectively[81]. Additionally, the fact that only minor conformational changes of Ung are induced upon binding Ugi (i.e. lower energetic costs) contributes to this superior affinity of Ung-Ugi complex.

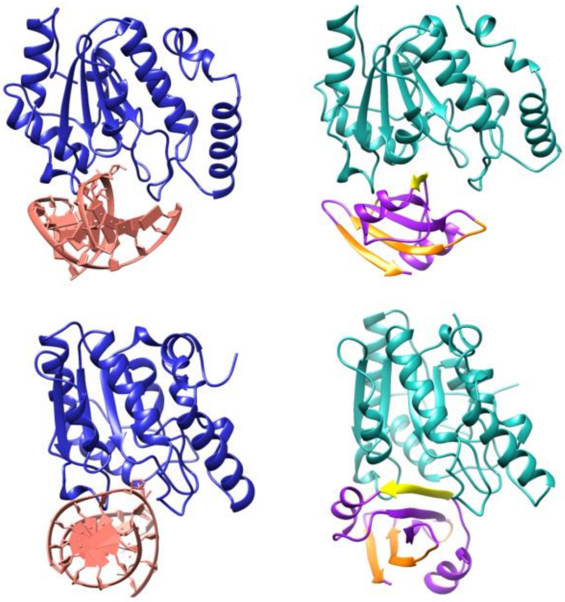


Figure 1.6.1.3. Structural comparison of Ugi and dsDNA when they are bound to Ung. Left column shows 2 different views of dsDNA-Ung complex [PDB ID:1SSP; dsDNA (chains A-B) is coloured salmon, Ung (chain E) is coloured medium blue]. Right column shows 2 different views of Ugi-Ung complex [PDB ID: 1UUG; Ugi (chain B) is coloured purple, Ung (chain A) is coloured light sea green]. The twisted β-sheet of Ugi mimics the bent shape of the dsDNA. Ugi 1st β-strand (coloured yellow) follows the path of the compressed DNA backbone. Ugi traces 2nd strand of DNA by the loop between 3rd and 4th strands, the edge of β-sheet with the 4th and 5th strands, and the loop between the 2nd and 3rd strand (coloured orange secondary structures). Cartoon depictions were rendered in Chimera.

1.6.2. Protein p56

The genome of *B. Subtilis* phage Φ29 is linear dsDNA that has a covalently linked terminal protein (TP) at both 5’ ends. Unlike phage PBS1, phage Φ29 genome does not include uracil residues. Replication of phage Φ29 DNA in infected cells, based on a protein-primed mechanism, generates different types of replication intermediates that have ssDNA regions of various lengths. Cytosine deamination or dUMP misincorporation can cause the emergence of uracil within these ssDNA regions. This uracil is vulnerable to the host BER process, and Ung activity followed by an AP endonuclease activity can subvert the replication intermediates by creating substrate sites for branchpoints. Phage Φ29 encodes a small protein of 56 amino acids called p56, this protein has an Ung inhibitory activity that ensures the integrity of BER-vulnerable replication intermediates[59].

1.6.2.1. p56 properties

Several members of Φ29 phage family, *Salasmaviridae*, are known to encode p56 homologous sequences with different protein lengths[31,60,84]. Protein p56 has been shown to be a dimer under physiological conditions. Each monomer comprises 3 anti-parallel β-strands and an α-helix, connected by three loops (Figure 1.6.2.1). The α-helix lies with parallel orientation against the β1-strand, forming a hydrophobic core. In the dimeric form, which includes a sheet of 6 β-strands, the hydrophobic core is formed by the 2 α-helices lying against each other at the same side of the β-sheet. The β3 strands of both monomers face each other on the inner side of the dimer[77].

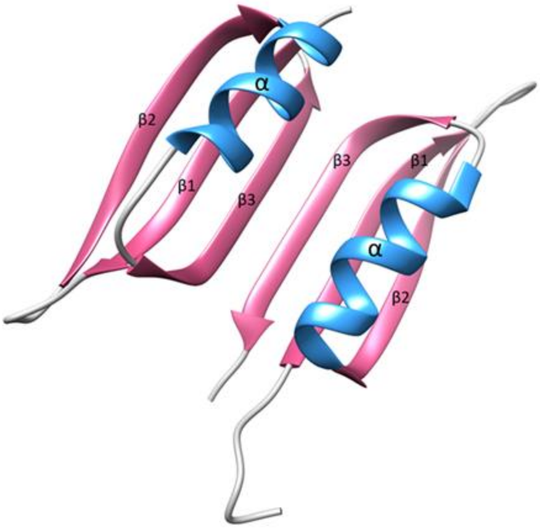


Figure 1.6.2.1. Secondary structure of p56 dimer in the p56-Ung complex (PDB ID: 4L5N, chains E and F). Three anti-parallel strands of each monomer are labelled and are coloured pink. The α-helices that are parallel to 1st strands of both monomers are coloured blue. Ribbon cartoon representation was rendered in Chimera[29].

1.6.2.2. p56-Ung interaction

Protein p56 inactivates Ung in a 2:1 stoichiometry. Previous isothermal titration microcalorimetry experiments showed that it forms a tight 2:1 complex with Ung[77]. Interestingly, the crystallization of p56-BsUng complex showed that the 2 subunits of p56 bind to BsUng in non-symmetrical way. One of the subunits contributes to more than 80% (~ 5/6) of the total buried area of the complex interface (1510 Å^2 of 1815 Å^2). The α-helix of this subunit locates in the DNA-binding cleft of Ung. The protruding leucine residue of the DNA minor groove intercalation loop of *E. coli* Ung and hUNG is replaced with a phenylalanine residue in BsUng. This phenylalanine residue (Phe191) is recognized by p56 and located in the hydrophobic pocket of the p56 dimer that is formed by symmetrical residues of p56 dimer (Phe36, Glu37, and Tyr40 from both α-helices; Figure 1.6.2.2). Glu37 and Tyr40 residues of each monomer also have important roles in dimer stability by making hydrogen bonds with the Tyr40 and Glu37 residues of the other monomer, respectively[85].

Additional to Phe191 sequestration, several hydrophobic interactions were observed between BsUng and p56 subunits. Moreover, 19 polar interactions were found in the complex, mainly involving the α-helix located in the BsUng DNA-binding cleft. These polar interactions strengthen the protein-protein interactions (PPIs) and contribute to the complex stability[85].

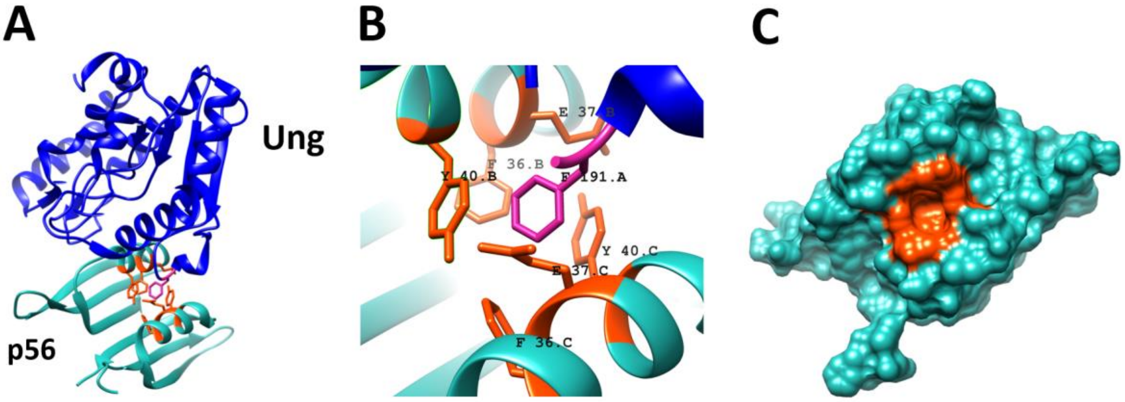


Figure 1.6.2.2. p56-Ung interaction (PDB ID: 3ZOQ). (A) Ribbon view of p56-Ung complex: p56 dimer (coloured light sea green) occupies Ung DNA-binding cleft by an α-helix of one of its subunits, and sequester the Ung (coloured medium blue) minor groove intercalation loop apical hydrophobic residue (Phe191, coloured violet) with six hydrophobic residues (coloured orange) on the 2 α-helices. (B) A focused view of p56 symmetric residues that make hydrophobic interactions with Ung minor groove intercalation loop apical hydrophobic residue (Phe191), all labelled with one letter identifier. (C) A surface view of p56 dimer showing the shape of hydrophobic pocket (orange) that is formed by the six symmetric residues on both α-helices. Depictions were rendered in Chimera[29]

1.6.2.3. p56 DNA mimicry

p56 binds Ung using DNA stereochemical mimicry. The 2 α-helices of p56 dimer fit in the positions usually located by DNA strands upon binding to Ung (Figure 1.6.2.3). Similar to the Ung-Ugi complex, the Ung-p56 complex possesses greater affinity than an Ung-DNA complex. The Ung-p56 complex sequesters a total accessible surface area of ~1815 Å^2 as compared with ~1690 Å^2, ~1640 Å^2, and ~1550 Å^2 for U:A, U:G, and abasic DNA substrates with hUNG, respectively[81,85]. Protein p56 also mimics the contacts of DNA backbone phosphates at position -1, 0, +1, and +2 of the flipped-out uracil site with Ung, using several residues spanning between Glu26 and Asn42. This Ung-specific DNA-mimicry of p56 is quite similar to the Ugi mimicry of DNA bound to Ung. However, in vitro, Ugi displaces p56 to dissociate it from a p56-Ung complex due to extended and tighter binding to Ung in comparison with p56[58,61].

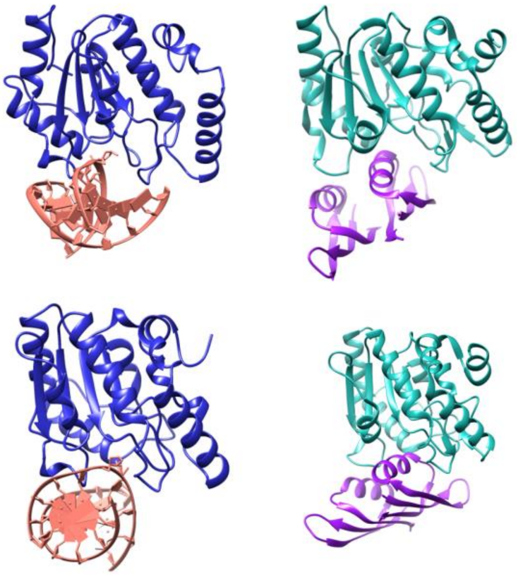


Figure 1.6.2.3. Structural comparison of p56 and dsDNA when they are bound to Ung. Left column shows 2 different views of dsDNA-Ung complex [PDB ID:1SSP; dsDNA (chains A-B) is coloured salmon, Ung (chain E) is coloured medium blue]. Right column shows 2 different views of p56-Ung complex [PDB ID: 3ZOQ; p56 (chains B-C) are coloured purple, Ung (chain A) is coloured light sea green]. The α-helices from both dimer subunits locate at similar positions that dsDNA strands are fitted to when binding Ung. Cartoon depictions were rendered in Chimera[29].

1.6.3. *Staphylococcus aureus* uracil-DNA glycosylase inhibitory protein (SAUGI)

Staphylococcus aureus, like all other eubacteria, encodes its own uracil-DNA glycosylase (SAUNG). Intriguingly, methicillin resistant *S. aureus* (MRSA) encodes also an Ung inhibitor (UngIn) called SAUGI[62]. Although it is the 1st reported UngIn to be encoded by a non-viral organism, SAUGI is encoded by a horizontally-transferred mobile genetic element that is possibly a phage-derived DNA[31]. SAUGI is a small protein of 112 amino acids, with uncertain biological function(s) in Ung-encoding microorganisms.

1.6.3.1. SAUGI properties

SAUGI is a conserved protein in all MRSA strains[86]. SAUGI shares the same fold with Ugi; However, the highly twisted β-sheet of SAUGI is composed of 6 antiparallel strands instead of the 5 strands found in the Ugi β-sheet. In addition, the SAUGI structure includes three 3_{10}-helices and 2 α-helices. The distribution of negative charge on the surface of SAUGI is similar to that on the surface of Ugi. However, the sequence homology between these 2 proteins is very weak (Figure 1.6.3.1).

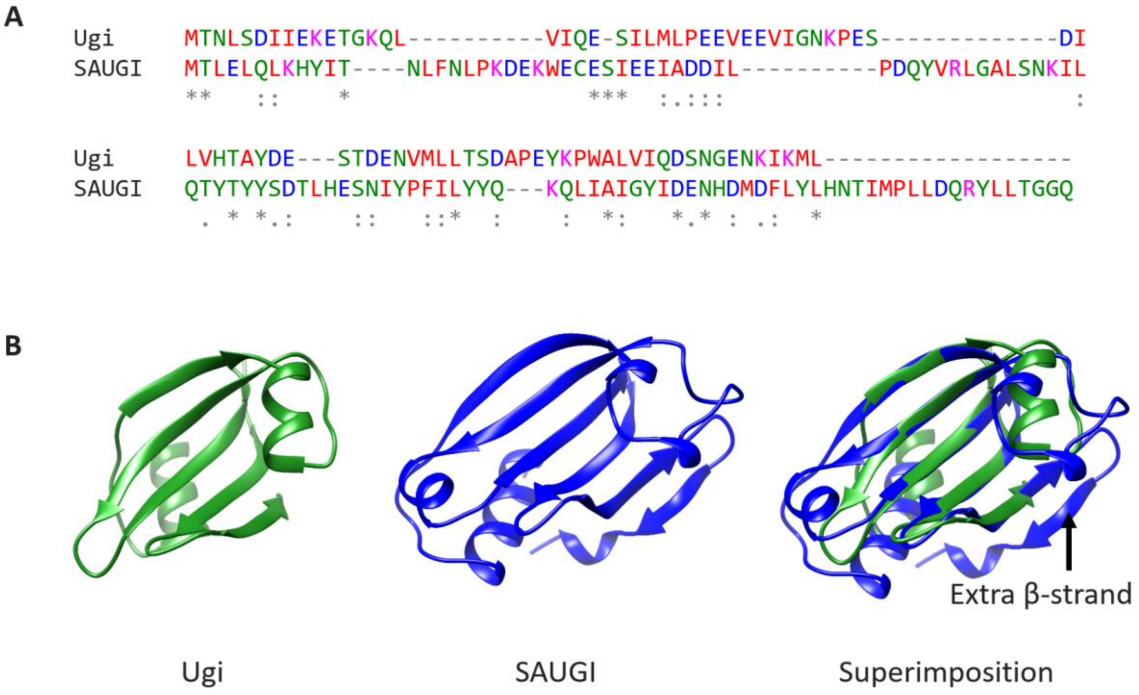

A
Ugi MTNLSDIIEKETGKQL----------VIQE-SILMLPEEVEEVIGNKPES-------------DI
SAUGI MTLELQLKHYIT----NLFNLPKDEKWECESIEEIADDIL----------PDQYVRLGALSNKIL
Ugi LVHTAYDE---STDENVMLLTSDAPEYKPWALVIQDSNGENKIKML-----------------
SAUGI QTYTYYSDTLHESNIYPFILYYQ---KQLIAIGYIDENHDMDFLYLHNTIMPLLDQRYLLTGGQ
B
Extra β-strand
Ugi
SAUGI
Superimposition

Figure 1.6.3.1. Comparison of Ugi and SAUGI sequences and structures. A) a structure-based sequence alignment of Ugi and SAUGI showing very poor sequence similarity. (B) The structure of Ugi (PDB:1UDI, chain I, coloured green), the structure of SAUGI (PDB ID: 3WDG, chain B, coloured blue), and a superposition of both structures. Ugi/SAUGI share the same fold at low level of sequence similarity.

1.6.3.2. SAUGI-Ung interaction

The crystal structure of an SAUGI-SAUNG complex showed inactivation of Ung via 1:1 stoichiometric complex formation. SAUGI targets the DNA-binding cleft of Ung via its 1st β-strand which mimics the Ugi 1st β-strand in Ung binding. About 20 hydrogen bonds are found in the complex of SAUGI-SAUNG, 12 of them are direct bonds without any water mediation. SAUGI has 5 hydrophobic residues that perform a hydrophobic sequestration of Ung minor groove intercalation loop apical hydrophobic residue (L184 in SAUNG). These conserved hydrophobic residues are I35, F69, L71, I83 and M89 (Figure 1.6.3.2). These hydrophobic interactions cement the SAUGI-SAUNG complex. Also contributing to complex stability are additional nonpolar interactions via SAUGI residues T55 and Y67 that lie against side chains of SAUNG residues P82 and P183, respectively. All these interactions together result in the tight binding of SAUGI and SAUNG. SAUGI associates rapidly with SAUNG, causing only minimal conformational changes upon Ung-binding.

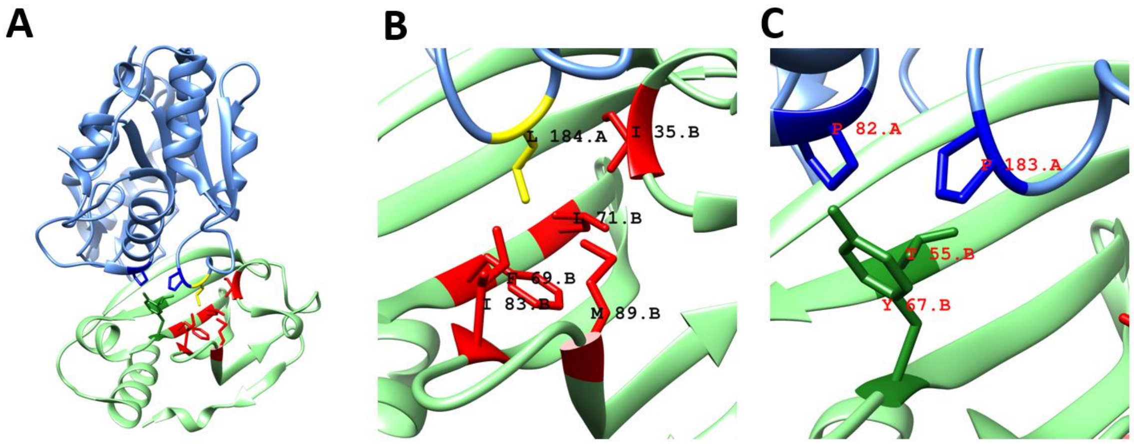


Figure 1.6.3.2. SAUGI-SAUNG interaction (PDB ID: 3WDG). (A) Ribbon view of SAUGI-SAUNG complex: SAUGI (coloured light green) occupies Ung DNA-binding cleft by its 1st β-strand, and sequesters the Ung (coloured cornflower blue) minor groove intercalation loop apical hydrophobic residue (coloured yellow) with five residues (coloured red). Two other nonpolar interactions are observed by T55 & Y67 residues (coloured forest green) of SAUGI that bind P82 & P183 residues (coloured blue) of SAUNG. (B) A focused view of hydrophobic sequestration of SAUNG minor groove intercalation loop apical hydrophobic residue (leucine) using the same colours as in (A) and labelling the involved residues of this hydrophobic sequestration. (C) A focused view of the other nonpolar interactions using the same colours as of (A) and labelling the involved residues in these interactions.

SAUGI has shown the ability to bind Ungs from *S. aureus*, human, Herpes simplex virus (HSV) and Epstein-Barr virus (EBV), with its greatest affinity in binding HSV and EBV Ungs[87,88]. SAUGI associates to UNG2 at a slow rate. However, Ugi showed a 12-13 fold greater affinity to binding UNG2 than SAUGI, and about twice the affinity in binding SAUNG[62].

1.6.4. Human Immunodeficiency Virus (HIV) Viral protein R (Vpr)

The Vpr encoding gene is conserved among HIV-1, HIV-2, and SIV immunodeficiency viruses[89]. Vpr was discovered more than three decades ago[90], it was known to bind UNG2 but its ability to inhibit UNG was not reported until 2016, when the structure of its multi-protein complex with UNG2, DNA damage-binding protein 1 (DDB-1), and CUL4A-associated factor 1 (DCAF1) was crystalized[63]. Interestingly, Vpr is the 1st reported UngIn to be encoded by viruses that infect mammals. Vpr is a small protein of 96 amino acids; despite its small size, it interacts with different host proteins to take part in multiple functions in the viral life cycle, thereby earning the analogy of a molecular Swiss-Army-Knife. These functions include pre-integration-complexes nuclear translocation, LTR transcription activation, cell-cycle arrest, and CD4$^+$ T cell dysfunction. Additionally, Vpr has shown to play a central role in reverse transcription of HIV-1 via the recruitment of UNG2, as well as downregulation of UNG2 mRNA transcripts[39,46,89].

1.6.4.1. The structure of Vpr

The secondary structure of Vpr includes mainly three amphiphilic α-helices accompanied by loops (Figure 1.6.4.1). The N-terminal domain, which binds Ung, is negatively charged, while the C-terminal domain is positively charged. Both N-terminal and C-terminal regions (spanning from M1 to N13 and from H78 to S96, respectively) are flexible[91].

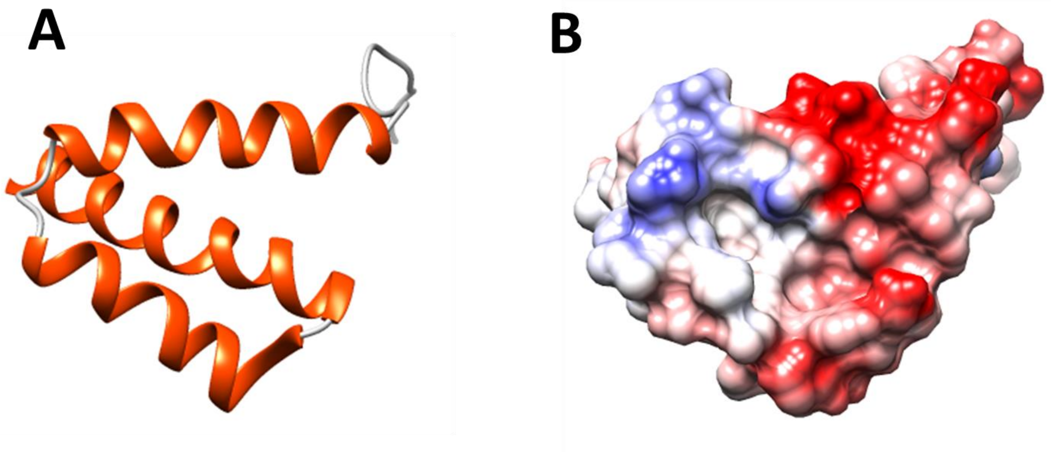


Figure 1.6.4.1. The structure of Vpr in the DDB1-DCFA1-Vpr-UNG2 complex (PDB ID: 5JK7, chains F). (A) Ribbon representation of Vpr structure. The bundled α-helices are coloured orange while the loop connectors are coloured light gray. (B) The negative charge distribution on the UNG2-binding interface of Vpr.

1.6.4.2. Vpr-Ung interaction

HIV-1 Vpr antagonises UNG2 by loading it onto the CRL4-DCAF1 E3 ubiquitin ligase complex for subsequent proteasome-mediated degradation[92]. Interestingly, like other UngIns, Vpr uses molecular mimicry of DNA and docks to the DNA-binding cleft of UNG2. Vpr is the only reported Ung inhibitory protein to antagonise Ung in multiple ways: The proteasome-mediated degradation of UNG2 via the molecular mimicry of DNA, and the downregulation of UNG2 mRNA[39,63]. The interaction of Vpr with different partners in multi-protein complex is achieved by 4 important structural motifs: A pair of these motifs (N-terminal tail and C-terminal region of helix α3) are used to interact with DCAF1 and another pair (a hydrophobic cleft between helices α1, α2, and the first turn of α3, and the loop connecting α2 and α3) are used to interact with UNG2 (Figure 1.6.4.2a).

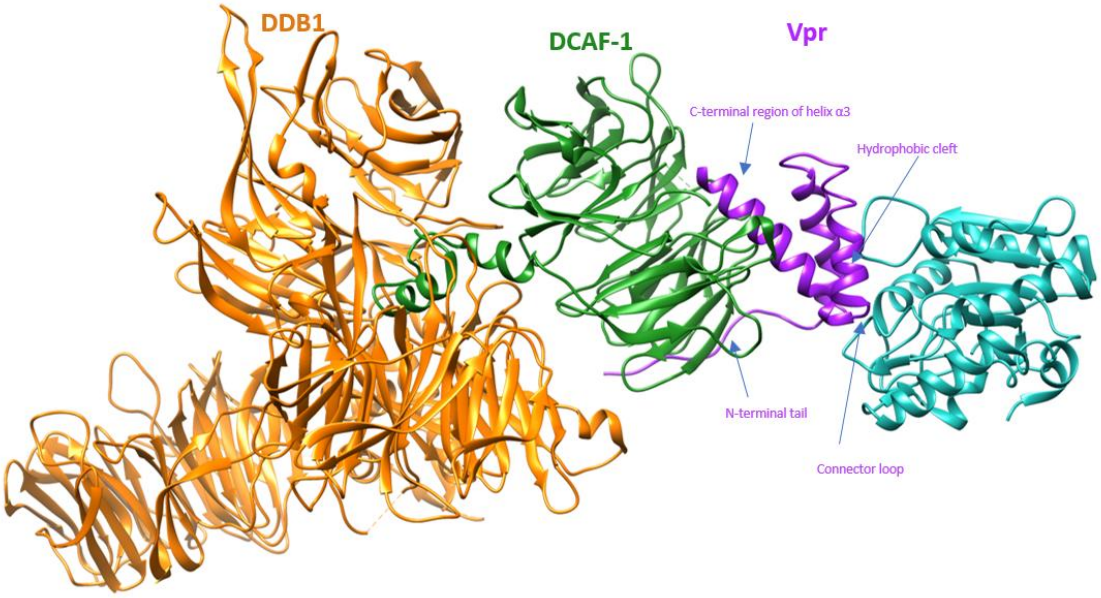


Figure 1.6.4.2a. The structure of DDB1-DCAF1-Vpr-UNG2 complex (PDB ID: 5JK7; chains B, D, E, and F). Ribbon view of the complex-structure shows how four Vpr motifs (referred to by purple arrows) are located to interact with DCAF-1 and UNG2. N-terminal tail and C-terminal region of helix α3 are used to interact with DCAF1. A hydrophobic cleft between helices α1, α2, and the first turn of α3, and the loop connecting α2 and α3 participate in hydrophobic sequestration of UNG2 minor groove intercalation loop apical hydrophobic residue (leucine). Cartoons depictions were rendered in Chimera.

Upon binding Vpr, UNG2 minor groove intercalation loop apical hydrophobic residue (L272) that is used to enter the minor groove of DNA in UNG2-DNA complex is inserted deeply to the hydrophobic cleft of Vpr (Figure 1.6.4.2b). This leucine residue is important for UNG2-Vpr interaction as L272D mutant of UNG2 has a significantly affected interaction with Vpr. In addition, Vpr binds the proline rich motif (165-PPPPS-169) of UNG2 via the loop connecting the α2 and α3 helices, creating a second interface. Y47, D52 and W54 residues of Vpr make stacking interactions with UNG2 (Figure 1.6.4.2b). The total buried area of Vpr-UNG2 interface is 940 Å^2. Intriguingly, W54R and W54G mutants of Vpr do not bind UNG2, revealing the importance of the Vpr W54 residue. Vpr mimics the DNA phosphate backbone around the abasic site in the UNG2-DNA complex via the loop connecting the α2 and α3 helices.

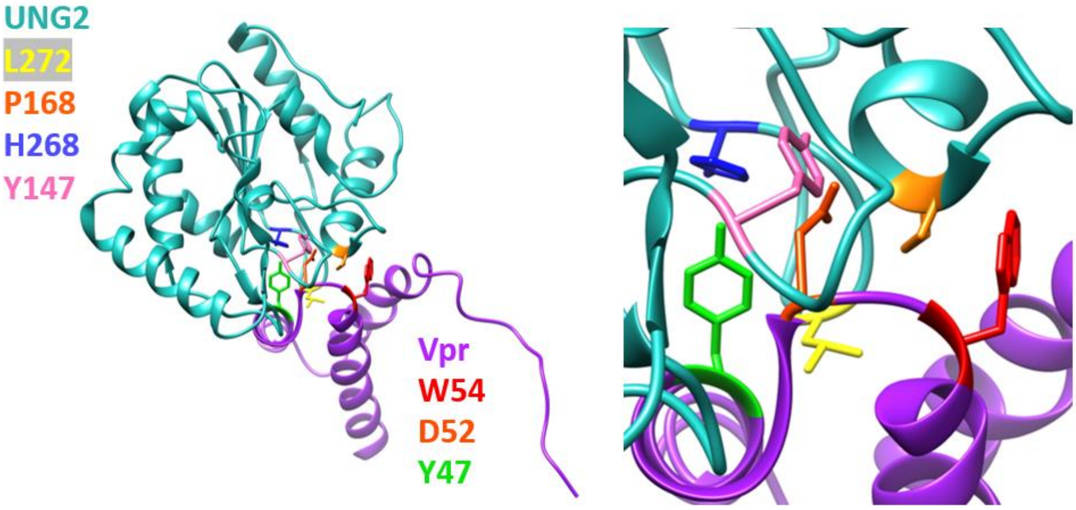


Figure 1.6.4.2b. HIV-1 Vpr-UNG2 interaction (PDB ID: 5JK7; chains E and F). UNG2 minor groove intercalation loop apical hydrophobic residue (yellow) is sequestered in the hydrophobic cleft of Vpr. UNG2 proline residue P168 (orange) interacts with key Vpr residues W54 (red) and D52 (orange red). Vpr residues Y47 (green) and D52 make hydrogen bonds with H268 (blue) and Y147 (hot pink) of UNG2, respectively. The left cartoon shows a total view of Vpr-UNG2 complex, while the right cartoon shows a focused view of the mentioned interaction-key-residues.

1.7. Motivations for this study

All the known UngIns are DNA mimic proteins, belonging to three architecturally discrete families converging upon a universal mechanism of Ung inhibition via hydrophobic sequestration of an essential residue for Ung catalytic activity[31], and even within a single architecture there is pronounced sequence diversity. Such heterogeneity has hindered the unambiguous identification of expected UngIns in the genomes of uracil-DNA phages closely related to PBS1 and AR9: *Yersinia* phage PhiR1-37[37], *Staphylococcus* phages[93,94], and *Listeria* phage LPJP1[95]. In addition, some Phi29-like phages including phages DK2, DK3, and vB_BthP-Goe4 have been identified without obvious Ung inhibitor encoding sequences in their genomes, despite pronounced similarities to the Phi29 phage genome[96,97]. There are also other organisms that are reported to encode functions that inhibit Ung such as *Escherichia* phage T5; however, no UngIn has been identified in its genome[98]. It is known that for most protein families, protein structures are more conserved than sequences[99]. It is possible, given their strategic utility to viruses, that UngIn sequence variants that have a known UngIn fold but undetectable sequence similarity may exist in other annotated genomes. It's also possible that novel types of UDG activity-modulating proteins exist in organisms without reported Ung inhibitory activity, such as the uracil-DNA roseophages that are not related to Ugi encoding uracil-DNA phages[100].

1.7.1. The aim of this study

The aim of this study was to identify new naturally occurring UngIns that are undetectable by simple sequence similarity search tools. We were able to produce an expanded repertoire of synthetic/natural-occurring UngIns that is more statistically powerful when used in searching for novel natural UngIns (chapter 3). Furthermore, we modelled known sequence variations in these proteins and generated libraries from that information to get a clearer insight into the nature of sequence plasticity at the structure-function level (chapter 3).

Understanding the way in which variant sequences can still underpin structures highly specific for Ung and its inhibition is important, because we still do not understand how widespread Ung targeting is by viruses or how such a promutagenic strategy as is consequent of Ung inhibition can be so advantageous as to have been evolved independently in several contexts. We also determined the structures of various distantly related sequences encoding an UngIn of the Ugi/SAUGI fold type and of the p56 fold type (chapter 4). A novel rapid bacterial conditional lethal assay for Ung inhibition was developed (chapter 3), this assay was used as an alternative to purely *in-silico* sequence-based searches to attempt to find novel types or new sequences of Ung inhibitors in the genomes known or expected to encode activities that inhibit uracil-DNA glycosylases (chapter 5).

Chapter 2

2. Materials and methods

All the techniques that were used during the course of this project are articulated in this chapter. The work presented in this thesis included laboratory work and computational work, hence this chapter is divided into 2 major sections: Laboratory methods and Computational methods. Wet-lab methods included DNA manipulation, protein manipulation, and laboratory X-ray crystallography. Computational methods included bioinformatics sequence search methods and computational structural biology methods. All these methods are articulated, along with a discussion of each technique.

2.1. Laboratory methods

2.1.1. DNA methods

DNA methods were based on both natural and synthetic DNA precursors: (1) Plasmid DNA and (2) oligonucleotides and assembled synthetic DNA. Plasmid DNA sources were generally from long term archives in the lab, but where sources are known they are indicated. Synthetic DNA was generally obtained from a limited number of sources and suppliers are indicated.

2.1.1.1. Polymerase Chain Reaction (PCR)

PCR is one of the most powerful technologies used in molecular biology to amplify specific sequences of genomic/synthetic DNA for cloning and mutagenesis purposes. PCR is based on repeated thermal cycles of heating and cooling to facilitate DNA replication by enzymatic reaction. Standard PCR include three main steps: (1) denaturation of the template dsDNA, (2) annealing of primers (oligonucleotides that bind specific DNA sequences to guide DNA polymerase replication) to ssDNA, and (3) extension of newly synthesised DNA. Since DNA polymerases elongate DNA only in the 5' to 3' direction, each pair of primers was designed to anneal on opposite strands at the 5' ends of amplification-targeted DNA. Extension with primers leads to duplication of desired region of template DNA.

Conventional PCR and variations of it were used in this project. Inverse PCR (iPCR), a variant PCR that uses circular DNA (cDNA) as a DNA template, was used to linearise some vectors for downstream cloning purposes and to perform site directed mutagenesis and library mutagenesis (section 2.1.1.10).

Target DNA sequence is represented in conventional PCR by one DNA template. However, in a PCR variant, overlap extension PCR (OE-PCR), 2 different DNA templates with overlapping fragments are included in the PCR reaction. OE-PCR was used as one of the molecular cloning methods in this thesis (section 2.1.1.11).

Another PCR variant, Touchdown PCR, which avoids the amplification of non-specific sequences, was used when conventional PCR failed to amplify the template DNA efficiently. In touchdown PCR, initial higher annealing temperature is used, then a gradual lowering of temperature to a permissive annealing temperature over thermal cycles leads to more specific annealing and hence to more efficient amplification of the target sequence.

Each PCR mixture includes a DNA template, a pair of primers, a DNA polymerase, blend of deoxynucleotide triphosphates (dNTPs), and a reaction-appropriate buffer. A Peltier block thermocycler, Primus 25 (MWG Biotech Inc), was used to perform PCRs.

Q5 High-Fidelity DNA Polymerase (NEB) was used for cloning and mutagenesis purposes as it has a high fidelity, low error rate, whilst *Taq* DNA Polymerase (NEB) was used for the preparation of uracil-DNA substrates, colony PCR as a construct verification tool, and for amplification from uracil-DNA genomes. Tables 2.1.1.1a and 2.1.1.1b summarize *Taq* and Q5 polymerase standard PCR procedures, respectively.

Table 2.1.1.1a. Components and conditions of *Taq* standard PCR protocol. Initial denaturation and final extension steps were used to increase the full-length copies of amplified target DNA.

Taq PCR – Reaction volume: 50 µl - quantities can be halved to make a 25 µl reaction						
Component	Vol (µl)	Final conc	Thermocycling programme			
10X Taq Reaction Buffer	5	1 X	Step		Time (s)	Temp (°C)
Fwd primer (10 µM)	1	0.2 µM	Initial denaturation		30	95
Rev primer (10 µM)	1	0.2 µM	30 cycles	Denaturation	20	95
dNTPs* (10 mM)	1	200 µM		Annealing	30	54-62
Template DNA	Variable	<1,000 ng		Extension	30 per Kb	68
Taq DNA polymerase	0.25	0.025 U/µl	Final extension		300	68
Nuclease-free ddH2o	to 50	N/A	Hold		infinite	4

* Normally dNTPs include equal molarity of dCTPs, dGTPs, dATPs, and dTTPs; dUTPs were used instead of dTTPs in uracil-DNA substrates preparation protocol.

Table 2.1.1.1b. Components and conditions of Q5 standard PCR protocol

Q5 PCR - Reaction volume: 50 µl - quantities can be halved to make a 25 µl reaction						
Component	Vol (µl)	Final conc	Thermocycler programme			
5 x Q5 buffer (NEB)	10	1 X	Step		Time (s)	Temp (°C)
Fwd primer (10 µM)	2.5	0.5 µM	Initial denaturation		30	98
Rev primer (10 µM)	2.5	0.5 µM	18 - 23 cycles	Denaturation	10	98
dNTPs (10 mM)	1	200 µM		Annealing	30	56-64
Template DNA	Variable	<1,000 ng		Extension	30 per Kb	72
Q5 DNA polymerase	0.25	0.02 U/µl	Final extension		120	72
Nuclease-free ddH2o	to 50	N/A	Hold		infinite	4

2.1.1.2. PCR primer design

PCR primer pairs were designed manually and were based upon the template DNA sequence to be amplified. Primers thus included a template-DNA homologous region that has a theoretical melting temperature (Tm) of 56-64°C. Tm was calculated via the generally useful approximation arrived at by the formula: [Tm= 2 (nA+nT) + 4 (nG+nC)].

The primer template-homology region contained a 3' G/C clamp. When required: Restriction sites, mutagenesis sites, or vector-overlap sequences were appended 5' of homologous regions.

2.1.1.3. Primer phosphorylation

The PCR variations iPCR and OE-PCR normally include a blunt-end ligation downstream of amplicon purification; primers are normally provided in unphosphorylated form at the 5' end unless otherwise requested. To this end, iPCR/OE-PCR primers were phosphorylated prior to performing PCR using a T4 Polynucleotide Kinase (T4 PNK; NEB) which catalyses the exchange of inorganic phosphate (P_i) from ATP to the 5' hydroxyl terminus of single- and double-stranded DNA. Phosphorylation was performed according to manufacturer's instructions using T4 Ligase reaction buffer (NEB).

2.1.1.4. Codon optimisation for synthetic genes

Coding sequences for candidate Ung inhibitory genes were optimised for recombinant expression in *E. coli* via manual adjustment of codon usage following initial analysis using the *E. coli* Codon Usage Analyzer 2.1 tool by Morris Maduro[101]. Any sequential codons considered sub-optimal were replaced with silent alternatives to limit local concentrations of sub-optimal codons from the 11th codon onwards (as previous reports showed that optimal codons are not heavily used in the first ten codons of even highly expressed genes[102]), to no more than one in any 5-codon window and no more than two in total per 105-nucleotides of sequence. Subsequent minimal silent manual adjustment to remove or insert restriction enzyme recognition sites was performed using NEBcutter V2.0[103]. Whilst maintaining earlier editing aims, further minimal manual silent adjustments were made to limit homo-polynucleotide runs and predicted mRNA hairpins to <7 nucleotides in length; these were performed guided by output from the UNAFold Web Server[104]. Synthetic desiccated gBlocks^{TM} Gene Fragments (IDT) were resuspended in autoclaved reverse osmosed deionised water to a concentration of 1 ng/µL.

2.1.1.5. Restriction enzyme digestion

All restriction endonucleases, and buffers were provided by New England Biolabs (NEB) Inc. Manufacturer's instructions were followed to perform restriction reactions, unless otherwise stated.

Restriction digestion is a reaction that is used in molecular cloning; additionally, it can be used as a diagnostic tool to verify constructs. Restriction-targeted DNA comes from gel-purified PCR products (section 2.1.1.8) and/or from purified plasmids (section 2.1.1.13). Digestion with multiple enzymes was performed via a single reaction when the buffers were compatible. CIP (Alkaline Phosphatase, Calf Intestinal; source: NEB) was added to vector digestion reactions for downstream cloning purposes. CIP dephosphorylates the 5' ends of DNA, thereby preventing self-ligation of digested vector upon downstream addition of DNA ligase. Digested products were run on agarose gels and purified/visualized for downstream ligation/construct restriction map analysis. Components of restriction digestion reactions are listed in Table 2.1.1.5.

Table 2.1.1.5. Double restriction-digestion reaction. Components and/or their quantities may differ slightly depending on the type of restriction-digestion reaction.

Component	Restriction reaction type		
	Construct verification	Vector digestion	Insert digestion
Purified DNA	500 ng	1000 ng	1000 ng
10X compatible buffer	2.5 µl	5 µl	5 µl
1st Enzyme	0.5 unit	1 unit	1 unit
2nd Enzyme	0.5 unit	1 unit	1 unit
CIP	-	0.5 µl	-
H_2O	to 25 µl	to 50 µl	to 50 µl

2.1.1.6. DNA precipitation

Alcohol precipitation was used to purify DNA between reactions requiring different buffers. DNA sample volume was adjusted to 250 µL by adding 10 mM Tris-Cl buffer at pH 8.5. A volume of 300 µL of 88% isopropanol, 0.2 M potassium acetate was added, then mixed sample was left for 10 minutes before centrifugation at 16100 x*g* for 5 minutes. Supernatant was poured away gently, tubes were recentrifuged at 16100 x*g* for 30 seconds then the remnant was removed out by pipetting. Pellet was dried at 50°C water bath for 2 minutes, then resuspended in 30 µL deionised water and placed in a 50°C water bath for 2 minutes then placed on ice for 3 minutes before storage or use in downstream experiments.

2.1.1.7. DNA separation via agarose gel electrophoresis

Agarose gel electrophoresis was used to resolve DNA fragments depending on their length in base pairs. DNA molecular weight markers (ladders) were used to estimate the size of DNA. Both 1 kb and 100 bp ladders were either generated in the lab from pPSU1 and pPSU2 plasmids[105], or were commercially sourced (NEB). For fragments shorter than 2.5 kb, 1% (w/v) agarose gels were used; while 0.8% (w/v) agarose gels were used for fragments greater than

2.5 kb. Gels utilised in this work were formed in 1xTAE buffer (40 mM tris base, 20 mM glacial acetic acid, 1 mM EDTA, pH 8.0) and stained using SybrSafe stain (Invitrogen); DNA samples were loaded using 6x purple dye (NEB). For DNA to be excised and purified from gels for later applications, visualisation was performed using a blue light illuminator (Clare Chemical). Terminal DNA visualisation in gels, for documentation/reference purposes, was performed in an ultraviolet light box camera system (BioDoc-It imaging system, UVP).

2.1.1.8. DNA extraction from agarose gels

DNA was extracted from gels using QIAquick Gel Extraction Kit (QIAGEN) according to manufacturer's instructions. In outline, excised gel fragments are solubilized in a high-salt buffer via incubation in a water bath at 50°C, and loaded onto a DNA-binding silica-membrane in a spin-column; applying buffers with high concentration of salts leads to DNA adsorption to silica-membrane while contaminants flow through the silica-membrane upon centrifugation. Washing with 80% ethanol buffers efficiently remove impurities, purified DNA is then eluted using deionized water. To increase the yield of eluted DNA, pre-warmed water at 50°C was used for elution. Eluted DNA was collected in 1.5 ml sterile (autoclaved) tubes and stored at -20°C.

2.1.1.9. Ligation

T4 DNA ligase (NEB) was used to perform sticky-/blunt-end ligations. Sticky-end ligation of restriction-digested vectors was performed using 200 ng of the digested vector and, by default a 1:3 vector to insert molar ratio. The volume of ligation reaction mixtures was adjusted to 10 μl either by adding nuclease-free water or by concentrating using a centrifugal vacuum evaporator. Blunt-end ligation of OE-PCR products was performed using 10-50 ng of purified

PCR product and adjusting the reaction volume to 30 µl. To achieve intramolecular blunt-end ligation the reaction volume is greater than that used in sticky-end ligation to favour intramolecular ligation.

DNA was added to the reaction mixture including 1x T4 DNA ligase buffer and 1 µl T4 DNA Ligase. Incubation time and temperature were 2 hours at 30°C for blunt-end ligations and 16-18 hours at 16°C for sticky-end ligations.

2.1.1.10. Library mutagenesis

Inverse PCR of the target construct was performed with oligonucleotides containing designed mutation propensities, phosphorylated prior to use. PCR reactions were performed using Q5® High-Fidelity DNA Polymerase, using 20 cycles. Amplicons were purified from 0.8% agarose following electrophoresis and adjusted to 30 µL at 2 ng/µL for circularisation (section 2.1.1.9). NEB® 5-alpha cells were transformed to propagate the ligation reaction (section 2.1.1.12.5).

2.1.1.11. Molecular cloning

Synthetic DNA of the target genes and oligonucleotides for PCR were provided by Integrated DNA Technologies Inc (IDT). Genes amplified via PCR were cloned into the target plasmid either by using restriction-ligation or by adding a vector-overlap to the genes and apply overlap extension PCR (OE-PCR) method (Figure 2.1.1.11).

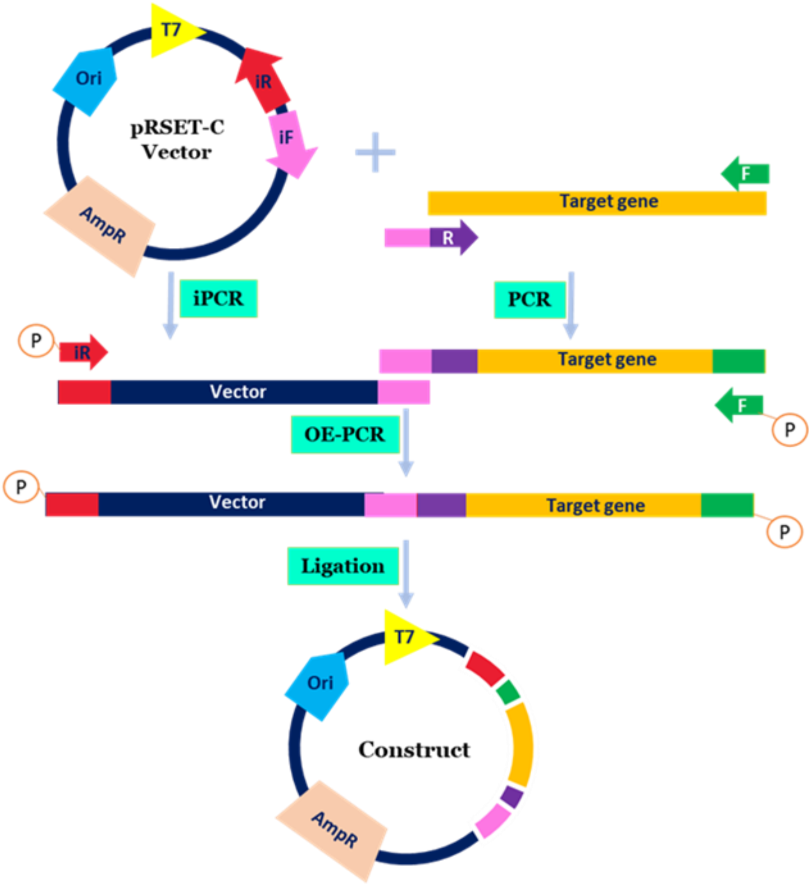


Figure 2.1.1.11. Schematic representation of molecular cloning workflow via OE-PCR. The target vector (pRSET-C as an example) is linearised by applying inverse PCR (iPCR) using inverse forward (iF) and inverse reverse (iR) primers. The target gene is amplified using forward primer (F) and reverse primer (R) that adds a vector-overlap (coloured pink) to the gene. OE-PCR is then performed using both the linearised vector and the amplified gene as DNA templates and utilizing phosphorylated iR and F primers. The extended linear DNA is then circularised by performing intramolecular blunt-end ligation reaction to form the required construct.

A T7 expression plasmid (pRSET-C; Life Technologies) was used to produce monocistronic protein expression constructs carrying the genes encoding potential Ung inhibitors. pRSET-C vector has a T7 RNA polymerase promoter and carries an AmpR gene that confers resistance to ampicillin via expression of beta-lactamase.

For OE-PCR cloning, a linear amplicon of pRSET-C was created by iPCR using the primers P1 and P2 (Table 2.1.1.11), and gel purified from 1% agarose subsequent to electrophoresis. A linear precursor of the construct was formed via overlap extension PCR using 2 ng of each of the plasmid/cassette purified DNA molecules. Reverse primers used for amplifying genes of interest incorporated overlap complementarity with P2 primer. Primer P1 and forward primer of gene amplification were utilised for OE-PCR and were pre-phosphorylated (section 2.1.1.3). Resulting linear amplicons were ligated (section 2.1.1.9).

Table 2.1.1.11. Primers used to linearise pRSET-C vector for downstream cloning via OE-PCR

Primer symbol	Primer name	Primer sequence
P1	LBA2delst2iR	CATATGTATATCTCCTTCTTAAAG
P2	dsblntpRS_iF	TAAGCTTGATCCGGCTGCTAAC

2.1.1.12 DNA microbial propagation

The uptake of the DNA of interest and the subsequent growth and maintenance of the cells containing that DNA is the link between the gene of interest and the isolation of its protein product. This section discusses the strains of *E. coli* and the growth media that were used in this project along with the transformation method used for these bacterial strains in order to uptake the DNA of interest.

2.1.1.12.1. Bacterial strains

For molecular cloning, NEB® 5-alpha competent *E. coli* cells were either provided by NEB or made in-house. This strain of *E. coli* cells permits transformation of unmethylated circular DNA efficiently; the elimination of nonspecific endonuclease I (*endA1*) activity in this strain provides plasmid preparations with high quality.

For expression of proteins, T7 Express *lysY/I*^{q} Competent *E. coli* cells (NEB) were used. This strain, a BL21 *E. coli* derivative for T7 expression, allows the protein expression of genes under T7 promoter with highest level of expression control.

For the *in vivo* lethal UDG assay, CJ236 Electrocompetent Cells were supplied by Lucigen. CJ236 is an *E. coli* strain that lack the activity of Ung and dUTPase genes (*ung* and *dut*), and hence has a uracilated DNA which is degraded into smaller fragments if an uninhibited *ung* gene is transformed to this strain. The genotypes of 5-alpha, T7 Express *lysY/I*^{q}, and CJ236 bacterial strains are summarized in Table 2.1.1.12.1.

Table 2.1.1.12.1. The genotypes of bacterial strains used in this study

Bacterial strain	Genotype
NEB® 5-alpha	*fhuA2 Δ(argF-lacZ)U169 phoA glnV44 Φ80 Δ(lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17*
T7 Express *lysY/I*^{q}	MiniF *lysY lacI*^{q}(Cam^R) */ fhuA2 lacZ::T7 gene1 [lon] ompT gal sulA11 R(mcr-73::miniTn10*--Tet^S*)2 [dcm] R(zgb-210::Tn10*--Tet^S*) endA1 Δ(mcrC-mrr) 114::IS10*
CJ236	[F’ Tra+ Pil+ (CamR)] *ung-1 relA1 dut-1 thi-1 spoT1 mcrA*

2.1.1.12.2. Liquid growth media

Lysogeny broth (LB) and all other broth components were obtained from Sigma-Aldrich. Sterilization of liquid-media was performed by autoclaving at 121°C and 15 psi for 15 minutes. A stock solution of α-D-glucose at 20% (w/v) was separately autoclaved. Antibiotics, and Isopropyl β-D-1-thiogalactopyranoside (IPTG) were prepared as stocks in autoclaved deionised water, then 0.2 µm syringe-filter sterilised (Sartorius). Appropriate sterile antibiotics and/or sterile α-D-glucose were added aseptically after cooling autoclaved liquid media to below 50°C. Components of liquid microbial growth media used in this study are given in Table 2.1.1.12.2.

Table 2.1.1.12.2. Components of broths used for bacterial growth

LB	SOB*	SOC**	2xTY
10 g/l Tryptone	20 g/l Tryptone	20 g/l Tryptone	16 g/l Tryptone
5 g/l yeast extract	5 g/l yeast extract	5 g/l yeast extract	10 g/l yeast extract
10 g/l NaCl	10 mM NaCl	10 mM NaCl	100 mM NaCl
	2.5 mM KCl	2.5 mM KCl	
	10 mM MgCl2	10 mM MgCl2	
	10 mM MgSO4	10 mM MgSO4	
		20 mM glucose	

*Super Optimal Broth - **Super Optimal broth with Catabolite repression

2.1.1.12.3. Solid growth media

LB-agar plates were prepared using Miller's LB-agar supplied by Sigma-Aldrich. For preparation of 10 LB-agar plates, 180 ml of deionized water was used to dissolve 6 g of LB-agar before autoclaving the solution at 121°C and 15 psi. Upon cooling to 50°C, 20 ml of 20% (w/v) sterile α-D-glucose stock solution and 200 µL of appropriate sterile 1000x antibiotic stock solution were added aseptically, with vigorous swirling, to autoclaved LB-agar. A volume of approximately 15-20 ml was poured into each sterile petri dish. Plates were allowed to cool at room temperature then stored at 4°C for later use within up to 14 days of preparation.

2.1.1.12.4. Preparation of chemically competent cells

Mix and Go! *E. coli* Transformation Kit (Zymo Research Corp.), was used to prepare chemically competent cells. Manufacturer's instructions were followed, with one exception: a 5 ml LB-grown overnight *E. coli* culture was used to inoculate SOB medium instead of ZymoBroth^{TM}. The last step of protocol was modified to improve the competence of cells[106]: pre-chilled 1.5 ml sterile (autoclaved) microcentrifuge tubes were used to collect 100 µL aliquots of cell suspension, these tubes were kept on ice for 4-8 hours before storing at -80°C for later use.

2.1.1.12.5. Chemical transformation

Frozen aliquots of *E. coli* chemically competent cells were allowed to thaw on ice for 10 minutes. Super-coiled plasmid DNA (10 ng) or ligated DNA (half of ligation reaction) was added to *E. coli* chemically competent cells. Tubes were flicked gently 4-5 times and incubated on ice for 20-30 minutes. Cells were heat-shocked in a water bath at 42°C for 35 seconds then were transferred back to ice for 5 minutes. A 200 µL volume of pre-warmed SOC media at 37°C was added to the cells, then cells were incubated at 37°C with shaking at 220 rpm for 1 hour. Sterile spreaders were used to spread 200 µL of cell culture onto a pre-warmed LB-agar plate (Section 2.1.1.12.3) at 37°C. Plates were then incubated for 12-16 hours at 37°C.

2.1.1.13. Plasmid DNA isolation from *E. coli* cultures

Recombinant *E. coli* cells were grown overnight in 5 ml Lysogeny Broth (LB) media (Section 2.1.1.12.2) containing appropriate antibiotics in a sterile 30 ml universal tube. Cells were incubated at 37°C with in an innova®42 shaking incubator (New Brunswick Scientific). Overnight (12-16h) cultures were harvested with centrifugation for 15 minutes at 3200 x*g*. The GenElute Plasmid Miniprep Kit (Sigma-Aldrich) was used to isolate plasmid DNA according to manufacturer's instructions.

2.1.1.14. Construct verification

Colony PCR was used as a preliminary tool for verification of a correct construct after ligation and transformation. Colonies were picked using sterile loops; each colony was streaked onto an LB agar plate (Section 2.1.1.12.3) then the loop was transferred into 10 µL deionized water to use 1 µL as a DNA template in PCR. Streaked plates were incubated at 37°C for 8 hours to use later for overnight culture inoculation of verified colonies. Primer pairs for colony PCR were designed to bind either upstream and downstream of the insert, or else upstream (or downstream) and also within the insert.

Restriction digestion was used to further verify the construct. Restriction enzymes used for verification were selected to give a distinct banding pattern (subsequent to agarose gel electrophoresis) that varies significantly from self-ligated vectors. Plasmids with positive results in both colony PCR and restriction map analysis were further scrutinised using Sanger DNA sequencing (Section 2.1.1.15).

2.1.1.15. DNA sequencing

Fluorescent DNA Sequencing was performed by Eurofins Genomics (Germany) using the Sanger sequencing method. In principle, *in vitro* DNA replication is performed, fluorescing dye labelled dideoxynucleotide triphosphates (ddNTPs) are selectively incorporated by DNA polymerase to terminate the chain. These ddNTPs emit light at specific wavelengths. Chain-termination happens at different lengths depending on the incorporated ddNTP. Capillary gel electrophoresis is used to separate the products; then depending on the size of fluorescent products, the original DNA sequence is generated. Sequencing data were analysed using the online editor Benchling (Benchling Inc).

2.1.2. Protein manipulation

In this section, protein manipulation methods including protein expression, cell lysis and soluble protein content isolation, protein content analysis, protein purification, and protein-DNA assays are discussed. The protein manipulation aimed to get a pure sample that is suitable for crystallisation and/or to test the Ung inhibition ability of potential variant UngIns.

2.1.2.1. Recombinant protein expression

To express proteins recombinantly, a target gene must be cloned into an expression vector that includes host-specific appropriate elements for plasmid propagation, selection, and recombinant expression. One of the most well-established widely used hosts for protein expression is *E. coli* which is the cell-factory of choice to express proteins whose correct folding does not require complex post-translational modifications[107]. The T7 expression system, in which a T7 RNA polymerase/promoter system is used, has huge advantages over relying on *E. coli* RNA polymerase. T7 RNA polymerase initiates transcription with high selectivity at its specific promoter sequence (a sequence unrecognisable by *E. coli* RNA polymerase), synthesises RNA at a higher rate, terminates transcription less frequently, and is resistant to some antibiotics that inhibit *E. coli* RNA polymerase (e.g., rifampicin) and could enable the exclusive expression of genes under the control of a T7 promoter by adding such antibiotics[108].

Several bacterial strains are genetically modified to encode T7 RNA polymerase under the control of the UV5 operon. Transcription of T7 RNA polymerase is inhibited in these strains by the lac repressor that binds to the UV5 operator. T7 protein expression is inducible by the addition of IPTG, a structural non-metabolizable analogue of allolactose, that binds the lac repressor and releases it from the UV5 operator, allowing transcription of T7 RNA polymerase which initiates the transcription of genes under the control of a T7 promoter (Figure 2.1.2.1). Specific tags can be designed to be inserted after the start codon and/or before the termination codon of the gene of interest to aid in purification downstream of protein expression[109].

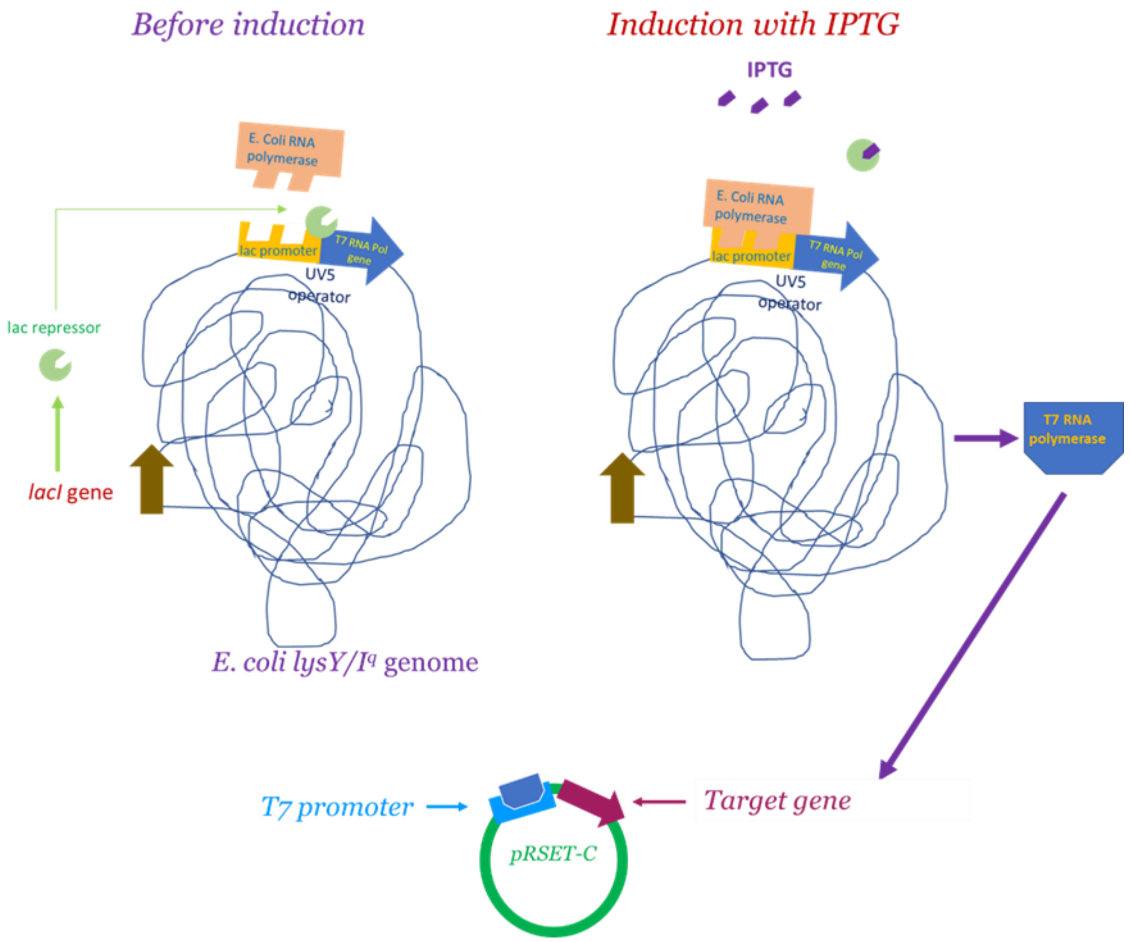


Figure 2.1.2.1. T7 expression system. A genetically modified *E. coli* strain encoding T7 RNA polymerase (e.g., *E. coli LysY/I*^{q}) is used. Before induction, lac repressor blocks the transcription of T7 RNA polymerase by binding to the UV5 operator and preventing *E. coli* RNA polymerase to synthesise T7 RNA polymerase. Induction with IPTG leads to the release of lac repressor from UV5 operator and hence allows E. coli RNA polymerase to initiate synthesis of T7 RNA polymerase which in turns initiates transcription at its own promoter (presented to *E. coli* cells via a T7 express vector such as pRSET-C, controlling the target gene expression). As a result, a high yield expression of the target protein, expressed under the control of T7 promoter, is achieved.

T7 Express *lysY/I*^{q} competent *E. coli* cells were transformed with the relevant expression vector. To perform small-scale expression, a single colony from each transformation plate was transferred aseptically to sterile 5 ml LB media containing 100 µg/ml Ampicillin and 2% (w/v) α-D-glucose in a sterile 30 ml universal tube. Inoculated tubes were incubated at 37°C with shaking at 220 rpm until a cell density measured at OD_{600} of 0.6-0.8 was reached (cell density was measured using a cuvette spectrophotometer; BioPhotometer 6131, source: Eppendorf). At this point, a 30 µL pre-induction sample was saved for SDS-PAGE analysis. Tubes were removed to a water-ice bath for 2 minutes. To induce recombinant expression, IPTG was added to cultures at a final concentration of 0.5 mM. Tubes were then swirled and placed in a 37°C water bath for 2 minutes, prior to their return to a shaker incubator at the required temperature for expression. Induced cells were allowed to grow for 16 hours at the required temperature for expression. Cell cultures were weighed in order to be resuspended in the correct volume of lysis buffer (Section 2.1.2.2); cells were then harvested by centrifugation at 4°C for 20 minutes at 4500 x*g* and stored at -20°C for later use.

Large-scale expression was performed in 2 L capacity plain Erlenmeyer flasks. 500 ml of LB was added to each flask before autoclaving at 121°C and 15 psi for 30 minutes. Ampicillin at 100 µg/ml and α-D-glucose at 2% (w/v) were added to the autoclaved flasks. Cells of overnight 5 ml cultures were harvested and resuspended in 5 ml fresh sterile (autoclaved) LB media before adding them at 1/100 (v/v) ratio to the flasks.

2.1.2.2. Cell lysis

At 5 ml scale, post-induction cultures were harvested as described (Section 2.1.2.1). Pellets were resuspended in a volume of lysis buffer equal to 1/8 OD_{600} for each gram of cell culture. A 30 µl sample was saved and labelled "post-induction" for SDS-PAGE analysis (Section 2.1.2.3). Sonication of a 0.5 ml resuspended sample was performed via a Sonics Vibra-cell ultrasonic processor for 2 minutes at 60 W amplitude with 3 second on/off pulses. The resulting lysate was centrifuged at 18000 x*g* for 30 minutes at 4°C, and the supernatant decanted for storage at -20°C to later visualise the soluble fraction of the cell in SDS-PAGE, or at 4°C if required for biochemical assay. The residual pellets were washed twice with 0.5 ml deionised water, the tube was pulsed-down and water was discarded each time. Pellets were then resuspended again in 0.5 ml of STE buffer (50 mM NaCl, 20 mM Tris, 0.5 mM EDTA, pH 8.0), sonicated, and stored at -20°C to later visualise the insoluble fraction of the cell in SDS-PAGE.

At 1 L scale, post-induction pellets were weighed and resuspended in lysis buffer at 1/5 (w/v) ratio. The sample was lysed by sonication using a Sonics Vibra-cell ultrasonic processor for 4 minutes at 60 W amplitude with 3 second on/off pulses. Sonication was repeated for three rounds leaving the sample for 5 minutes on ice between rounds. A sample of 30 µl was stored at -20°C to later visualise the whole cell lysate in SDS-PAGE. Lysates were then centrifuged at 45000 x*g* for 60 minutes at 4°C, and 30 µl of the supernatant was decanted for storage at -20°C to later visualise the soluble fraction of the cell in SDS-PAGE.

2.1.2.3. Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE)

SDS-PAGE was used to separate and visualise proteins in collected samples. Gels and running buffers were either provided by Thermo Fisher Scientific or prepared in-house. Samples were mixed with loading buffers and heated at 95°C for 5-10 minutes to achieve protein denaturation, then placed at room temperature for 5 minutes before loading onto SDS-PAGE gels.

For commercial gels, the proprietary MES running buffer (recipe not available) was used with 4-12% gradient Bis-Tris Plus gels. Gels were run at 200 V constant voltage for 23 minutes.

For in-house prepared gels, the Mini-PROTEAN III SDS-PAGE system (BIO-RAD) was used. Gels included a resolving region, which was topped with a stacking region. Composition of resolving and stacking gels, and the running buffer is given in Table 2.1.2.3. All components were obtained from Sigma-Aldrich. Gels were run at 100 V constant voltage for 20 minutes, or until loading dye was observed at the interface of stacking and resolving gels; then gels were run for additional 50 minutes at 180 V constant voltage, or until the loading dye was observed to reach the gel bottom edge.

All SDS-PAGE gels were stained for 10 minutes with Instant Blue staining solution (Expedeon), which incorporates Coomassie brilliant blue dye. Gels were then de-stained in double distilled water for up to 16 hours before inspection of protein content.

Table 2.1.2.3. Components of glycine SDS-PAGE gels and running buffer.

Resolving gel – 12% (5 ml)		**Stacking gel (4 ml)**		**Running buffer**
ddH_2O	1.64 ml	ddH_2O	2.32 ml	25 mM Tris
Acrylamide	2 ml	Acrylamide	0.52 ml	192 mM glycine
1.5 M Tris-HCl (pH 8.8)	1.25 ml	0.5 M Tris-HCl (pH 6.8)	1 ml	0.1% (w/v) SDS
10% SDS	50 µl	10% SDS	100 µl	
10% APS	50 µl	10% APS	40 µl	
TEMED	14 µl	TEMED	12 µl	

2.1.2.4. Purification from small-scale expressions

For purification from a 5 ml scale growth, step-wise fractionation of the soluble fraction from cell lysates (Section 2.1.2.2) was performed in centrifugal filter units of 5.0 µm pore-size filled with 400 µL of Q-Sepharose™ Fast Flow (Amersham Biosciences). Ion exchange chromatography and elution of proteins applying step-wise incremental salt concentration was preferred at small-scale purifications in this thesis over affinity chromatography and elution of proteins using step-wise incremental imidazole concentrations as all of the proteins purified at small scale were markedly acidic and were not histidine-tagged proteins.

The filter unit resin bed was equilibrated using 4 ml of a low salt concentration buffer (50 mM NaCl, 20 mM Tris, pH 8.0). The soluble fraction of lysed cells was loaded onto the equilibrated Q-Sepharose, then the filter units were gently flicked-inverted for 5-6 times before letting the resin bed set for 2 minutes. The filter units were spinned down at 300 x*g* for 2 minutes. Elution proceeded via stepwise application of buffers containing incremental amounts of salt (Buffers 1 to 11; Table 2.1.2.4). In each step, 250 µL of buffer was added; note that buffers 1 and 11 were each, respectively, added three steps in succession. Filter units were centrifuged at 300 x*g* at each step for 2 minutes. Eluted fractions were analysed by SDS-PAGE to identify fractions with >90% protein purity.

Table 2.1.2.4. Partial fractionation buffers used for small-scale growth purification.

Buffer	Composition
1	50 mM NaCl, 20 mM Tris, pH 8.0
2	100 mM NaCl, 20 mM Tris, pH 8.0
3	150 mM NaCl, 20 mM Tris, pH 8.0
4	200 mM NaCl, 20 mM Tris, pH 8.0
5	250 mM NaCl, 20 mM Tris, pH 8.0
6	300 mM NaCl, 20 mM Tris, pH 8.0
7	350 mM NaCl, 20 mM Tris, pH 8.0
8	500 mM NaCl, 20 mM Tris, pH 8.0
9	750 mM NaCl, 20 mM Tris, pH 8.0
10	1000 mM NaCl, 20 mM Tris, pH 8.0
11	1300 mM NaCl, 20 mM Tris, pH 8.0

2.1.2.5. Protein purification from large-scale expressions

The expressed protein must be separated from other host expressed proteins before downstream experiments. To obtain a highly pure sample that is suitable for protein crystallography, multiple steps of chromatographic purifications are usually required to minimize the presence of contaminants.

For tagged proteins, tag affinity chromatography is the first step performed in fast protein liquid chromatography (FPLC) due its high specificity. This technique involves binding the tagged protein to a stationary-phase with high affinity for a specific tag, then removing the non-bound soluble molecules then applying a material with higher affinity to the stationary phase to elute the target tagged protein. A hexa-histidine tag is a widely used tag to purify proteins via immobilized metal affinity chromatography (IMAC, Figure 2.1.2.5). Different protein properties can be used for other chromatographic purifications; these properties include the charge (ion exchange chromatography), the size (size exclusion chromatography), and the hydrophobicity (hydrophobic interaction chromatography).

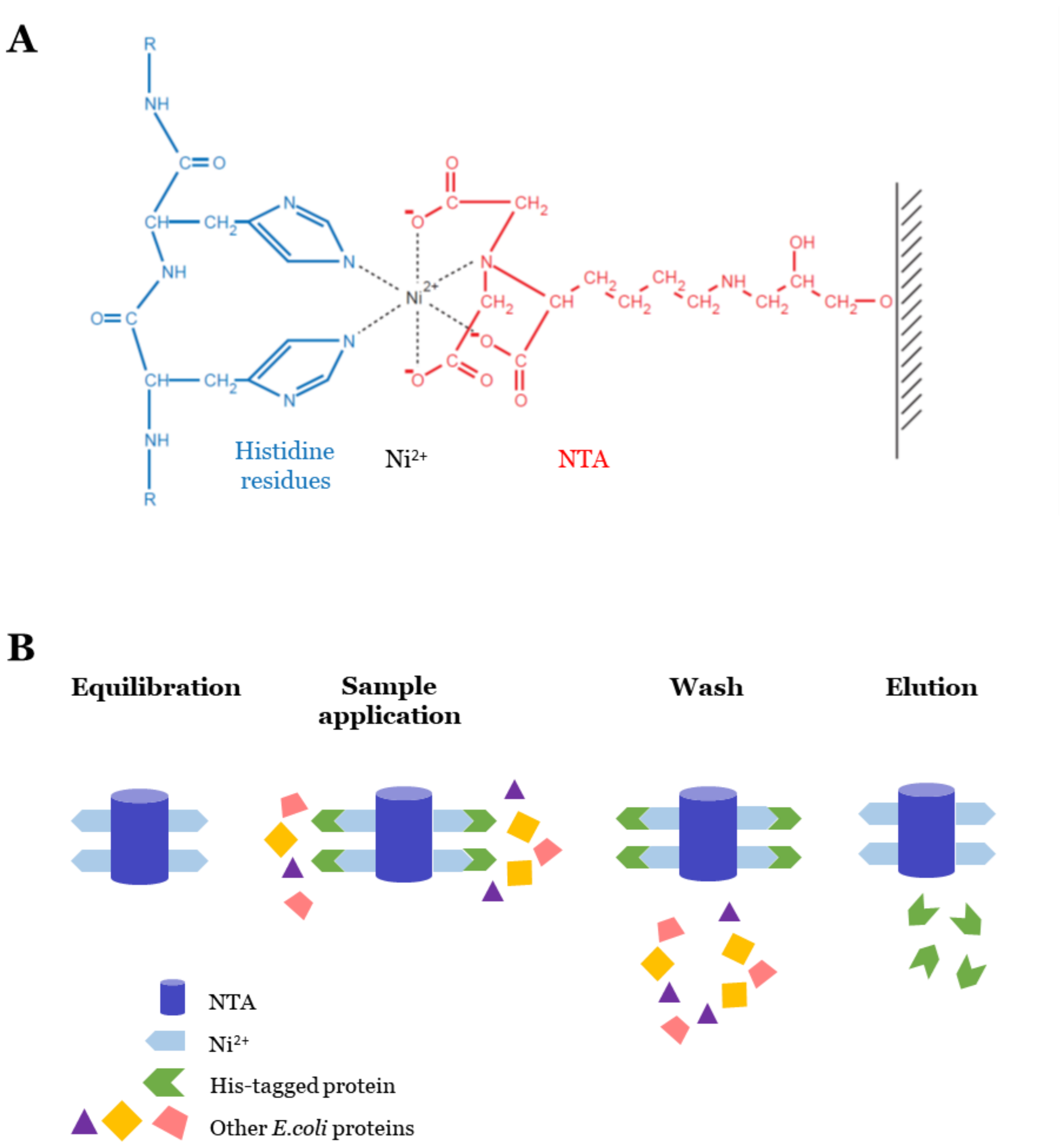


Figure 2.1.2.5. Principles of Immobilized-metal affinity chromatography (IMAC). (A) Interaction between neighbouring residues in the 6xHis tag and Ni-NTA matrix; Nitrilotriacetic acid (NTA) occupies four of the six ligand binding sites of the nickel ion (Ni2+), leaving two sites free to interact with 2 histidine residues. (B) Schematic of IMAC workflow: The His-tagged protein selectively binds to the stationary-phase (Ni-NTA), unbound proteins can be removed by applying a washing step, the 6xHis-tag protein is eluted with high concentration of imidazole which displaces the His-tag from nickel ions.

The details of the utilised purification buffers are listed in Table 2.1.2.5. Cell paste obtained from large-scale expressions (Section 2.1.2.1) was resuspended in Buffer A at 1/5 (w/v) ratio. Cells were lysed by sonication as described (Section 2.1.2.2). Purification was performed via sequential chromatographic steps: immobilized metal affinity chromatography (IMAC; using 1ml HisTrap HP, GE), ion exchange chromatography (IEC; using 1 ml HiTrap Q HP, GE) and size exclusion chromatography (SEC; using 120 ml Superdex 75, GE). The AKTA fast protein liquid chromatography (FPLC) system was used for all chromatography steps. Fractions corresponding to peaks in chromatography steps were analysed by SDS-PAGE and fractions that contained protein content with the expected size were included in next chromatography step.

Table 2.1.2.5. Purification buffers. All buffers were filtered prior to use using papers with 0.45 µm pore-size

Buffer	Composition
A	300 mM NaCl, 20 mM Tris, 20 mM Imidazole, 0.5 mM EDTA pH 8.0
B	300 mM NaCl, 20 mM Tris, 20 mM Imidazole, pH 8.0
C	300 mM NaCl, 20 mM Tris, 500 mM Imidazole, pH 8.0
D	50 mM NaCl, 20 mM Tris, pH 8.0
E	1 M NaCl, 20 mM Tris, pH 8.0
F	20 mM Tris, pH 8.0
G	200 mM NaCl, 20 mM Tris, pH 8.0

2.1.2.5.1 Immobilized metal affinity chromatography

In IMAC, 1 ml HisTrap HP column was equilibrated with 20 column volumes of buffer B. The resuspended sample in buffer A was loaded on the equilibrated column. The column was then washed with 20 column volumes of buffer B. Samples were eluted by applying a 30-column volume linear gradient of buffer B to buffer C. The eluted fractions corresponding to absorbances significantly above baseline were collected and diluted with buffer F to a 50 mM NaCl concentration.

2.1.2.5.2. Ion exchange chromatography

In IEC, the column was equilibrated with 20 column volumes of buffer D before loading the IMAC diluted sample (Section 2.1.2.5.1). The column was then washed with 20 column volumes of buffer D. Samples were eluted by applying a 30-column volume linear gradient of buffers D to E. The eluted fractions corresponding to absorbances significantly above baseline were collected and concentrated at 3200 x*g* to a final volume of 1 ml using an Ultracel 3K centrifugal filter (Millipore).

2.1.2.5.3. Size exclusion chromatography

The concentrated sample obtained from IEC (Section 2.1.2.5.2) was applied to 120 ml Superdex 75 column previously washed with 1 column volume of double distilled water followed by equilibration with 3 column volumes of buffer G. Eluted fractions containing the pure target protein were concentrated at 3200 x*g* using an Ultracel 3K centrifugal filter (Millipore) to reach the required protein concentration.

2.1.2.6. Protein concentration measurements

DS-11 Spectrophotometer (DeNovix) was used to measure protein concentrations depending on sample absorbance at 280 nm wavelength. A volume of 2 µL of each sample was applied onto the test plate per measurement. ProtParam tool on the ExPaSy webserver[110] was used to estimate both molecular mass and extinction coefficient (ec_{280}) of protein. Measured concentration (mg/ml) was divided by ec_{280} to calculate the corrected protein concentration (mg/ml). To assess protein samples contamination with DNA, a ratio of absorbance at 260 nm: 280 nm (A260/A280, given with protein concentration) was considered. Any value of A260/A280 smaller than 1.0 indicates no detectable nucleic acid contamination.

2.1.2.7. Protein-DNA assays

2.1.2.7.1. *In vitro* UDG activity/inhibition assay

To monitor Ung activity and its inhibition, a semiquantitative general UDG agarose gel assay was performed[58], in the presence of gel purified 600bp *Taq* polymerase PCR amplicons (deoxynucleotide pools were, respectively, ACGU in the substrate PCR and ACGT in the control PCR). A serial dilution in STE buffer was used to discover the minimum Ung enzyme concentration required to unambiguously process 5 µL of uracil-DNA substrate relative to the control. Inhibitory profiles of putative Ung-inhibitory proteins were investigated using 3 µL of partially purified small-scale expressed proteins (section 2.1.2.4) or the soluble lysate fraction of respective recombinants harvested 12-16 hours after induction. Previously purified Ung inhibitor proteins were used as controls: Ugi from *Bacillus subtilis* phage PBS1[111], SAUGI[62], and bacteriophage PZA p56[58]. Cell lysates from plasmid-free T7 Express *lysY/I^q* cells, and from the *dut$^-$/ung$^-$* deficient *E. coli* strain CJ236, were used as cell lysate controls.

2.1.2.7.2. Protein-based analysis of Ugi mutants

The construct pBUGI8, built previously[111], and its library mutagenesis products were expressed at small-scale. Pellets from 250 µL aliquots of cell culture were resuspended in 16 µL 1x UDG buffer (NEB) containing lysozyme (50 µg mL^{-1}), RNase (40 µg mL^{-1}), and 1x Protease Inhibitor Cocktail Tablet solution (Roche). Resuspended pellets were then processed with three freeze-thawing cycles at -80°C (3 min each) and 37°C (30 sec each). Streptomycin sulphate (Sigma) was added from a 10% (w/v) stock to a final concentration of 1% (w/v), with mixing by repeated gentle inversion and tubes were further incubated on ice for 30 minutes. Tubes were then centrifuged at 8,000 x*g* for 2 minutes at 4°C, 1.8 µL of the supernatant was extracted and diluted tenfold in UDG buffer supplemented with 100 mM EDTA and 1 unit of *E. coli* UDG. After allowing UDG-lysate interaction at room temperature for 3 minutes, 1 ng µL^{-1} of thymine-DNA or 2 ng µL^{-1} of uracil-DNA substrate were added to a final volume of

18 µL and incubated at 37°C for 15 minutes. Samples were heated up at 85°C for 10 minutes, then cooled down over 20 cycles to 25°C using the Primus 25 thermocycler (MWG Biotech Inc). Samples were visualised via running on 1% agarose gel.

2.1.2.7.3. Selection-based analysis of UngIn functionality

A *dut*$^{-}$ *ung*$^{-}$ *E. coli* strain, CJ236, was supplied by Lucigen. This strain has a uracilated genome, if competent CJ236 cells are transformed with plasmids encoding *ung* gene, Ung-induced genomic fragmentation results in cell death unless Ung activity is inhibited. This is the basis of rapid plate cellular survival Ung inhibition assay that was developed in this thesis.

Constructs encoding Ugi, Ung, both Ugi and Ung, or both non-functional Ugi mutant and Ung were developed and used to control the assay. Control transformations were performed using: (1) vectors carrying Ung inhibitory gene to produce colonies, (2) vectors carrying both Ung gene and Ung inhibitory gene to produce compromised colonies; and a vector carrying uninhibited Ung gene resulting in no colonies. This assay was used to test potential UngIn variants and to analyse Ugi library mutagenesis products.

Potential UngIn genes were cloned into constructs carrying Ung gene to test the Ung inhibition ability of those potential UngIns via transformation of the plasmid DNA of the verified formed dual constructs into CJ236. Plasmid DNA from any survival colonies was sequenced to verify UngIn and Ung sequences.

To analyse Ugi mutants, Ugi library mutagenesis (section 2.1.1.10) of Ugi-Ung carrying constructs was performed. Circularised products were transformed into NEB® 5-alpha cells. An agar plate, containing a well dispersed lawn of transformed NEB® 5-alpha colonies, was treated by spreading 1.5 mL of LB media over the colonies and pipetting the resuspension into a clean tube. The pellet was retained following centrifugation at 8,000 x*g* for 2 minutes, and plasmid DNA was isolated using the PureLink™ Quick Plasmid Miniprep Kit (Invitrogen). CJ236 aliquots were transformed with 50 ng of the plasmid DNA. Resulting colonies were grown separately in liquid LB media and plasmid DNA was subjected to SupremeRun Sanger sequencing services from GATC were used to obtain data on functional Ugi mutants (Eurofins Genomics).

2.1.3. Protein X-ray crystallography

2.1.3.1. Brief theory of X-ray crystallography

The brief theory of protein X-ray crystallography given in this part of thesis is derived from textbooks by Bernhard Rupp and Gale Rhodes[112,113].

2.1.3.1.1 Protein crystallisation

A protein crystal represents an array of protein molecules ordered in identical orientation. X-ray crystallography determines the atomic arrangement in a protein molecule. Crystals consist of repeating units, the simplest of which is called the unit cell. A three-dimensional lattice is formed by translationally arranged unit cells. The asymmetric unit is the smallest integer volume of the macromolecule; applying symmetry operators to the asymmetric unit can be performed to obtain reconstruction of the entire unit cell. Typically, 30-80% of crystals content is formed by solvent; protein molecules (forming 20-70% of crystal content) are held together in a crystal by different types of non-covalent bonds including: electrostatic and hydrophobic interactions, hydrogen bonds, and salt bridges.

In protein crystallisation trials, homogenous soluble protein sample is required; sample solubility is gradually decreased by subjecting it to vapour-diffusion conditions using the hanging drop or sitting drop format. Dehydration of protein sample is performed by mixing protein sample with a reservoir solution (a mother liquor) that includes a precipitant.

Precipitants can be organic precipitants, such as polyethylene glycols (PEGs), or salts; precipitant makes hydrogen bonds with water molecules and by that competes with protein molecules for water molecules. Dehydration induces protein molecules to interact with each other in order to neutralize charges on their surfaces.

A drop of the mixture of protein and mother liquor is added to a chamber enclosing a larger reservoir filled with mother liquor. The mixture drop could be suspended above the reservoir on a cover slip (hanging drop), or sat on a chamber next to the large reservoir (sitting drop) as presented in Figure 2.1.3.1.1a.

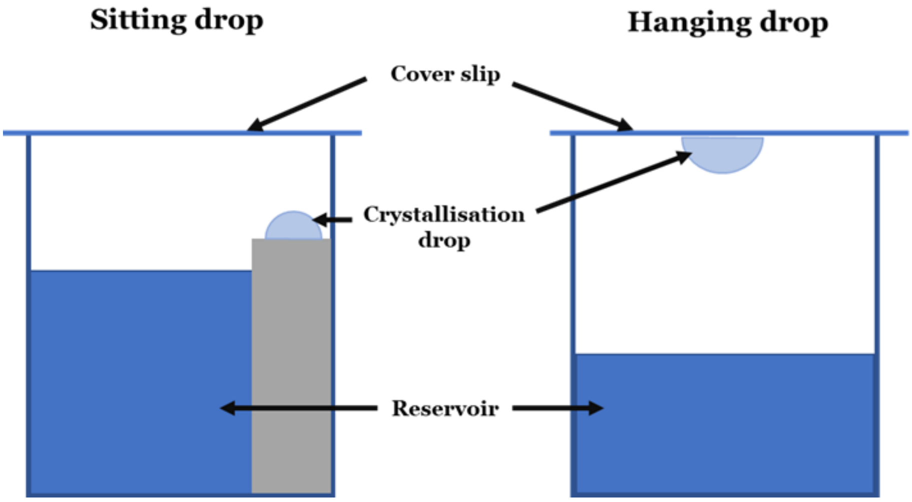


Figure 2.1.3.1.1a. Common vapour-diffusion crystallisation techniques. In the sitting drop method, the protein sample is pipetted into a drop containing the reservoir on a platform then a well containing both the reservoir and the sitting drop is sealed with a cover slip. In the hanging drop method, the crystallisation drop is set directly on a cover slip that is then inverted above the reservoir in a tightly sealed well.

In vapour-diffusion, water molecules evaporate from the drop and the mother liquor until equilibration is achieved. In order to equilibrate this system, a net water transfer from the drop to the large reservoir is achieved until equal concentration of the major solute present in the system, the precipitant, is reached in both the drop and the reservoir. A supersaturated state of the protein is reached, at which the protein becomes metastable (Figure 2.1.3.1.1b). At this point, the supersaturated system is not thermodynamically equilibrated due to kinetic barriers. High concentration of supersaturated solution leads to frequent collisions between protein molecules; hence, protein molecules can frequently interact with each other and form small soluble aggregates. These aggregates can act as nuclei to generate a nucleation event that is necessary for the activation barrier between a supersaturated system and biphasic system to be overcome. In the nucleation event, protein molecules attach to a soluble aggregate to result in a growing nucleus before forming a biphasic system via precipitation of protein. The drop splits into a protein-rich phase and a saturated solution phase, and the system reaches equilibrium. The protein-rich phase could take the form of an amorphous precipitate; however, at optimal conditions, a crystal can form in an ordered manner.

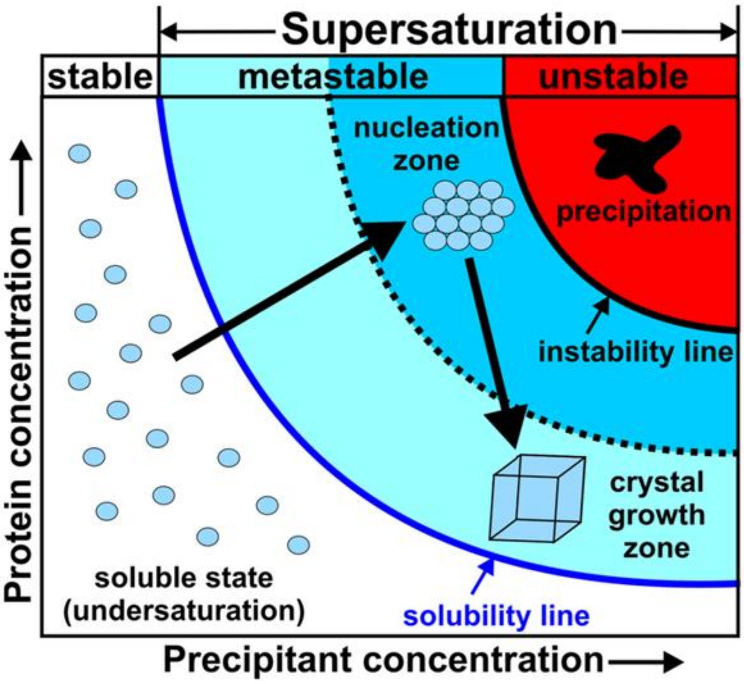


Figure 2.1.3.1.1b. Representation of protein crystallisation phase diagram. The protein solution in the soluble state (undersaturation zone, white background) is stable and comprises a single phase. The solution in supersaturation zone is either metastable or unstable. In the metastable region nucleation events can occur, and if supersaturation proceeds (arrow from soluble state zone to nucleation zone), the nuclei reach the critical size and become stable. As the size of nuclei increases, crystals start to form. At this stage, protein concentration decreases and the crystal growth zone is reached (arrow starting from nucleation zone), where crystals continue to grow. At higher concentrations of protein or precipitant, the unstable zone is reached and amorphous protein precipitation occurs (adapted from Bijelic & Rompel, 2018[114]).

In practice, optimal protein-specific crystallisation conditions are unpredictable. A sparse matrix high-throughput approach is used typically to sample a wide variety of previously successful crystallisation conditions to get insights into promising hits that show a crystalline material or a phase separation state. Hits are then optimized by inspecting around the condition pH, protein concentration, and precipitant concentration aiming to get optimal crystals. Typically, multiple variant screens are used, with protein to mother liquor ratios of 1:1 and 1:2 (v/v) for each condition, duplicate trays of these screens are usually incubated at different temperatures, commonly at 4°C and at 16°C.

2.1.3.1.2. X-ray diffraction

An X-ray diffraction experiment involves irradiating a crystal with an X-ray beam and measuring the pattern of X-ray diffraction. An X-ray diffraction is generated by arrays of atoms due to the fact that X-ray wavelength (0.5-1.5 Å) is comparable to spacing between atomic planes. Scattered X-rays by atoms of a crystal interfere in phase or out of phase with one another to form constructive or destructive interreference events, respectively. Constructive events are achieved when the difference of path length between two waves satisfies Bragg's law (equation A.1 and Figure 2.1.3.1.2). To observe constructive interference events for atoms in different unit cells, the Bragg planes must penetrate identical positions of every unit cell of the crystal. In a diffraction experiment, 2D diffraction images are collected at angles differing by 0.1°-1° in order to get complete data about unique reflections of a crystal at each angle.

$$n\lambda = 2d\sin\theta \qquad \text{(A.1)}$$

Bragg's equation: λ is the wavelength of X-rays, d is the distance between atomic planes, and θ is the scattering angle.

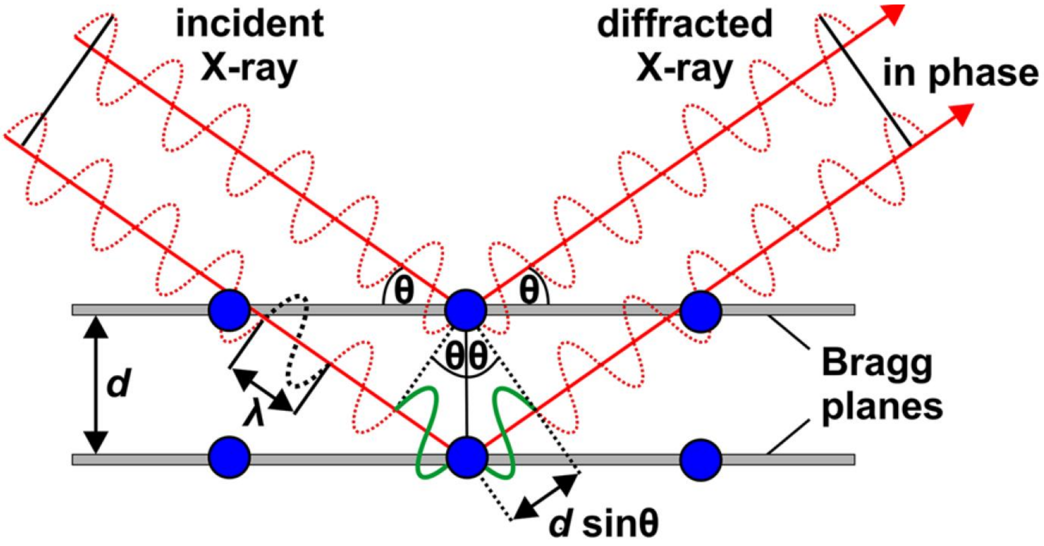


Figure 2.1.3.1.2. Schematic of 2 waves interfering constructively according to Bragg's law. The path length difference, 2dsinθ, is shown in green solid line, which equals an integer number of wave lengths λ (in this case 2dsinθ=2λ) and therefore a constructive diffraction event is achieved (adapted from Bijelic & Rompel, 2018[114]).

2.1.3.1.3. The structure factor, phasing, and refinement

Recorded patterns of diffraction in X-ray crystallography represents the Fourier transform of the protein electron density. The electron density map can be constructed from the recorded diffraction pattern of a crystal. Reflections can be defined by their positions (coordinates called Miller indices h, k, and l), and due to the fact that reflections are generated by waves, they have amplitude, frequency, and phase that characterise them.

The amplitude and phase of a crystal lattice planes diffracted wave are described by the structure factor, a function that represents the Fourier sum of individual atomic structure factors (equation A.2).

$$F_{(hkl)} = \sum_{j=1}^{n} f_j e^{2\pi(hx_j + ky_j + lz_j)} \tag{A.2}$$

The structure factor equation. $F_{(hkl)}$: the structure factor; n: the number of atoms; f_j: the amplitude; x_j, y_j, and z_j are the coordinates of atom *j* in the unit cell; *h*, *k*, and *l* are the Miller indices.

Intensities are detected in X-ray crystallography; however, phase information cannot be measured directly in diffraction experiments as the detector is not able to measure the phase angles (ϕ) of the reflections. Hence, this so-called phase problem in crystallography must be solved to calculate electron density derived from diffraction patterns. Phase estimation of a data set, often called phasing, can be performed using variety of methods including direct methods, experimental phasing, and molecular replacement.

Direct methods employ statistical relationships between sets of structure factors, these relationships become weak when the resolution is lower than 1.2 Å and when the number of atoms increases; therefore, such methods are rarely used for determination of protein structures. Experimental phasing employs isomorphous replacement in which different types of crystals are produced, native crystals and heavy atom derivative crystals. The heavy atoms introduce perturbation to the diffraction pattern, their positions can be identified and used to deduce the phases of the data set. Molecular replacement involves the use of an atomic model of a homologous protein whose structure is already known. Typically, the RSMD between the α-carbon atoms of the used atomic model and the target protein should be <2.0 Å. Molecular replacement involves placing the atomic model into the unit cell of the target protein crystal, then rotating and moving the model to find the best fit position and orientation based on either Patterson-function (used by MOLREP software[115]) or maximum likelihood function (used by PHASER software[116]).

Once a best-fit position and orientation is found, an electron density map is calculated and a target protein can be built to best-fit this map. An initial atomic model of a target protein is built automatically when molecular replacement is used to estimate phases. To minimise the initial atomic model bias, improving diffraction quality of protein crystals might be needed to get a solution at a higher resolution. The initial model then has to be improved via a refinement process. The goal of refinement is to improve phase estimates and hence improve the electron density map to define the atomic model structure more accurately. Refinement can be achieved locally in COOT[117], and next globally using a software such as REFMAC5 that includes restraints such as bonds length and torsion angles[118]. An important factor that measures phase estimate improvement is the R-factor, often referred to as R_{work}. The R-factor reflects the difference between the structure factor amplitudes calculated from the current atomic model and the measured amplitudes observed in the diffraction experiment (equation A.3).

$$R = \frac{\sum ||F_{obs}| - |F_{calc}||}{\sum |F_{obs}|} \tag{A.3}$$

The R-factor equation: R is the R-factor, $|F_{obs}|$ is the structure factor magnitude from diffraction data, $|F_{calc}|$ is structure factor amplitude calculated from model.

Ideally, R_{work} would be zero; however, prior to refinement, a model produced by molecular replacement may in practice have an R_{work} of 0.3-0.45, such R_{work} may improve to 0.15-0.3 in a fully refined model; which has a substantial difference from an R_{work} of 0.63 given by a random set of atoms. Another important measure that guarantees the model is not over-fitted to the data is R_{free}, which is the R-factor calculated from a subset of reflections that were not used neither in phasing nor in the refinement processes, and hence R_{free} is never model-biased. An atomic model is considered not over-fitted when $R_{free} - R_{work} \leq 0.05$.

2.1.3.2. Protein crystallisation

Screening for crystallisation conditions using commercial vapour-diffusion screens utilised a 96-well sitting drop format. Protein samples were automatically dispensed in volumes ranging from 50 nL to 200 nL using a Mosquito Nanolitre Liquid Handler (TTP Labtech). Each reservoir comprised 75 µL of mother liquor; sub-wells included protein and mother liquor in 1:1 or 1:2 (v/v) ratios. Clear plastic film was used to seal each tray before incubating in a room at 16°C where a Minstrel Desktop Crystal Imaging System (Rigaku) captured photos of crystallisation drops at regular time intervals.

For hanging-drop crystallisation, 0.5-1 µL of protein samples were pipetted onto glass cover slips. Commercial and/or in-house made solutions were prepared in advance to be dispensed in wells using volumes ranging from 300 to 500 µL. Well solution was pipetted onto the sample drop to achieve protein to mother liquor ratios of 1:1 and 1:2 (v/v). Cover slips were sealed with grease suspending drops above well solution. Plates were incubated at 16°C.

2.1.3.3. Crystal mounting

Crystal mounting and cooling was done with the help of Dr. Claire Bagneris (Rosalind Franklin Lab manager). Litho loops or nylon loops with approximately same diameter of crystals were employed for the transfer of crystals from the droplets in which they had grown into a cryo-buffer drop. Cryo-buffers contained the mother liquor supplemented with cryo-protectants: 20% (v/v) glycerol or 20% (v/v) ethylene glycol. The loops were then transferred to pucks covered with liquid nitrogen for cryo-cooling and storage until pucks were shipped in dry Dewars to a synchrotron for data collection.

2.1.3.4. Data collection and structure determination

Data collection was done with help of Dr. Claire Bagneris and Dr. Nikos Pinotsis (Crystallography Lab manager). X-ray diffraction data were collected on beamline IO4 at the Diamond light source facility or on beamline ID30B at the European Synchrotron Radiation Facility (ESRF). Data collection was performed using a PILATUS detector[119] with a fine slicing methodology and oscillation of 0.10°, exposure of 0.10 s and a transmission of 75% for 1800 images. Automated beam-line processing was conducted using the XIA2 pipeline[117,120]. Further processing used the programs XDS[121] and DIALS[122], data reduction was performed using AIMLESS[123]. The structures were initially phased by molecular replacement with MOLREP[115]. REFMAC5[118] was used for structure refinement steps, and model building used the program COOT[117].

2.2. Computational methods

The work in this project included using computational tools in order to search for sequence similarities and structural homologs of known UngIns in certain taxonomic families or target organisms. In addition, state-of-the-art structure prediction tools were used to get some structural insights of target proteins. This section discusses all the computational methods used in this project.

2.2.1. Bioinformatics sequence search methods

2.2.1.1. *In silico* identification of potential homologs of Ugi

An online PSI-BLAST search at the NCBI was performed on the non-redundant protein sequences database using the Ugi sequence (Accession: YP_009664501.1) from *Bacillus* phages PBS1/AR9 as a template. Default search parameters were used. Iterations were run until no new sequences below the E-value threshold (0.05) were obtained.

2.2.1.2. *In silico* identification of potential homologs of SAUGI

An online PSI-BLAST search at the NCBI was performed on the non-redundant protein sequences database using the SAUGI sequence (UniProtKB - Q936H5_STAAU) as a template. Default search parameters were used, with maximum target sequences set at 20,000. Iterations were run until no new sequences below the E-value threshold were obtained. After each iteration, all the sequences below the E-value threshold and selected sequences above the E-value threshold (satisfying the condition: query cover >80% and percent identity >25%) were included to build the Position-Specific Scoring Matrix (PSSM) for the next iteration.

Hits generated from this procedure, that were encoded by species other than *Staphylococcus* with >80% query cover and <35% percent identity were further investigated as potential distant homologues of SAUGI.

2.2.1.3. *In silico* identification of potential homologs of p56

An online PHI-BLAST search at the NCBI was performed using the p56 sequence (GenBank: ACE96021.1) encoded by *Bacillus* phage Phi29 as a template and E-X(2)-Y-X(0,2)-G as a PROSITE pattern. The search was iterated until no new sequences were obtained. Hits generated from this procedure, found to be encoded by phages, were further investigated as potential p56 homologues.

Additionally**,** A PSI-BLAST search was performed on the non-redundant protein sequences database, selecting the *Salasmaviridae* organism dropdown (p56 is known to be encoded by *Salasmaviridae* phages). The search was iterated until no new sequences were obtained.

2.2.1.4. Targeting UngIn searches in specific genomes

BLASTP, PSI-BLAST and PHI-BLAST searches were performed to search within phages PhiR1-37 and T5 genomes for Ung inhibitors using each of Ugi (PDB: 1UGI), SAUGI (PDB: 3WDG), and p56 (PDB:4L5N) sequences as queries. The PROSITE pattern E-X-[ILVMF] was used for Ugi/SAUGI PHI-BLAST search, while the PROSITE pattern E-X(2)-Y-X(0,2)-G was used for p56 PHI-BLAST search. Results that showed a plausible sequence alignment, acidic isoelectric point, and a protein length of <150 amino acids were further investigated. For T5 phage, a manual sequence alignment was performed for each of the open reading frames in the fragment of its genome that is reported to encode the Ung inhibition activity with each of the known Ung inhibitors; attractive potential sequence homologs were further investigated.

2.2.1.5. Heuristics analyses of Ugi, SAUGI, and p56 Sequences

Sequences generated from the SAUGI PSI-BLAST search with >80% query cover, and the sequences generated from Ugi PSI-BLAST search were aligned using Clustal Omega[124] or MUSCLE[125] as part of the MEGA X suite[126]. Sequences generated from the p56 PHI-BLAST and PSI-BLAST searches were aligned using Clustal Omega. Sequence attributes for Ugi, SAUGI, and p56 were plotted using the matplotlib package[127], and sequence logos were made using Weblogo 3[128].

2.2.1.6. Phage genome database filtering

Filter scripts were written in the VSCode IDE[129] in Python by Naail Kashif-Khan, a fellow PhD student in the Savva research lab, using tools from the Biopython module[130]. Scripts were deposited on Birkbeck College server and were run on individual genomes or bulk genome lists, downloaded from the NCBI database[131]. Genomes were initially translated into all six reading-frames, with any continuous sequence between two stop codons treated as a putative polypeptide (with a minimum length of 40 amino acids). Any N-terminal sequence before the first valid start codon in each sequence was removed – start codons defined as M, I, V, or L[132]. While there is evidence of additional start codon usage in bacteria[133], these were excluded in the interests of stringency in sequence processing.

Processed sequences were then passed through a set of filters that were set based on the heuristics analysis, with sequences passing each filter check written to an output FASTA file. For bulk genome inputs, additional filtering and binning was applied - genomes were binned based on GC content and screened for additional motifs. Parameters for each filter were optimized such that the known Ugi or p56 sequence was returned in the final output, along with as few false positive sequences as possible.

2.2.2. Structural computational methods

2.2.2.1. Structural homology search for potential Ung inhibitors

The four known types of Ung inhibitors [Ugi, PDB code: 1UDI, chain I; SAUGI, PDB code: 3WDG, chain B; p56, PDB code: 4L5N, chain C; and Vpr: PDB code: 5JK7, chain F] were used as starting queries to search on the Dali server[134] and on PDBeFold server[135] for any structurally homologous proteins. Proteins with loosely similar structures (Z-score>4.0 and RMSD <3.5 Å) were considered for further investigation. Next, additional constraints deduced partially from Ung inhibitors' mutual properties were applied to shorten the list of candidate proteins: (i) a protein length of <200 amino acids; (ii) a theoretical isoelectric point (pI) of < 7.0; (iii) a protein of unidentified function (i.e., annotated: hypothetical protein, uncharacterized protein...etc). A list of candidate proteins that passed all the filtration steps were considered for structural superposition to look for the essential motif structural conservation and surface negative charge distribution pattern.

2.2.2.2. Structure predictions of potential UngIn homologs

Structure predictions for monomers were performed using AlphaFold2[136] and/or RoseTTAFold[137]. Dimer and protein-complex structure predictions were performed using AlphaFold-Multimer[138]. All predictions were run using the Google Colab notebook ColabFold[139]. For each predicted structure, the model with the highest predicted local distance difference test (pLDDT) score was used for structural analysis.

Chapter 3

3. Analyses of UngIn sequence variation

The goal of the work presented in this chapter is to define new UngIn specific sequence space more completely, using a combined library-based directed evolution and bioinformatics approach, involving selection by a conditional lethal assay and validation in biochemical assays. In this chapter, uncovering synthetic Ugi variants and the identification of distant UngIn sequences will be discussed, along with the discussion of inability of some plausible sequence/structural homologs to inhibit Ung.

3.1. UngIn verification assay

3.1.1. Introduction

The *E. coli* strain CJ236 contains genomic mutations that result in a phenotype deficient in Ung and dUTPase activity. Due to the inability of CJ236 cells to prevent accumulation of dUTP in the nucleotide pool (*dut*$^-$) or repair DNA uracil arising by replicative incorporation or spontaneous cytosine deamination (*ung*$^-$) a significant proportion of DNA positions will be occupied by deoxyuridine[140]. If competent cells of CJ236 are transformed with plasmids encoding a functional Ung, bacterial colonies do not survive due to catastrophic Ung-induced genomic DNA fragmentation resulting in cell death. In contrast, if plasmids are introduced from which both an Ung and an Ung Inhibitor (UngIn) are encoded, bacterial colonies will be able to form due to sufficient nullification of Ung activity by the UngIn (Figure 3.1.1). This is the basis of a rapid agar plate assay, from which plasmids of surviving colonies of CJ236 may be DNA sequenced to establish the identities of UngIns encoded by them. This assay was employed to ascertain UngIn functionality among distant candidate UngIn sequence homologs identified from database accessions, and to ascertain the extent of tolerable sequence substitution at various positions in the primary structures of validated UngIns. The main aim is to discover whether sequence plasticity might underly the curious apparent absence of an UngIn in genomes such as *Yersinia* phage PhiR1-37 and to validate whether sequences fulfilling certain heuristic signatures might support UngIn functionality.

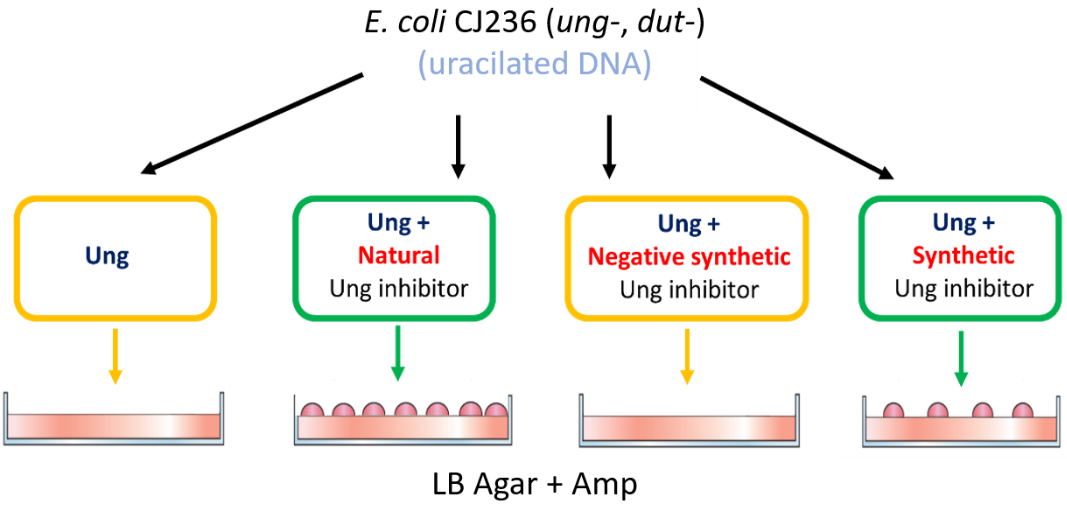


Figure 3.1.1. Schematic representation of the bacterial conditional lethal assay for Ung inhibitory activity. If Ung is transformed into CJ236, cell death will be a result of Ung-induced catastrophic disintegration of the genome due to proximal uracil residues on both strands. However, transforming a natural/synthetic UngIn along with Ung protects the cells of that disintegration and leads to colonies survival.

Multiple constructs were used to control and assess assay robustness: **(i)** pRSET-B-UGI (pBUgi.8), a vector based upon pRSET-B (Invitrogen), pBUgi.8 carries the phage PBS1 Ugi gene under control of a T7 promoter[141]; **(ii)** pRSET-B-U12, developed previously by Mr Daniele Mestriner, formerly an undergraduate project student in the Savva research lab (appendix A), this vector is similar to pBUgi.8, but instead of phage PBS1 Ugi, it carries a non-functional Ugi mutant (known as U12) in which the 1st β-strand, required for docking into the Ung-DNA binding cleft, residues are substituted; **(iii)** pTS106.1: a vector based upon the commercial vector pTrc99A (Pharmacia LKB Biotechnology), pTS106.1 carries the conserved Ung catalytic domain encoding portion of the Herpes Simplex Virus UL2 gene under control of a Trc promoter[142]; **(iv)** pSDM4-Ugi: carries the phage PBS1 Ugi gene under control of a Trc promoter; **(v)** pSDM4_Ugi_Ung: carries both the conserved Ung catalytic domain encoding portion of the UL2 gene and phage PBS1 Ugi gene under the control of one Trc promoter each **(vi)** pSDM4_U12_Ung: carries both conserved Ung catalytic domain encoding portion of the UL2 gene and the U12 Ugi mutant gene under the control of one Trc promoter each. The T7 promoter was used in pBUgi.8 and pRSET-B-U12 constructs in order to control protein expression and assay small-scale expressed Ugi and U12 proteins (appendix A), while the Trc promoter was used for the other constructs to allow a high level of basal transcription and assay candidate UngIns without the need of protein expression induction or basal transcription control.

3.1.2. Vector construction

A high-copy-number vector, designated pBpST-CAT was developed by Dr James Horton, formerly an MRes student in the Savva research lab (appendix A).

The resulting construct, pBpST-CAT, was linearised via a 1-hour HindIII-HF / NdeI double digest at 37°C, in the presence of 0.5-unit CIP alkaline phosphatase, and purified following electrophoresis from 0.8% agarose. The Ugi gene was obtained from the pBUGI8 construct[142], via a 1-hour HindIII-HF / NdeI double digest at 37°C, and purified following electrophoresis from 1% agarose. The plasmid (100ng) and insert were ligated for 2-hours at 25°C in a 1:3 molar ratio using T4 DNA ligase and recombinants were isolated from clonal colonies of transformed NEB® 5-alpha cells. The sequence-verified construct was designated pSDM4-Ugi.

The construct pSDM4-Ugi was linearised, via a 1-hour BspEI digest at 37°C in the presence of 0.5-unit CIP alkaline phosphatase and purified following electrophoresis from 0.8% agarose. The HSV1-UNG gene was obtained from pTS106.1 via PCR using primers P3 and P4 (the sequences of all the oligonucleotides used in this chapter are listed in Table A.1, Appendix A). The amplicon was digested for 1-hour in the presence of BspEI and AgeI at 37°C and purified following electrophoresis from 1% agarose. The plasmid (100ng) and insert were ligated for 2-hours at 25°C in a 1:3 molar ratio and recombinants were isolated from clonal colonies of transformed NEB® 5-alpha cells. The sequence-verified construct was designated pSDM4_Ugi_Ung.

To build construct pSDM4_U12_Ung, the procedure of pSDM4_Ugi_Ung preparation was followed using construct B instead of A (Figure 3.1.2) in the initial NdeI/HindIII-HF digestion.

CJ236 cell viability was tested with transformation of different constructs, these transformations confirmed good growth (>200 colonies per agar plate) with vectors not carrying an Ung gene (pBUgi.8, pSDM4-Ugi, and pRSET-B-U12), compromised growth (50-150 colonies per agar plate) with vectors that carry both an Ung gene and an Ugi gene (pSDM4_Ugi_Ung), and no growth with vectors that carry only an Ung gene (pTS106.1 and pSDM4_U12_Ung). Construct pSDM4_U12_Ung was utilised as the parental vector for constructing new libraries, thus safeguarding the CJ236 assay against false positives that would results from trace parental DNA contamination if pSDM4_Ugi_Ung was used to construct new libraries.

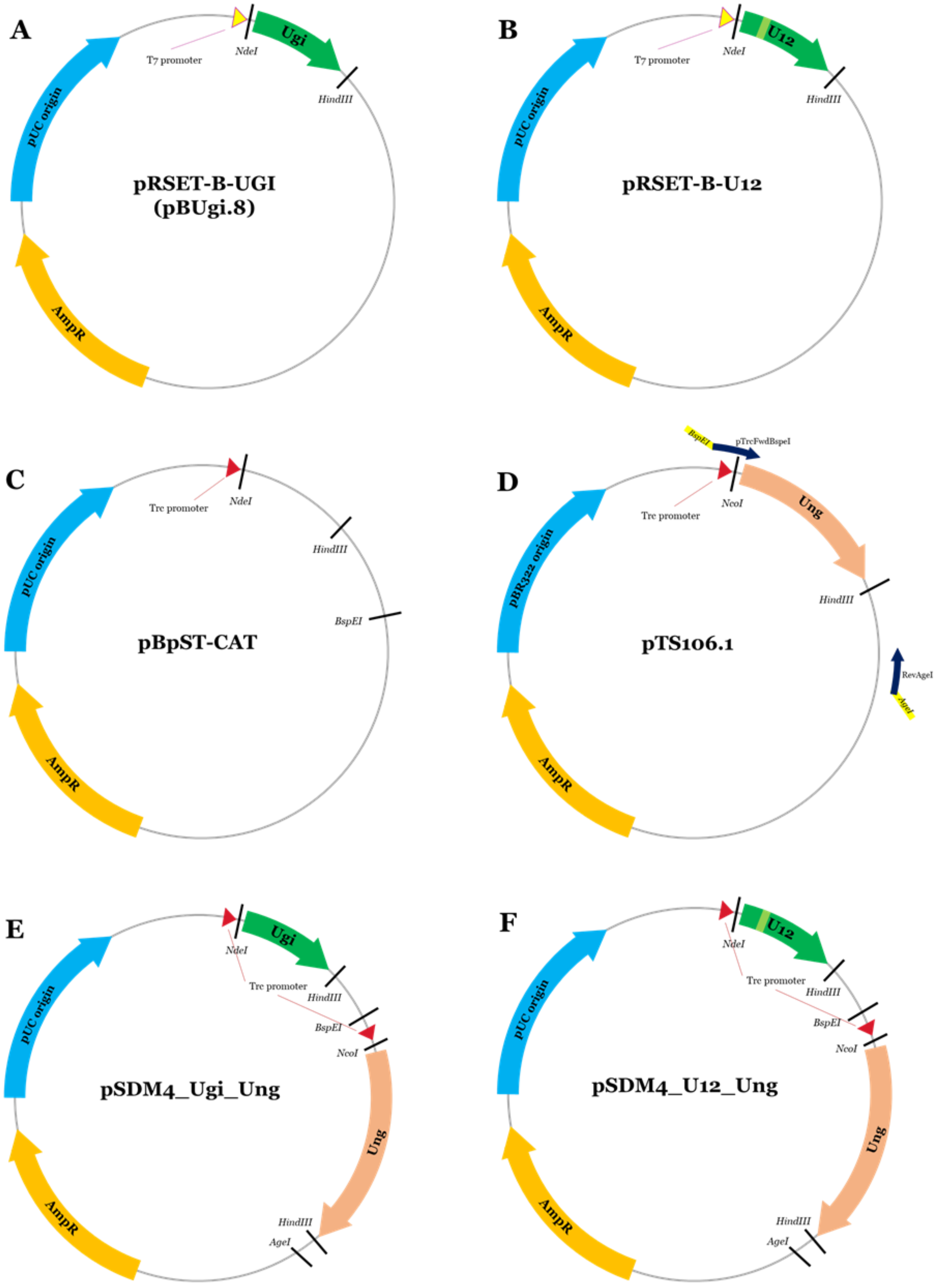


Figure 3.1.2. Constructs developed for the CJ236 bacterial lethal assay. Blue elements represent the origins of replication, orange elements represent ampicillin resistance gene (AmpR), red triangles represent Trc promoters, yellow triangles represent T7 promoters, coral elements represent Ung gene, and green elements represent Ugi or the mutant Ugi variant U12 as labelled.

3.2. Engineering synthetic Ugi variants

Since Ugi and SAUGI share a common fold underpinned by distantly homologous sequences exhibiting high levels of plasticity, a structure-based sequence alignment approach (section B.2.1) was used as the basis for constructing libraries of random substitutions at defined positions in the sequence (Figure 3.2). The deposited crystal structures of HSV1-UNG complexes with PBS1 Ugi[22] and SAUGI[62] were used as the templates for the alignment of these respective UngIns[29].

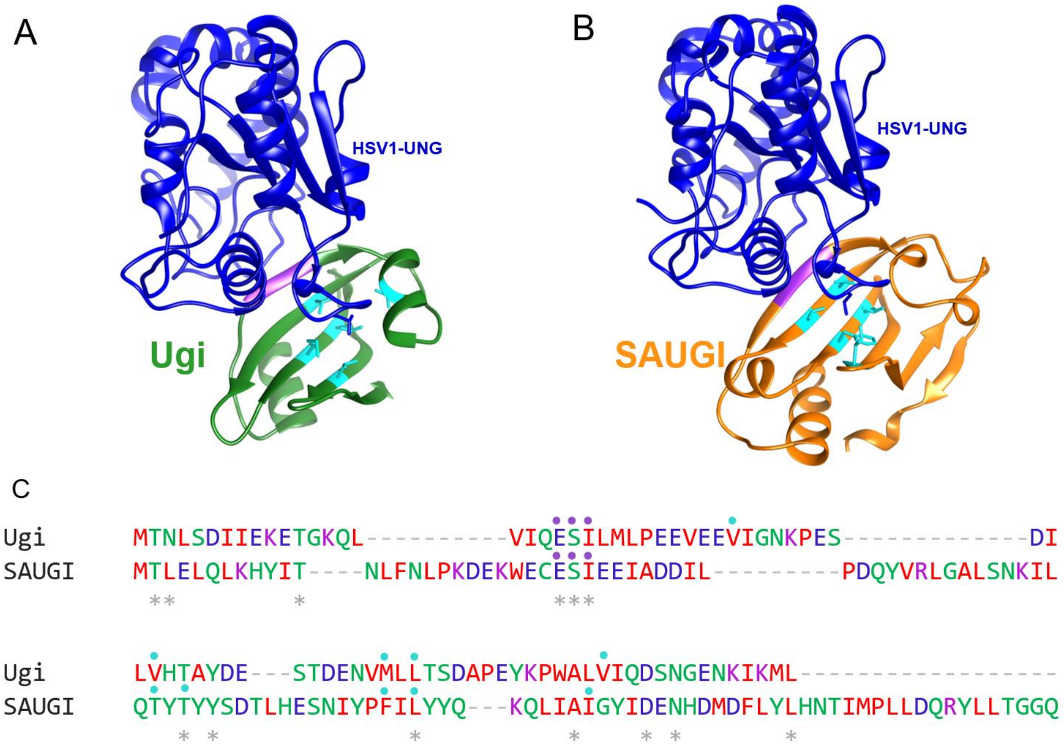

A
HSV1-UNG
Ugi
B
HSV1-UNG
SAUGI
C
Ugi MTNLSDIIEKETGKQL----------VIQESILMLPEEVEEVIGNKPES-------------DI
SAUGI MTLELQLKHYIT----NLFNLPKDEKWECESIEEIADDIL---------PDQYVRLGALSNKIL
** * ***
Ugi LVHTAYDE---STDENVMLLTSDAPEYKPWALVIQDSNGENKIKML------------------
SAUGI QTYTYYSDTLHESNIYPFILYYQ---KQLIAIGYIDENHDMDFLYLHNTIMPLLDQRYLLTGGQ
* * * * * * *

Figure 3.2. Structure-based sequence alignment of Ugi and SAUGI. HSV1-UNG complexes with (A) Ugi (PDB: 1UDI). (B) SAUGI (PDB: 5AYS). (C) Structure-based sequence alignment of Ugi (84 aa) and SAUGI (112 aa). The apical residue [leucine in HSV1-UNG, shown as a stick] of the Ung minor groove DNA intercalation loop is sequestered by UngIn residues coloured cyan in panels A and B and topped with cyan dots in panel C. Although the structures of Ugi and SAUGI share a common fold, their sequences are heterologous. These proteins share only 13 identical residues (indicated by asterisks in panel C); the only conserved motif, ESI, coloured purple in panels A and B topped with purple dots in panel C, is located on the 1st β-strand of each inhibitor, which docks in the Ung-DNA binding cleft.

Interestingly, only one motif comprised of three amino acids (ESI) was found to be identical in Ugi and SAUGI, this motif is found in the 1st β-strand of these inhibitors, which docks in the Ung-DNA binding cleft. In the SAUGI sequence family, the ESI motif is observed to tolerate mutations at positions 2 (Ser) and 3 (Ile), according to the pattern E-[APST]-[FILMV][31]. The 1st β-strand of Ugi and the ESI motif, were respectively targeted for library mutagenesis to screen for tolerated variations.

Construct pSDM4_U12_Ung was used as the starting vector to generate mutant libraries of Ugi via iPCR site directed mutagenesis. Libraries were generated to (1) randomly mutate the residues comprising the Ung-binding β-strand of Ugi (library L1; primers P5 and P6, Table A.1) or (2) to shuffle the ESI motif according to observed variation in SAUGI sequence data (library L2; primers P7 and P8) or entirely randomly at positions 2 and 3 of that motif (library L3; primers P9 and P10). All primers were pre-phosphorylated (section 2.1.1.3), iPCR products were gel-purified (section 2.1.1.8), ligated (section 2.1.1.9), and transformed into NEB® 5-alpha cells (section 2.1.1.12.5). Potential sizes of libraries were calculated based on the different possible amino acid sequence variations in each library; library L1 size is 20 amino acids variation in each of the Ung-binding β-strand 7 residues (i.e. $20^7 = 128 \times 10^7$ sequences), library L2 size is 2 × 4 × 5 = 40 sequences, library L3 size is 2 × 20 × 20 = 800 sequences.

Transformation of NEB® 5-alpha isolated plasmid DNA of library L1 into CJ236 returned no surviving colonies (i.e., no synthetic Ugi sequences with a novel 1st β-strand sequence). Transforming plasmid DNA of libraries L2 and L3 into CJ236, yielded a total of 16 unique nucleotide sequences encoding 11 novel Ugi variants. One other variant encoded the wild-type Ugi sequence, however, via a synonymous encoding nucleotide sequence. Importantly, the remaining 15 synthetic sequences showed 11 novel motif sequences substituting for the wild-type ESI motif without compromising the UngIn functionality of Ugi (Table 3.2).

Table 3.2. Library mutagenesis targeting the Ung-binding 1st β-strand of Ugi

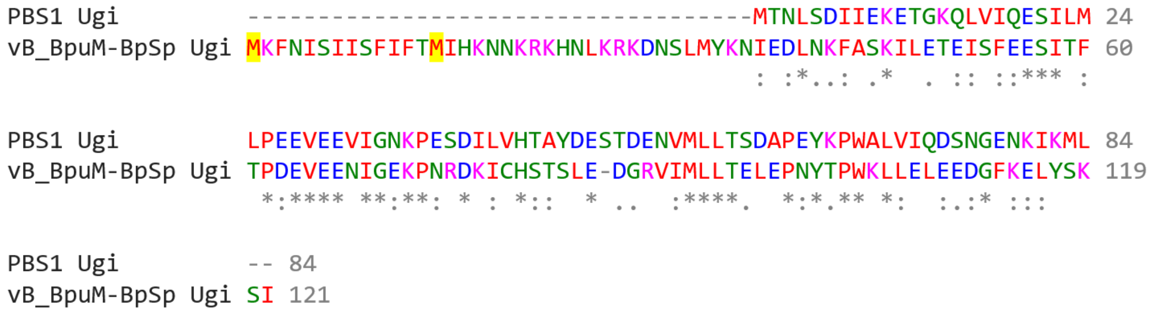
Sequences of Ugi 1st β-strand, and library range	
Ugi:	QESILML
L1:	XXXXXXX
L2:	Q[ED][APST][FILMV]LML
L3:	Q[ED]XXLML
Discovered synthetic variant Ugi sequences	
L1:	No surviving colonies
L2:	QESILML, QEAMLML
L3:	QEALLML, QESTLML, QESVLML, QETCLML, QESWLML, QEAPLML, QETVLML, QEVTLML, QETMLML, QETILML

3.3. Bioinformatics searches for Ugi homologs

3.3.1. PSI-BLAST search for Ugi homologous sequences

Three homologous sequences, annotated as uracil-DNA glycosylase inhibitor, were output from a PSI-BLAST search using a PBS1 Ugi sequence. These sequences included the search sequence (i.e., Ugi from phages PBS1, PBS2, AR9), a close homolog (93% ID) encoded by *Bacillus* phage vB_BspM_Internexus, and a more distant homologous sequence (32% ID) encoded by the *Bacillus* phage vB_BpuM-BpSp.

The homology of *Bacillus* phage vB_BpuM-BpSp encoded Ugi (designated Ugi-2 in this thesis; locus tag: Bp8pS_259) with PBS1 Ugi (84 aa) starts after the 3rd methionine in the annotated sequence of Ugi-2 (121 aa; Figure 3.3.1). Based on Ugi/SAUGI insights, two sequences allowing translation from the 2nd methionine or the 3rd methionine of Ugi-2 annotated sequence were cloned, and these 2 sequences were designated Ugi-2$_{108}$ (108 aa) and Ugi-2$_{89}$ (89 aa), respectively.



```
PBS1 Ugi         ------------------------------------MTNLSDIIEKETGKQLVIQESILM 24
vB_BpuM-BpSp Ugi MKFNISIISFIFTMIHKNNKRKHNLKRKDNSLMYKNIEDLNKFASKILETEISFEESITF 60
                                                     : :*..: .*  . :: ::*** :

PBS1 Ugi         LPEEVEEVIGNKPESDILVHTAYDESTDENVMLLTSDAPEYKPWALVIQDSNGENKIKML 84
vB_BpuM-BpSp Ugi TPDEVEENIGEKPNRDKICHSTSLE-DGRVIMLLTELEPNYTPWKLLELEEDGFKELYSK 119
                  *:**** **:**: * : *::  * ..  :****.  *:*.** *:  :.:* :::

PBS1 Ugi         -- 84
vB_BpuM-BpSp Ugi SI 121
```

Figure 3.3.1. Sequence alignment of PBS1 Ugi and identified Bp8Sp_259 Ugi (Ugi-2). Sequence alignment was performed with default settings at Clustal Omega web server. Methionine residues of Ugi-2 preceding homology with Ugi are highlighted yellow. PBS1 Ugi first residue aligns four residues after the 3rd methionine residue of Ugi-2.

3.3.2. Cloning of Ugi-2 variants

The synthetic gene sequence for Ugi-2 (locus tag: Bp8pS_259) was designed *in silico* (Appendix A) with *E. coli* optimized codon usage[143], using the *E. coli* Codon Usage Analyzer 2.1 tool[101]. Synthetic DNA was obtained (IDT), amplification of Ugi-2_{108} was performed via PCR using Q5 DNA polymerase and the primers P11 and P12. The Ugi-2_{108} amplicon was isolated from 1% agarose and purified (section 2.1.1.8) and was cloned to pRSET-C using OE-PCR/ligation approach (section 2.1.1.11). The construct was designated pRSCUgi-2_{108}. Truncation to the 3rd methionine position was achieved by iPCR, pre-phosphorylated primers (P1 and P13) were used to amplify pRSCUgi-2_{108} to generate (following purification and ligation) the construct pRSCUgi-2_{89}.

3.3.3. Small-scale protein expression and purification of Ugi-2 variants

Protein expression of both variants Ugi-2_{108} and Ugi-2_{89} was induced by the addition 0.5 mM IPTG. Small-scale expression was tested using variable temperatures to decide upon the most suitable conditions (Figure 3.3.3a). Partial purification of the soluble fraction following initial cell lysis was performed according to the step-wise fractionation protocol (section 2.1.2.4). Eluates at 500 mM NaCl (Figure 3.3.3b) contained the peak band of target protein and hence were used in the UDG assay.

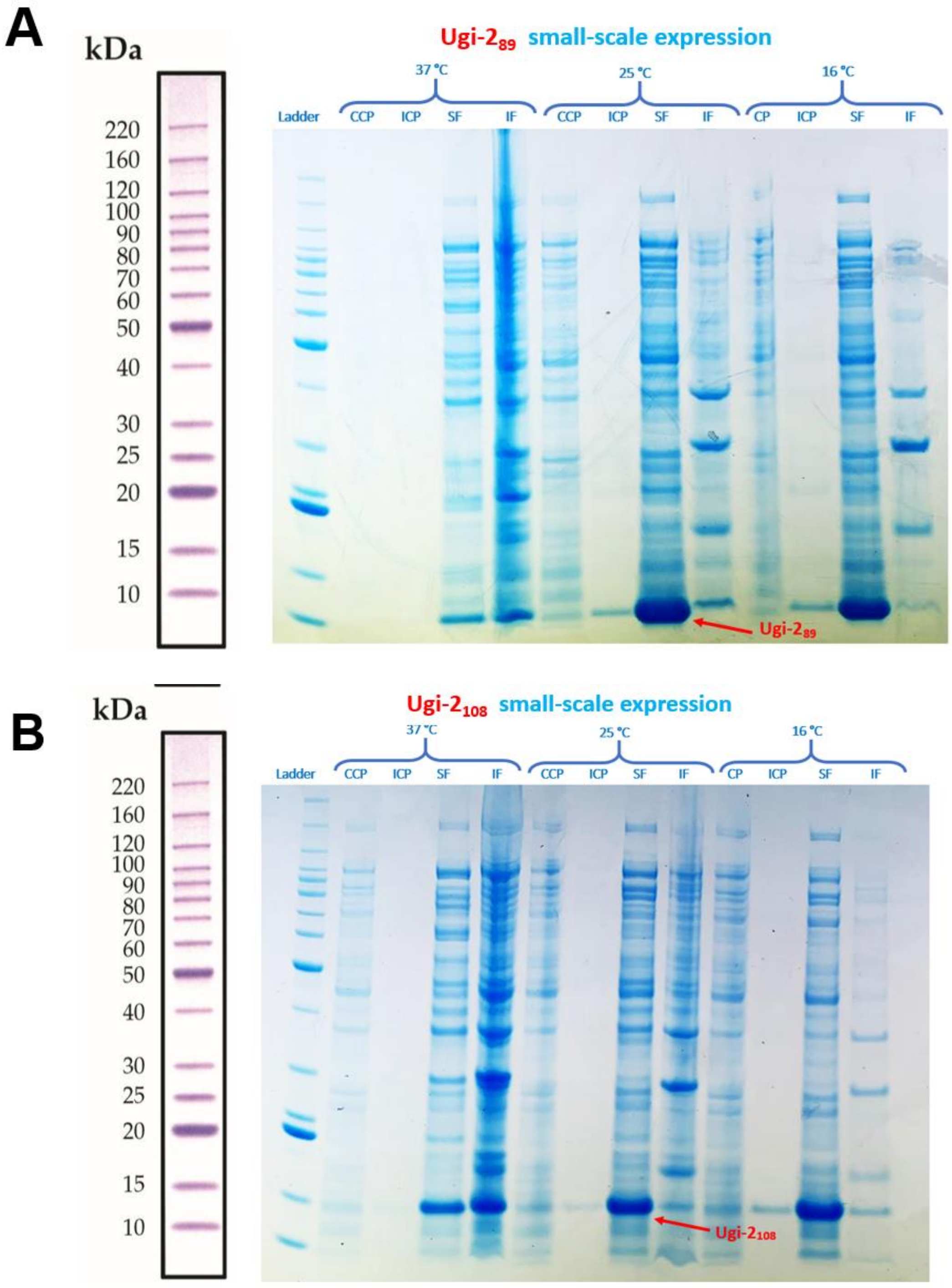


Figure 3.3.3a. SDS-PAGE showing Ugi-2$_{89}$ (A) and Ugi-2$_{108}$ (B) proteins using variable experimental temperatures and induction at 0.5 mM IPTG. The optimal temperature for induced expression is 25 °C. CCP: control cell pellet, ICP: induced cell pellet, SF: Soluble fraction, IF: insoluble fraction.

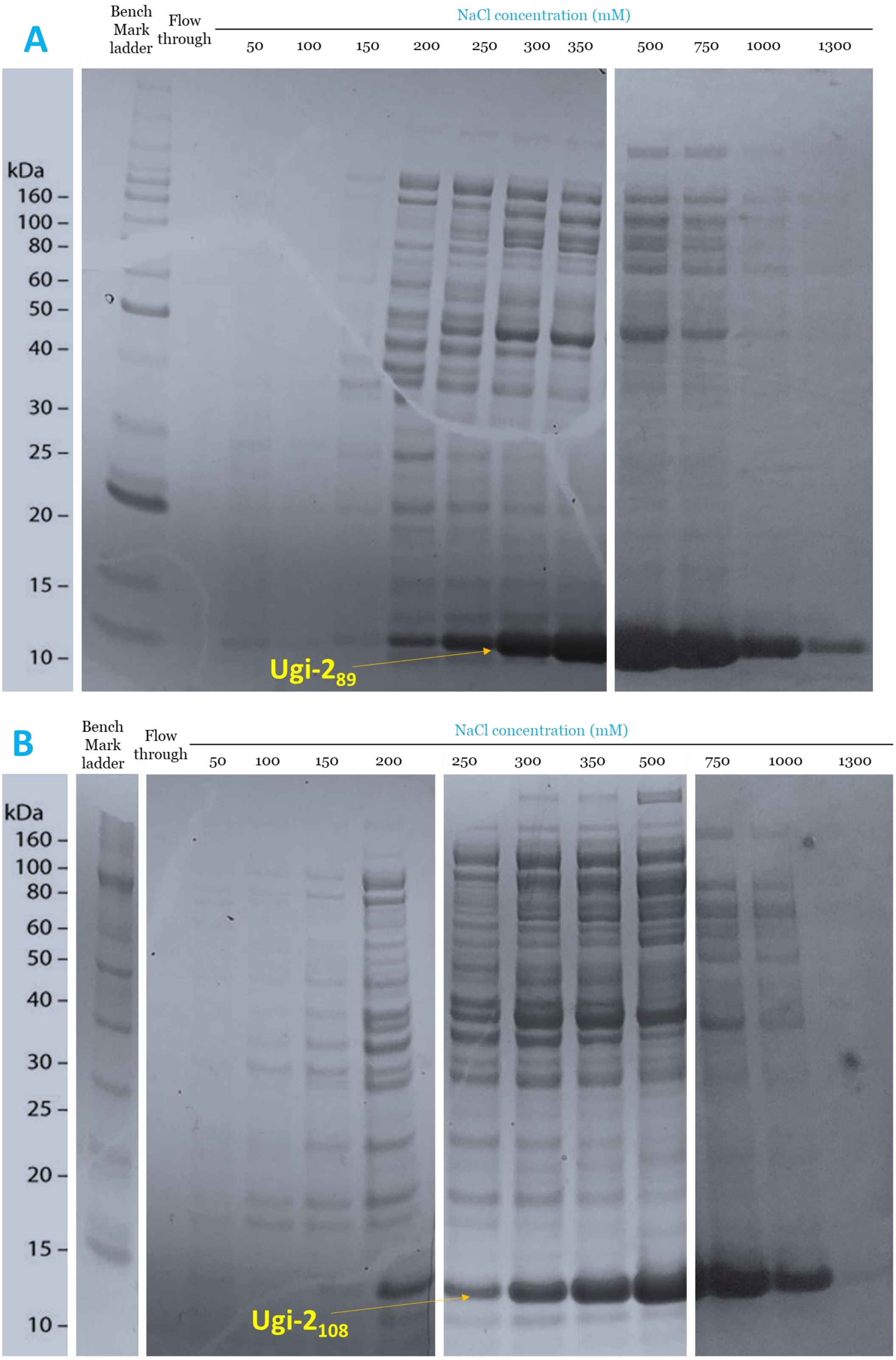


Figure 3.3.3b. SDS-PAGE analysis of Resin-bound Ugi-2$_{89}$ (A) and Ugi-2$_{108}$ (B) filtration on centrifugal filters (Ultrafree - MC) using a gradient of NaCl concentration buffer (Tris 20mM, pH 8). Most of both Ugi-2$_{89}$ and Ugi-2$_{108}$ variants were eluted at 500 mM NaCl concentration. Eluates at 500 mM were chosen for use in the UDG assay. This figure is formed of 3 cropped images. Original uncropped images are available and stored electronically.

3.3.4. UDG inhibition assay of Ugi-2_{89} and Ugi-2_{108}

Partially purified Ugi-2_{89} and Ugi-2_{108} proteins (section 2.1.2.4) were utilised to test the protein ability to inhibit Ung. In vitro UDG assay was used (section 2.1.2.7.1). *Staphylococcus aureus* UNG (SAUNG) activity on DNA substrates was tested when assayed alone or with either PBS1 Ugi or the homologous Ugi-2 samples. PBS1 Ugi, Ugi-2_{89} and Ugi-2_{108} showed ability to inhibit SAUNG activity (Figure 3.3.4b).

The PBS1 Ugi is known to retain its fold and activity when heated to 95 °C for 10 minutes and allowed to cool down to room temperature[142]. The potential thermostability of Ugi-2_{89} and Ugi-2_{108} was therefore tested. Running SDS-PAGE showed that both Ugi-2_{89} and Ugi-2_{108} remained in the soluble fraction after heating to 95 °C for 10 minutes (Figure 3.3.4a). The UDG assay was repeated to test the Ung-inhibition ability of heated proteins. The results showed that PBS1 Ugi and Ugi-2_{89} but not Ugi-2_{108} retain the property of Ung inhibition after heating (Figure 3.3.4b).

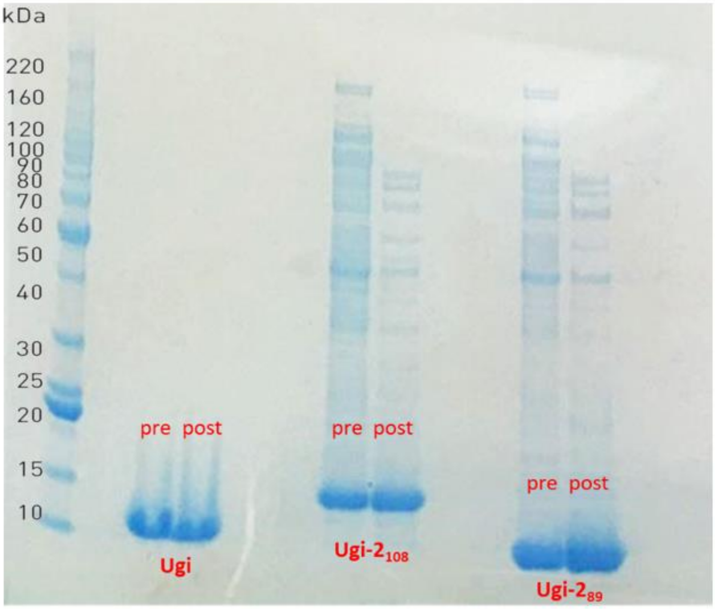


Figure 3.3.4a. SDS-PAGE analysis of Ugi (left), Ugi-2_{108} (Middle), and Ugi-2_{89} (right) soluble samples before (pre) and after (post) heating to 95 °C for 10 minutes. All three variants remained soluble after heat treatment.

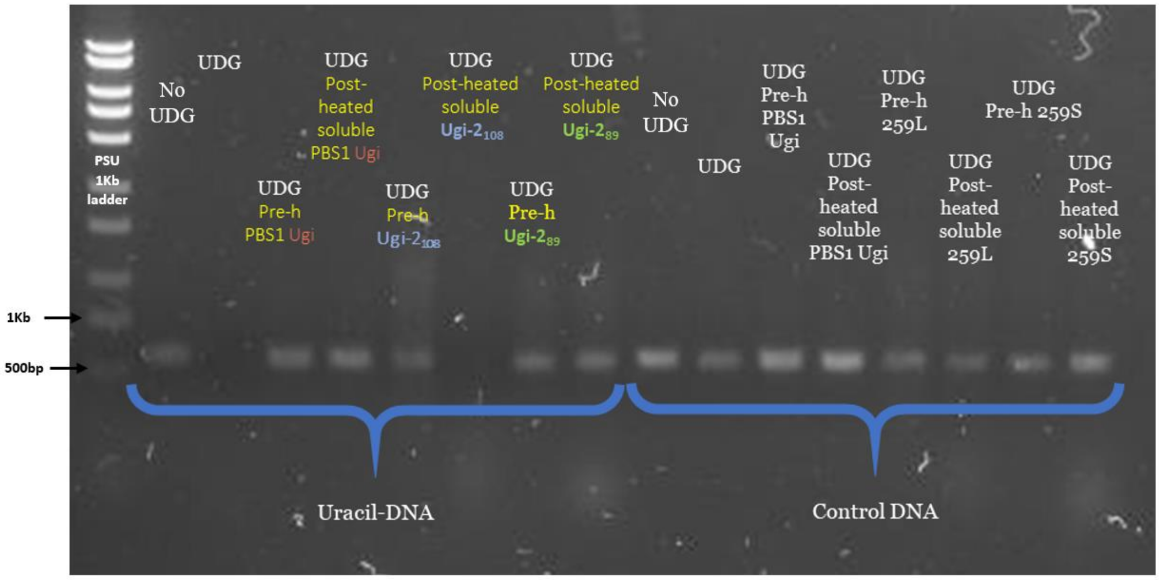

PSU 1Kb ladder
No UDG
UDG
UDG Pre-h PBS1 Ugi
UDG Post-heated soluble PBS1 Ugi
UDG Pre-h Ugi-2108
UDG Post-heated soluble Ugi-2108
UDG Pre-h Ugi-289
UDG Post-heated soluble Ugi-289
No UDG
UDG
UDG Pre-h PBS1 Ugi
UDG Post-heated soluble PBS1 Ugi
UDG Pre-h 259L
UDG Post-heated soluble 259L
UDG Pre-h 259S
UDG Post-heated soluble 259S
1Kb
500bp
Uracil-DNA
Control DNA

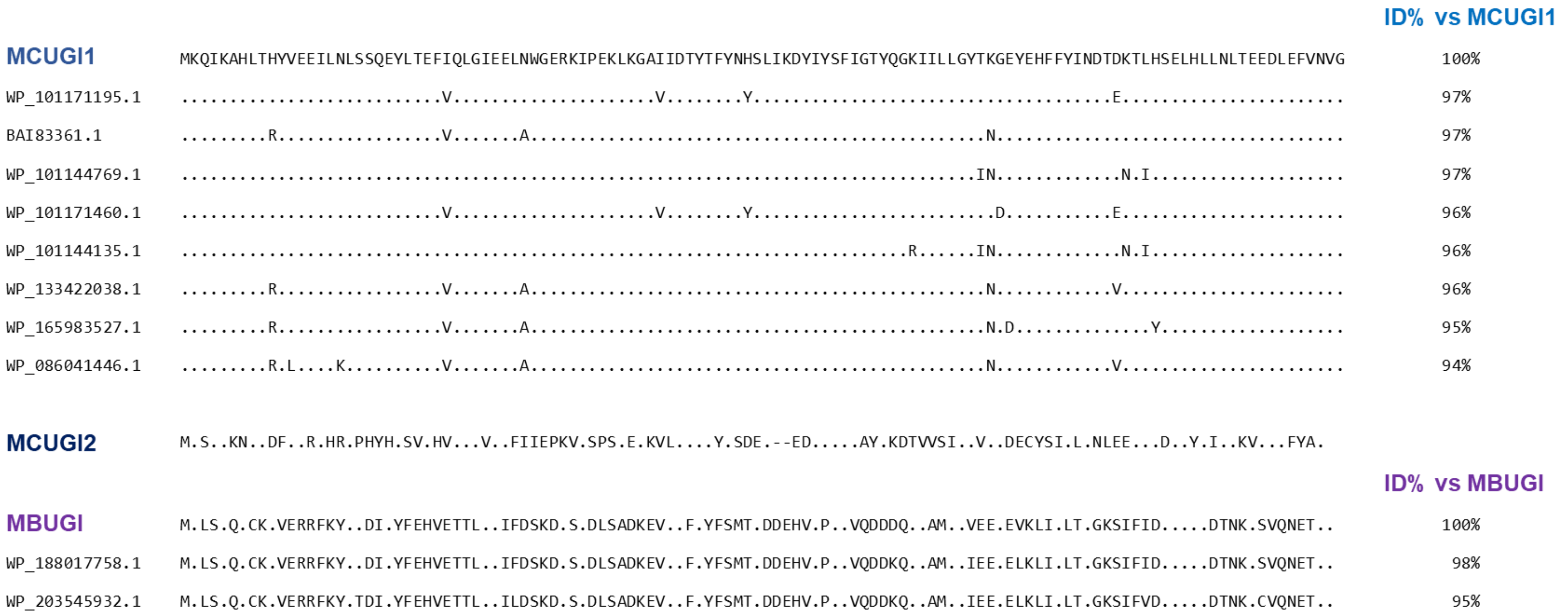
Figure 3.3.4b. UDG assay of Ugi-2 variants. Assay showed that both Ugi-2_{89} and Ugi-2_{108} were able to inhibit Ung activity. However, Ugi-2_{89} but not Ugi-2_{108} retained inhibitory ability after heating the protein to 95 °C for 10 minutes and ambient cooling to room temperature. The gel represents three independent replicates.

3.4. Bioinformatics searches for SAUGI homologs

3.4.1 *In silico* Identification

In contrast to PBS1/AR9 phage-encoded Ugi, a PSI-BLAST search with the SAUGI (PDB: 3WDG) encoded by the staphylococcal cassette chromosome *mec* (SCC*mec*), the mobile genetic element that carries mecA (methicillin-resistant gene), output 977 hits. Limiting the search to *Staphylococcaceae* increased the sensitivity and output 1024 sequences. Applying thresholds of >80% query cover and <35% percent identity to the query, 15 non-redundant sequences were selected. Thirteen of these sequences are encoded by *Macrococcus* species, one sequence is encoded by *Salinicoccus* sp. YB14-2, and one sequence is encoded by *Jeotgalicoccus meleagridis*. Sequences encoded by *Macrococcus* were divided into three groups according to their shared sequence identity (Figure 3.4.1a); Three representative sequences were designated MCUGI1 (WP_101156358.1), MCUGI2 (WP_101143899.1), and MBUGI (WP_165958605.1). The sequences encoded by *Salinicoccus* sp. YB14-2 and *Jeotgalicoccus meleagridis* were designated SYUGI (WP_052256111.1) and JMUGI (WP_185124884.1), respectively. No additional sequences satisfying the set criteria could be recovered from a subsequent HMMER search or HHblits search. Importantly, SYUGI and JMUGI sequences were not output by a PSI-BLAST search unless limiting the search to the *Staphylococcaceae* family. However, SYUGI and JMUGI sequences were found in the outputs of HMMER search, and JMUGI but not SYUGI was found in the outputs of HHblits search (Appendix B).

Multiple Sequence alignment was performed online using MAFFT at the EMBL-EBI webserver[124,144]. MCUGI1, MCUGI2, MBUGI, SYUGI, and JMUGI exhibit high sequence plasticity when compared with SAUGI or with each other with identities ranging from 25% to 42% (Figure 3.4.1b).



```
                                                                                                                                               ID% vs MCUGI1
MCUGI1           MKQIKAHLTHYVEEILNLSSQEYLTEFIQLGIEELNWGERKIPEKLKGAIIDTYTFYNHSLIKDYIYSFIGTYQGKIILLGYTKGEYEHFFYINDTDKTLHSELHLLNLTEEDLEFVNVG        100%
WP_101171195.1   ...........................V.....................V........Y.....................................E.......................         97%
BAI83361.1       .........R.................V.......A...............................................N....................................         97%
WP_101144769.1   ..................................................................................IN.............N.I....................         97%
WP_101171460.1   ...........................V.....................V........Y.........................D...........E.......................         96%
WP_101144135.1   ...........................................................................R......IN.............N.I....................         96%
WP_133422038.1   .........R.................V.......A...............................................N............V.......................         96%
WP_165983527.1   .........R.................V.......A...............................................N.D..............Y...................         95%
WP_086041446.1   .........R.L....K..........V.......A...............................................N............V.......................         94%

MCUGI2           M.S..KN..DF..R.HR.PHYH.SV.HV...V..FIIEPKV.SPS.E.KVL....Y.SDE.--ED.....AY.KDTVVSI..V..DECYSI.L.NLEE...D..Y.I..KV...FYA.
                                                                                                                                               ID% vs MBUGI
MBUGI            M.LS.Q.CK.VERRFKY..DI.YFEHVETTL..IFDSKD.S.DLSADKEV..F.YFSMT.DDEHV.P..VQDDDQ..AM..VEE.EVKLI.LT.GKSIFID.....DTNK.SVQNET..        100%
WP_188017758.1   M.LS.Q.CK.VERRFKY..DI.YFEHVETTL..IFDSKD.S.DLSADKEV..F.YFSMT.DDEHV.P..VQDDKQ..AM..IEE.ELKLI.LT.GKSIFID.....DTNK.SVQNET..         98%
WP_203545932.1   M.LS.Q.CK.VERRFKY.TDI.YFEHVETTL..ILDSKD.S.DLSADKEV..F.YFSMT.DDEHV.P..VQDDKQ..AM..IEE.ELKLI.LT.GKSIFVD.....DTNK.CVQNET..         95%
```

Figure 3.4.1a. MSA view (Flat query-anchored with dots for identity) for SAUGI homologs encoded by *Macrococcus* species. Three groups can be observed according to sequence similarity. Group I include MCUGI1 and 8 homologous sequences with at least 94% Identity; Group II includes one sequence, MCUGI2; and Group III include MBUGI and 2 homologous sequences with at least 95% identity. Representative sequences of these groups (MCUGI1, MCUGI2, and MBUGI) were selected for verifying Ung inhibition ability.

A

```
SYUGI       ---MHQKLKQYITRHLKKS-EDEYLSESFVLPSTETFQSPQFQ-RLFDDQSLSHQLYYST
JMUGI       ---MNTKLKIYIKKYFPELSTLTWSDEAVSMSGDELFEDTKLK-SLYENESLDTRLYYPI
SAUGI       -MTLELQLKHYITNLFNLPKDEKWECESIEEIADDILPDQYVRLGALSNKILQTYTYYSD
MBUGI       -MSLSEQLCKFVERRFKYLND-IWYFEHVETTLGEIFDSKDLSGDLSADKEVDTFTYFSM
MCUGI1      MKQIKAHLTRYLEEILKLSSQ-EYLTEFVQLGIEELAWGERKIPEKLKGAIIDTYTFYNH
MCUGI2      -MSIKKNLTDFVERIHRLPHY-HYSVEHVQLGVEEFIIEPKVISPSLEGKVLDTYTYYSD
               :  :*  :: .         :  * .     :             .  :.   ::

SYUGI       TD-DEPFFPFEVYQDDTLIALGYMEED-KQHILYLKHDDEILVEEL-------------
JMUGI       EI-NSAILPFEIYKEETLVALGYTNDE-SQKIIYFKHGAETLINHL*------------
SAUGI       TLHESNIYPFILYYQKQLIAIGYIDENHDMDFLYLHNTIMPLLDQRYLLTGGQ-------
MBUGI       TLDDEHVYPFIVQDDDQIIAMGYVEEE-EVKLIYLTDGKSIFIDELHLLDTNKESVQNET
MCUGI1      SLIKDYIYSFIGTYQGKIILLGYTNGE-YEHFFYINDTVKTLHSELHLLNLTEEDLEFVN
MCUGI2      ELE--DIYSFIAYYKDTVVSIGYVKGD-ECYSIYLNNLEETLHDELYLINLKVEDLFYAN
                  .  *    .  :: :** . :     :*: .    : ..

SYUGI       ----
JMUGI       ----
SAUGI       ----
MBUGI       VG--
MCUGI1      VG--
MCUGI2      FDVG
```

B

ID%	SAUGI	MCUGI1	MCUGI2	MBUGI	SYUGI	JMUGI
SAUGI	100	29	31	29	30	25
MCUGI1	29	100	42	27	28	25
MCUGI2	31	42	100	35	28	26
MBUGI	29	27	35	100	35	25
SYUGI	30	28	28	35	100	33
JMUGI	25	25	26	25	33	100

Figure 3.4.1b. Sequence plasticity in homologs of SAUGI. (A) Clustal format structure-based multiple sequence alignment of SAUGI and the identified distant homologs. Multiple Sequence alignment was performed online using MAFFT at the EMBL-EBI webserver. Among the 112 amino acid residues in SAUGI sequence, only 6 residues, excluding the start codon translated methionine, remain identical in all six variants. (B) Sequence identity matrix between SAUGI HOMOLOGUES. This matrix highlights that no pair among these six distant homologs share > 42% identity. Sequence identity amongst these SAUGI HOMOLOGUES can be as low as 25%.

3.4.2. Cloning of SAUGI homologues

The laboratory work reported in sections 3.4.2, 3.4.3, and 3.4.4 was performed by Ms Rosalia Santangelo, formerly an undergraduate project student in the Savva research lab.

The synthetic gene sequences of SAUGI, MCUGI1, MCUGI2, MBUGI, SYUGI, and JMUGI were designed *in silico* (Appendix A) as described (section 2.1.2). Primers P14-P25 were used to amplify the synthetic genes to be cloned into a pRSET-C vector using an overlap extension PCR/ligation cloning strategy as described (section 2.1.1.11). Sequencing results verified the fidelity of constructs for MCUGI2, MBUGI, SYUGI, and JMUGI. However, a double mutant MCUGI1 (G39V/L80V) was cloned instead of the annotated sequence. The missense mutation of the MCUGI1 double mutant at residue 39 was corrected via iPCR/ligation using pre-phosphorylated primers P26 and P27. Both the double mutant and the single mutant were investigated (section 3.4.4), assuming that the conserved mutation in the single mutant is not likely to affect the protein function. The wild type SAUGI sequence (PDB: 3WDG) was cloned to be used as an assay control.

3.4.3. Expression of SAUGI homologues

SAUGI homologues were expressed at small-scale (section 2.1.2.1). All the homologues produced soluble proteins (Figure 3.4.3), the soluble fraction of the *lysY/I^q E. coli* cell lysate was assayed for UDG activity.

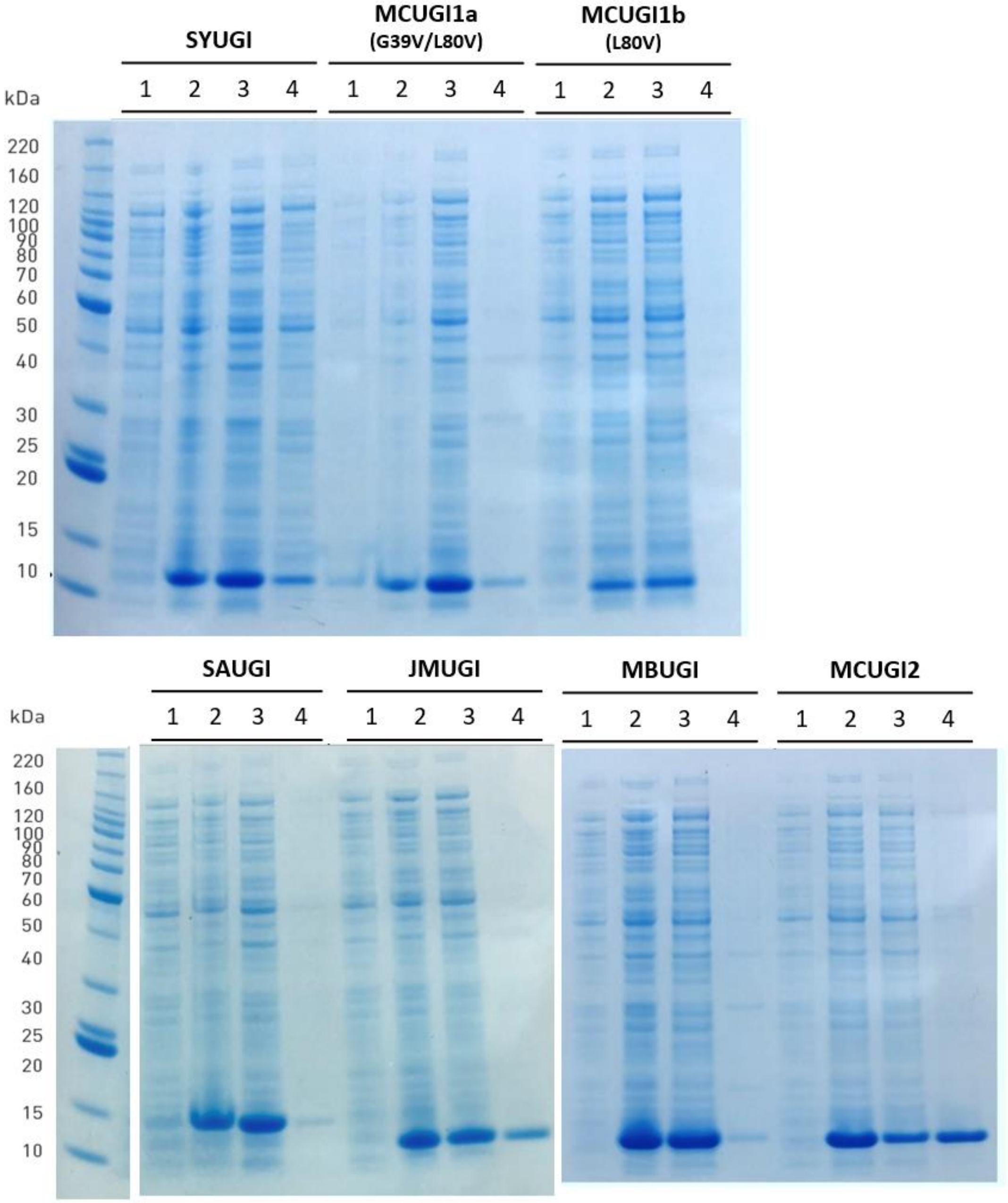


Figure 3.4.3. SDS-PAGE gels for SAUGI homologs. Marker lane: Benckmark^{TM} protein ladder (Invitrogen)
Lanes for each named sample set are as follows:
1) pre-induction
2) cell harvest post induction + 16 hours
3) clarified supernatant fraction
4) Insoluble fraction
All the expressed genes produced a soluble protein as shown in lane 3 of each set.

3.4.4. Ung inhibition assay of SAUGI homologues

A serial dilution in STE buffer was used to discover the minimum *Staphylococcus aureus* UNG enzyme concentration required to unambiguously process 5 µL of uracil-DNA substrate relative to the thymine-DNA control (Figure 3.4.4a). The UDG assay (section 2.1.2.7.1) was validated with different controls (Figure 3.4.4b). The minimum concentration of SAUNG that showed activity on uracil-DNA is 39.7 nM (Figure 3.4.4a). SAUGI and its distant homologues MCUGI2, MBUGI, SYUGI, and JMUGI showed ability to inhibit Ung. Additionally, MCUGI1 mutant (L80V) but not the double mutant (G39V/L80V) was able to inhibit Ung (Figure 3.4.4c).

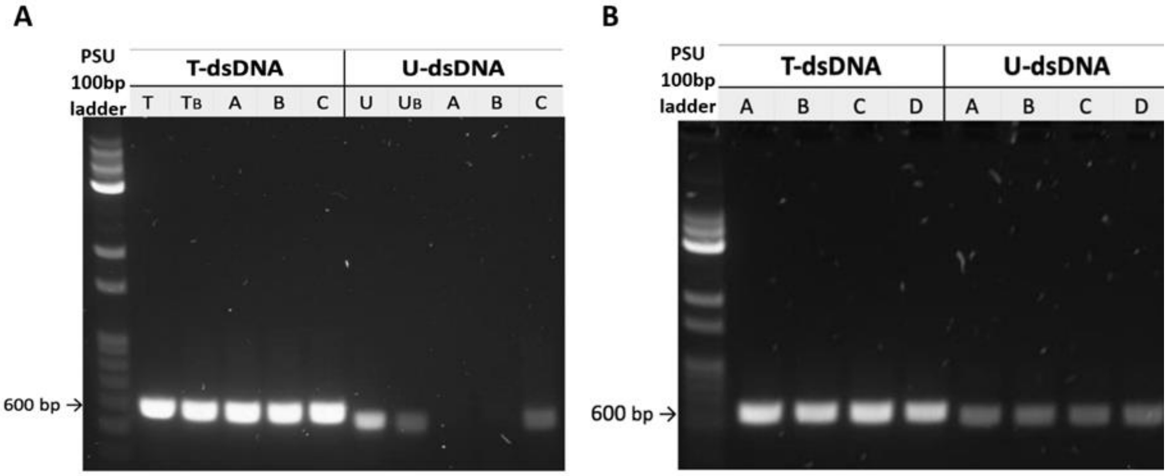


Figure 3.4.4a. Visual U-DNA attrition assay [*Staphylococcus aureus* UNG activity] - dilution series of Ung. The Ung assay reaction products to verify two serial dilution of Ung from S. aureus [0.1 mg/mL, 3.97 µM] are shown on 1 % agarose gels. **A)** The reaction products are split according to the DNA substrate been used: thymine-DNA (T-dsDNA) and uracil-DNA (U-dsDNA). In each section the first two lanes represent the controls: DNA untreated with the assay buffer and thermocycling conditions (lanes T and U); and DNA treated with the assay buffer and thermocycling conditions in the absence of Ung (lanes T_B and U_B). (A) The first serial dilution follows the order: Ung in 1:10 dilution (lane A), 1:100 dilution (lane B) and 1:1000 dilution (lane C). **B)** The reaction products following the second serial dilution are displayed in the following order: 1:200 dilution (lane A), 1:400 dilution (lane B), 1:600 dilution (lane C), 1:800 dilution (lane D). The minimum concentration of Ung that showed activity on U-DNA is 39.7 nM (1:100 dilution). The gels represent three independent replicates.

PSU 100bp ladder	A								B							
	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
Control DNA substrate	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-
Uracil-DNA substrate	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+
Treatment with reaction buffer and conditions	-	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+
1 µL *LysY/I*^{Q} cell lysate	-	-	+	+	+	-	-	-	-	-	+	+	+	-	-	-
1 µL CJ236 cell lysate	-	-	-	-	-	+	-	-	-	-	-	-	-	+	-	-
SAUNG	-	-	-	+	+	-	+	+	-	-	-	+	+	-	+	+
Ugi	-	-	-	-	+	-	-	+	-	-	-	-	+	-	-	+

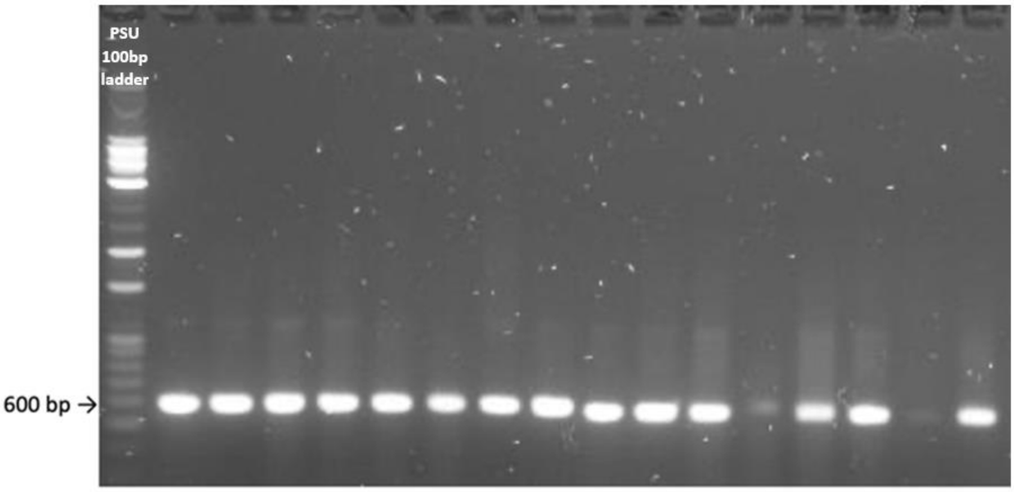


Figure 3.4.4b. Visual U-DNA attrition assay [*Staphylococcus aureus* UNG activity], validation with controls, 1% (w/v) agarose gel. The UDG assay control is divided into two sub-sections. **(A)** Thymine-DNA (3 µL) is used in all lanes. **(B)** Uracil-DNA (5 µL) is used in all lanes. SAUNG was added at [39.7 nM]. PBS1 Ugi was added at [2.6 mM]. The gel represents three independent replicates.

Uracil-DNA substrate	+	+	+	+	+	+	+	+	+	+	+	+
Treatment with reaction buffer and conditions	X	√	√	√	√	√	√	√	√	√	√	√
SAUNG	-	-	+	+	+	+	+	+	+	+	+	+
UngIn/potential UngIn	-	-	Ugi	SAUGI	-	SA	MC1a	MC1b	MC2	MB	SY	JM

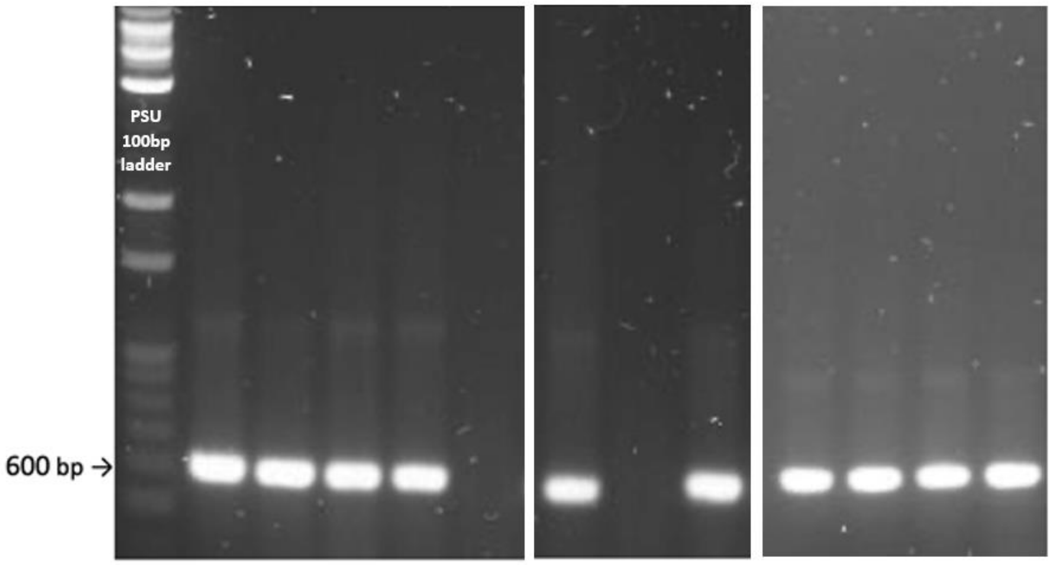
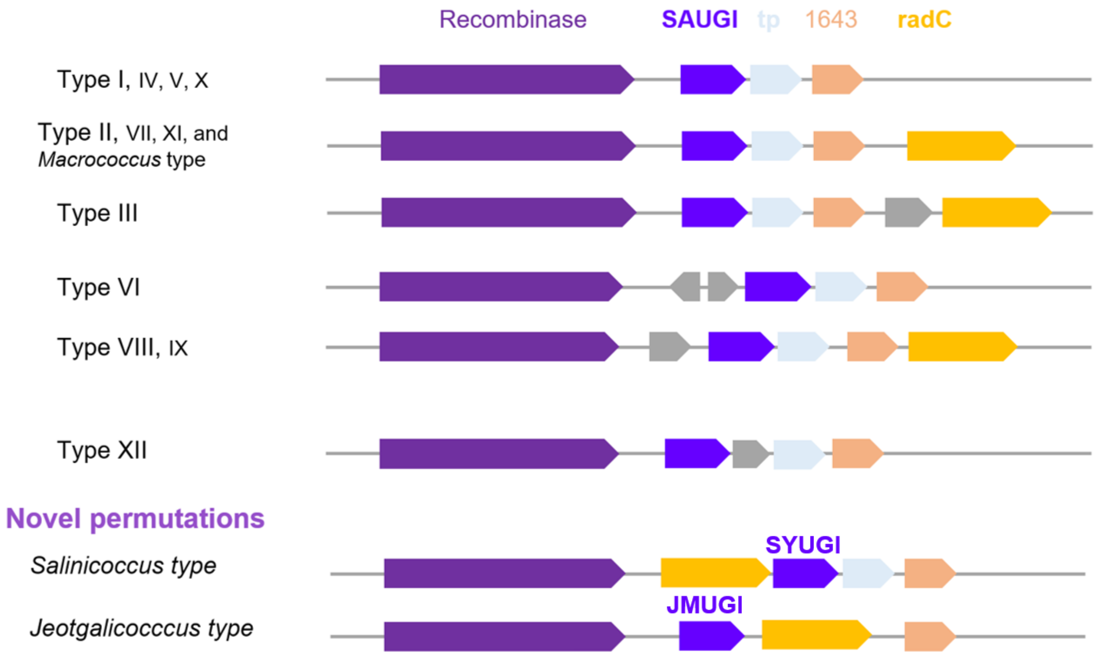


Figure 3.4.4c. Visual U-DNA attrition assay [*Staphylococcus aureus* UNG activity], to assess candidate SAUGI-type Ung inhibitors. All reactions contained substrate DNA (~600 bp) in which all thymidine is supplanted by deoxyuridine. All reactions, excluding the leftmost control lane were incubated at 37 °C for 30 minutes. SAUNG was added at [39.7 nM]. UngIns/potential UngIns were added either as a purified protein (Ugi and SAUGI lanes) or as the cell lysates of expressed actual/potential SAUGI-type Ung inhibitors: SA = SAUGI; MC1a = MCUGI1a (G39V, L80V); MC1b = MCUGI1b (L80V); MC2 = MCUGI2; MB = MBUGI, SY = SYUGI; JM = JMUGI. The UDG assay showed that SAUGI, MCUGI1b, MCUGI2, MBUGI, SYUGI, and JMUGI are Ung inhibitors while the double mutant MCUGI1 (MCUGI1a) is not able to inhibit Ung. This figure is formed of 3 cropped images. Original uncropped images are available and stored electronically. The data shown is this figure represents three independent replicates.

3.4.5. Newly discovered homologs of SAUGI reside in novel permutations within the SCC*mec* Cassette

SAUGI is a conserved gene in the transposable genomic pathogenicity island SCC*mec* of Methicillin Resistant *Staphylococcus aureus* (MRSA) strains, specifically on the cassette chromosome recombinase (ccr) complex[86]. Genomic mapping of the newly identified SAUGI homologue flanking regions confirmed that all of them are located in the ccr complex of SCC*mec*. Unusually, SYUGI and JMUGI were found to be located adjacent to the DNA repair protein radC (Figure 3.4.5).

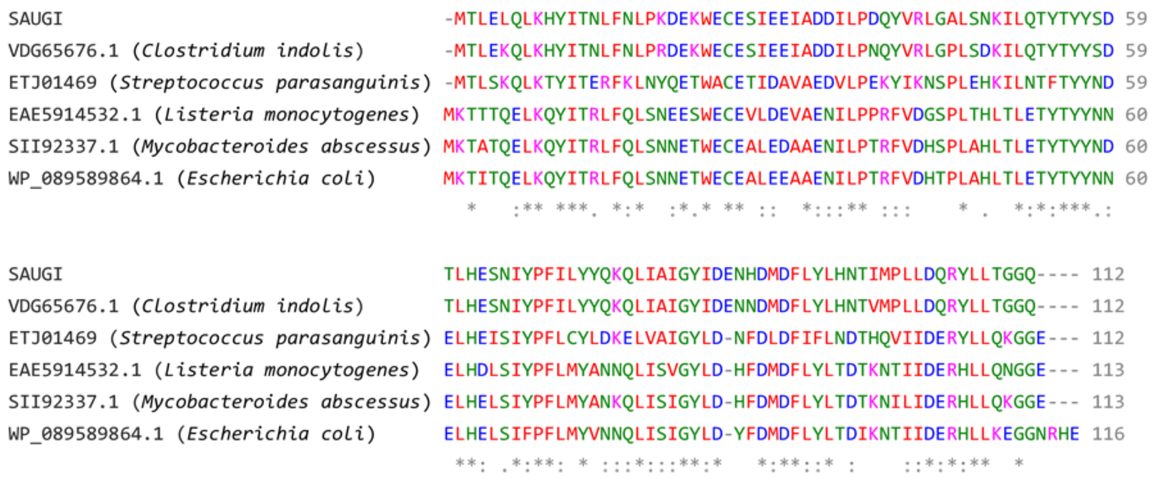


tp: a transposase usually annotated as DUF960, 1643: DUF1643, radC: DNA repair protein radC.

Figure 3.4.5. Genomic mapping of SAUGI HOMOLOGUES in different types of SCC*mec*. All SCC*mec* elements contain a *mec* gene complex (*mec*) and a cassette chromosome recombinase complex (*ccr*). Five classes of *mec* and eight types of *ccr* have been reported. Based on the combination of *mec* and *ccr*, SCC*mec* elements are classified into different types. Twelve types of SCC*mec* elements in *S. aureus* have been identified; all these types include a trio of genes: (1) *SAUGI*, usually annotated as DUF950; (2) *tp*, usually annotated as DUF960; and (3) DUF1643. This DUF 950-960-1643 trio, is always preceded by a recombinase gene and is often followed by a fourth gene, annotated as *DNA repair protein radC*. The genomic context of *MCUGI* shows similarity to known SCC*mec* types of *S. aureus*. However, the *SYUGI* and *JMUGI* genomic contexts are novel permutations in the *ccr* complex, wherein *radC* either precedes the DUF 950-960-1643 trio (in *Salinicoccus*) or separates *SAUGI* and DUF1643 (in *Jeotgalicoccus*). Gray coloured ORFs are non-conserved genes in SCC*mec*.

3.4.6. Proteins with an annotated SAUGI domain

SAUGI is usually annotated as DUF950[62] (a domain that represents a sequence motif of 112 amino acids). Searching the NCBI protein database for proteins with an annotated DUF950 sequence region, outputs sequences from the *Staphylococcaceae* family (including *Staphylococcus*, *Macrococcus*, *Salinicoccus*, and *Jeotgalicoccus* species) and sequences encoded by other taxonomic families including sequences highly homologous to SAUGI encoded by some strains of *Escherichia. coli*, *Streptococcus pneumoniae*, *Listeria monocytogenes*, *Mycobacteroides abscessus* and other bacterial species (Figure 3.4.6a). Genomic mapping of those highly homologous sequences shows that they are also located on an SCC*mec* ccr complex. This analysis indicates that the SCC*mec* cassette might transfer between different bacterial species via conjugation, which involves transfer of the cassette in the form of ssDNA, which is vulnerable to Ung activity if uracil is present. This suggests a biological context for the apparently strict conservation of SAUGI in the SCC*mec* cassette.



```
SAUGI                                  -MTLELQLKHYITNLFNLPKDEKWECESIEEIADDILPDQYVRLGALSNKILQTYTYYSD 59
VDG65676.1 (Clostridium indolis)       -MTLEKQLKHYITNLFNLPRDEKWECESIEEIADDILPNQYVRLGPLSDKILQTYTYYSD 59
ETJ01469 (Streptococcus parasanguinis) -MTLSKQLKTYITERFKLNYQETWACETIDAVAEDVLPEKYIKNSPLEHKILNTFTYYND 59
EAE5914532.1 (Listeria monocytogenes)  MKTTTQELKQYITRLFQLSNEESWECEVLDEVAENILPPRFVDGSPLTHLTLETYTYYNN 60
SII92337.1 (Mycobacteroides abscessus) MKTATQELKQYITRLFQLSNNETWECEALEDAAENILPTRFVDHSPLAHLTLETYTYYND 60
WP_089589864.1 (Escherichia coli)      MKTITQELKQYITRLFQLSNNETWECEALEEAAENILPTRFVDHTPLAHLTLETYTYYNN 60
                                         *   :** ***. *:*  :*.* ** ::  *:::** :::    * .  *:*:***.:

SAUGI                                  TLHESNIYPFILYYQKQLIAIGYIDENHDMDFLYLHNTIMPLLDQRYLLTGGQ---- 112
VDG65676.1 (Clostridium indolis)       TLHESNIYPFILYYQKQLIAIGYIDENNDMDFLYLHNTVMPLLDQRYLLTGGQ---- 112
ETJ01469 (Streptococcus parasanguinis) ELHEISIYPFLCYLDKELVAIGYLD-NFDLDFIFLNDTHQVIIDERYLLQKGGE--- 112
EAE5914532.1 (Listeria monocytogenes)  ELHDLSIYPFLMYANNQLISVGYLD-HFDMDFLYLTDTKNTIIDERHLLQNGGE--- 113
SII92337.1 (Mycobacteroides abscessus) ELHELSIYPFLMYANKQLISIGYLD-HFDMDFLYLTDTKNILIDERHLLQKGGE--- 113
WP_089589864.1 (Escherichia coli)      ELHELSIFPFLMYVNNQLISIGYLD-YFDMDFLYLTDIKNTIIDERHLLKEGGNRHE 116
                                        **: .*:**: * :::*:::**:*   *:**::* :    ::*:*:**  *
```

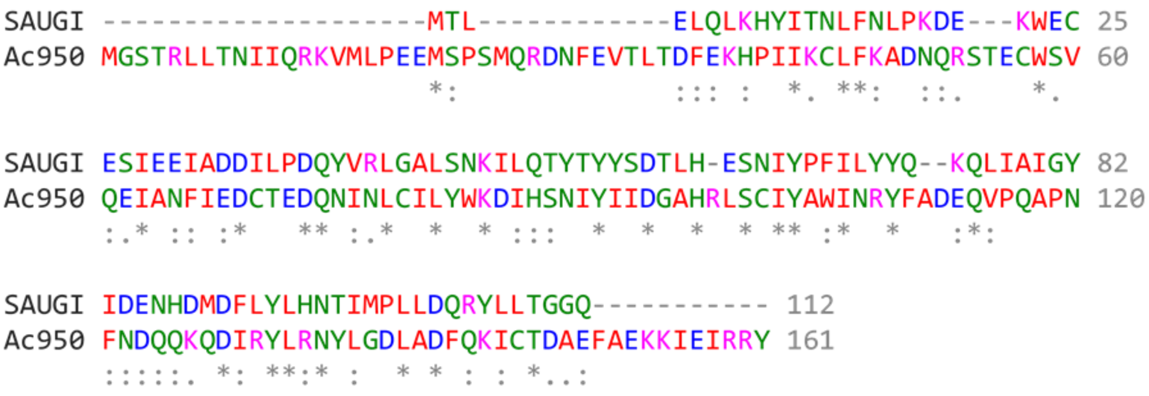
Figure 3.4.6a. MSA of SAUGI and its homologous sequences from strains of other bacterial species.

The NCBI protein database search for DUF950 also output 7 proteins less homologous to SAUGI with an annotated partial DUF950 domain region (Table 3.4.6). Six of these proteins have a DUF950 region with length <57aa out of 112aa domain; the seventh protein, encoded by *Acinetobacter* species (designated in this thesis Ac950), has a DUF950 region spanning 108 of 161 amino acids in that sequence.

Table 3.4.6. The NCBI database proteins with partial DUF950 domain. The second row represents protein Ac950 encoded by *Acinetobacter* species.

Organism	Protein accession (protein name)	Length of encoded protein with DUF950 region	DUF950 region spanning sequence
Cupriavidus taiwanensis	WP_018004428	128	19-74
Acinetobacter pittii	KQD32686.1 (Ac950)	161	33-141
Jatropha curcas	XP_012085129	283	139-188
Elusimicrobia bacterium	PIX14766	296	136-197
Aestuariibacter salexigens	WP_026377246	191	81-175
Halomonas sp. KO116	AJY53241	159	106-144
Pseudocohnilembus persalinus	KRX02449	867	736-855

The pairwise sequence alignment of Ac950 and SAUGI shows 25% sequence identity (Figure 3.4.6b). Interestingly, the InterPro protein families and domains database[145] lists Ac950, but none of the other 6 proteins with DUF950 region, in the *S. aureus* uracil-DNA glycosylase inhibitor family (IPR009295).



```
SAUGI --------------------MTL------------ELQLKHYITNLFNLPKDE---KWEC 25
Ac950 MGSTRLLTNIIQRKVMLPEEMSPSMQRDNFEVTLTDFEKHPIIKCLFKADNQRSTECWSV 60
                          *:             ::: :  *. **:  ::.    *.

SAUGI ESIEEIADDILPDQYVRLGALSNKILQTYTYYSDTLH-ESNIYPFILYYQ--KQLIAIGY 82
Ac950 QEIANFIEDCTEDQNINLCILYWKDIHSNIYIIDGAHRLSCIYAWINRYFADEQVPQAPN 120
      :.* :: :*   ** :.*  *  * :::  *  *  *  * ** :*  *   :*:

SAUGI IDENHDMDFLYLHNTIMPLLDQRYLLTGGQ----------- 112
Ac950 FNDQQKQDIRYLRNYLGDLADFQKICTDAEFAEKKIEIRRY 161
      :::::. *: **:* :  * * : : *..:
```

Figure 3.4.6b. Pairwise sequence alignment of Ac950 and SAUGI. Interestingly, the 2 sequences share 25% identity and 58% homology (which represents the sum of identity and conserved and semi-conserved mutations).

Ac950 was cloned into pSDM4_U12_Ung using OE-PCR/ligation strategy. The Ac950 gene (Appendix A) was amplified from synthetic DNA using primers P28 and P29 and pSDM4_U12_Ung was linearised using primers P1 and P2. The verified construct was transformed into CJ236 to assay Ac950 via the bacterial lethal UDG inhibition assay (section

2.1.2.7.3); Ac950 apparently did not inhibit Ung. SAUGI-type Ung inhibitors (SYUGI and MBUGI) were used as controls for this assay and apparently inhibited Ung. If Ac950 is related to SAUGI this work would suggest that it is not an UngIn; it has four important differences in comparison to other SAUGI homologues (i.e., MCUGIs, MBUGI, SYUGI, and JMUGI): (1) Performing a PSI-BLAST search for Ac950 output several sequences annotated as HNH endonucleases; (2) the 25% sequence identity that Ac950 shares with SAUGI is distributed across various SAUGI secondary structure features rather than just the core-forming secondary structure, typical of the verified UngIn variants (Figure 3.4.6c); (3) none of the main secondary structure breaking residues: glycine and proline, is conserved between Ac950 and SAUGI, and (4) SCC*mec* is not the genetic locus for Ac950. A 3D-structure model of Ac950 was generated using AlphaFold: the predicted model (high confidence prediction, pLDDT score=86) has a different fold/structure in comparison with SAUGI (Figure 3.4.6d).

These results strongly imply that signature motif conservation and genomic context are more important than the sum sequence identity in SAUGI homologous sequences.

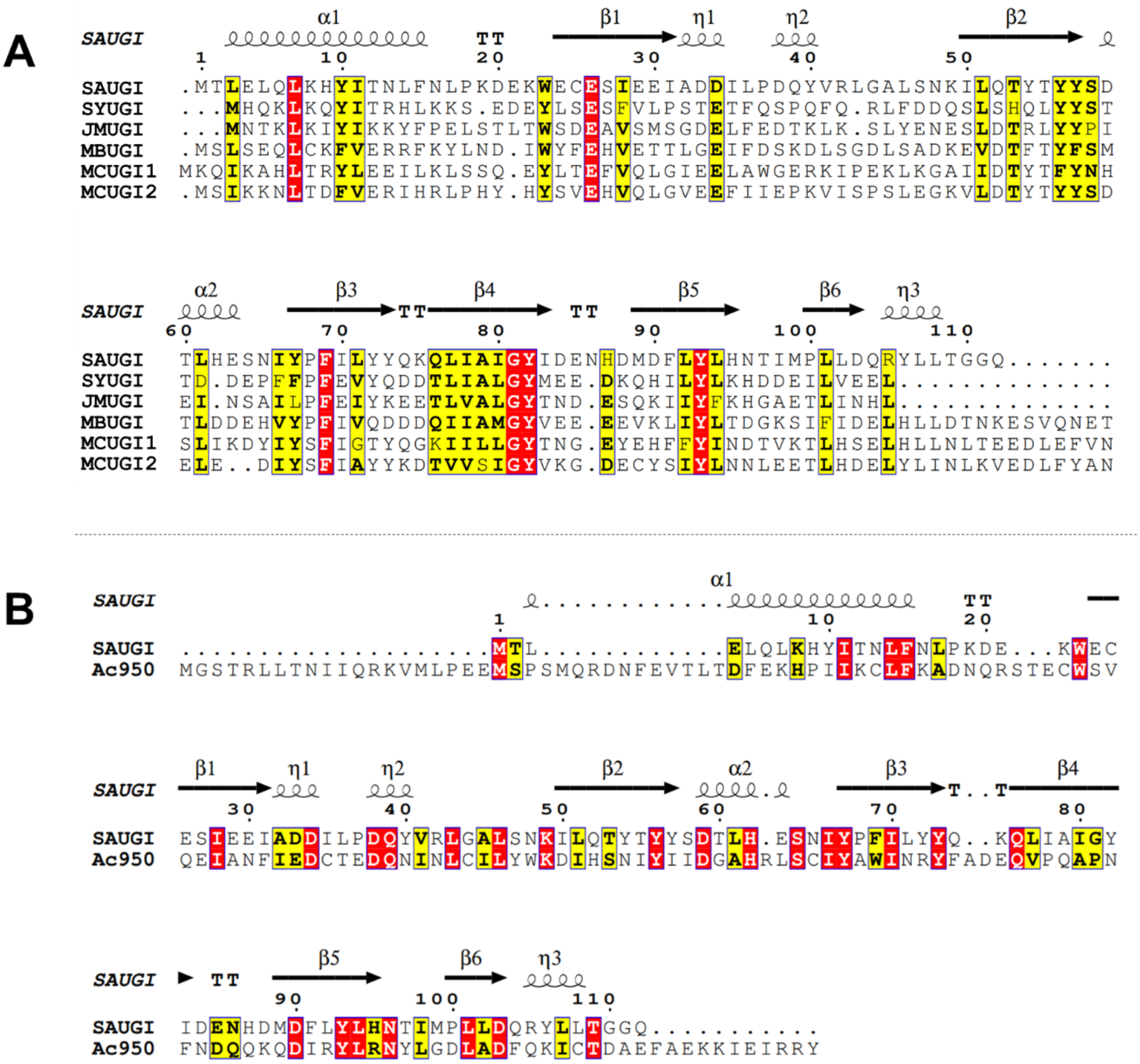

A
SAUGI
α1
TT
β1
η1
η2
β2
1
10
20
30
40
50
SAUGI .MTLELQLKHYITNLFNLPKDEKWECESIEEIADDILPDQYVRLGALSNKILQTYTYYSD
SYUGI ...MHQKLKQYITRHLKKS.EDEYLSESFVLPSTETFQSPQFQ.RLFDDQSLSHQLYYST
JMUGI ...MNTKLKIYIKKYFPELSTLTWSDEAVSMSGDELFEDTKLK.SLYENESLDTRLYYPI
MBUGI .MSLSEQLCKFVERRFKYLND.IWYFEHVETTLGEIFDSKDLSGDLSADKEVDTFTYFSM
MCUGI1 MKQIKAHLTRYLEEILKLSSQ.EYLTEFVQLGIEELAWGERKIPEKLKGAIIDTYTFYNH
MCUGI2 .MSIKKNLTDFVERIHRLPHY.HYSVEHVQLGVEEFIIEPKVISPSLEGKVLDTYTYYSD
SAUGI
α2
β3
TT
β4
TT
β5
β6
η3
60
70
80
90
100
110
SAUGI TLHESNIYPFILYYQKQLIAIGYIDENHDMDFLYLHNTIMPLLDQRYLLTGGQ.......
SYUGI TD.DEPFFPFEVYQDDTLIALGYMEE.DKQHILYLKHDDEILVEEL.............
JMUGI EI.NSAILPFEIYKEETLVALGYTND.ESQKIIYFKHGAETLINHL.............
MBUGI TLDDEHVYPFIVQDDDQIIAMGYVEE.EEVKLIYLTDGKSIFIDELHLLDTNKESVQNET
MCUGI1 SLIKDYIYSFIGTYQGKIILLGYTNG.EYEHFFYINDTVKTLHSELHLLNLTEEDLEFVN
MCUGI2 ELE..DIYSFIAYYKDTVVSIGYVKG.DECYSIYLNNLEETLHDELYLINLKVEDLFYAN
B
SAUGI
α1
TT
1
10
20
SAUGI ....................MTL.............ELQLKHYITNLFNLPKDE...KWEC
Ac950 MGSTRLLTNIIQRKVMLPEEMSPSMQRDNFEVTLTDFEKHPIIKCLFKADNQRSTECWSV
SAUGI
β1
η1
η2
β2
α2
β3
T..T
β4
30
40
50
60
70
80
SAUGI ESIEEIADDILPDQYVRLGALSNKILQTYTYYSDTLH.ESNIYPFILYYQ..KQLIAIGY
Ac950 QEIANFIEDCTEDQNINLCILYWKDIHSNIYIIDGAHRLSCIYAWINRYFADEQVPQAPN
SAUGI
TT
β5
β6
η3
90
100
110
SAUGI IDENHDMDFLYLHNTIMPLLDQRYLLTGGQ...........
Ac950 FNDQQKQDIRYLRNYLGDLADFQKICTDAEFAEKKIEIRRY

Figure 3.4.6c. Structure-based sequence alignment of Ac950 with SAUGI. (A) MSA of SAUGI and its distant verified homologues in *Staphylococcaceae* family. (B) Sequence alignment of Ac950 and SAUGI. The sequence alignment was performed using Clustal Omega and was rendered in EsPript[146].

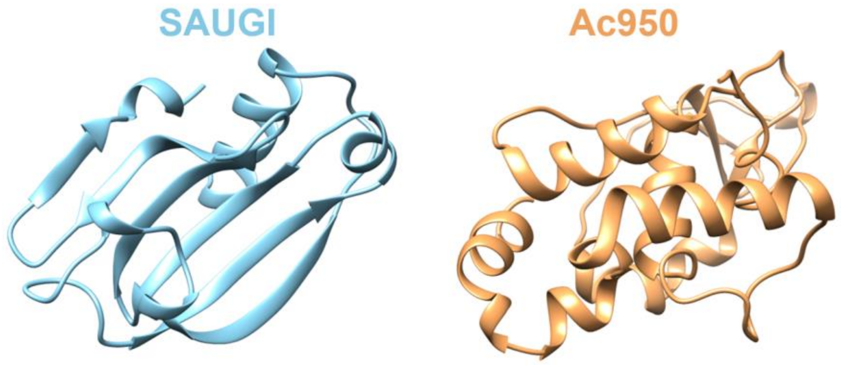


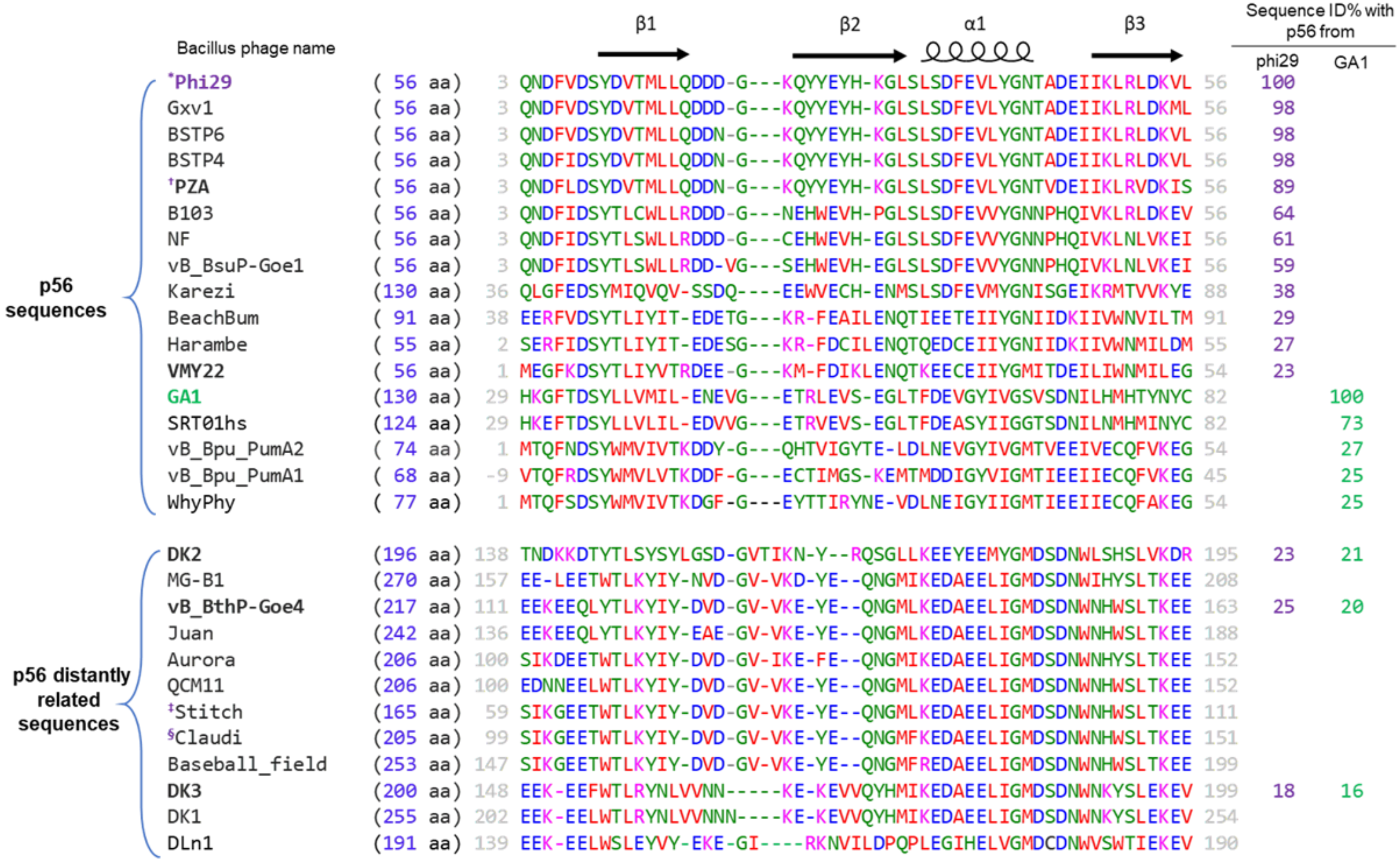
Figure 3.4.6d. Structure comparison of SAUGI and Ac950. SAUGI structure on the left (PDB: 3WDG-B) shows significant difference from the AlphaFold2 predicted model of Ac950 structure. The significant structural differences explain why Ac950, with a relatively similar sequence, is not able to act as an Ung inhibitor.

3.5. Bioinformatics searches for p56 homologs

3.5.1. Identification of potential p56 homologs

Nineteen homologous sequences were output from a PHI-BLAST search using the PROSITE pattern E-X(2)-Y-X(0,2)-G and the template Phi29 p56 sequence. These sequences were divided into Phi29-type p56-like sequences (having EXXYG motif at the α-helix) and GA1-type p56-like sequences (having EXXYXXG motif at the α-helix) (Figure 3.5.1). Out of these hits, the most distantly related hits to a known p56 sequence are encoded by *Bacillus* phage VMY22 and *Bacillus* phage WhyPhy. The VMY22 hit shares only 23% sequence identity with Phi29 p56, and the WhyPhy hit shares 25% sequence identity with GA1 p56.

Following verification of the VMY22 p56 as an Ung inhibitor (section 3.5.2), it was used as a PSI-BLAST template to search for new p56 proteins in the non-redundant database. This search showed no new potential p56 sequences; however, performing PSI-BLAST only within genomes of *Salasmaviridae* resulted in additional potential sequences after 4 iterations of the search (Figure 3.5.1). Interestingly, phages encoding these hits have significant genomic and proteomic similarity to known p56 encoding phages[96,97,147,148]. The hit sequences vary in length and partly align with known p56 sequences (Figure 3.5.1). Hits from *Bacillus* phages: DK2, DK3, and vB_BthP-Goe4 were selected for cloning, expression, and UngIn assay (section 3.5.2).



	Bacillus phage name					Sequence ID% with p56 from	
						phi29	GA1
p56 sequences	*Phi29	(56 aa)	3	QNDFVDSYDVTMLLQDDD-G---KQYYEYH-KGLSLSDFEVLYGNTADEIIKLRLDKVL	56	100	
	Gxv1	(56 aa)	3	QNDFVDSYDVTMLLQDDD-G---KQYYEYH-KGLSLSDFEVLYGNTADEIIKLRLDKML	56	98	
	BSTP6	(56 aa)	3	QNDFVDSYDVTMLLQDDN-G---KQYYEYH-KGLSLSDFEVLYGNTADEIIKLRLDKVL	56	98	
	BSTP4	(56 aa)	3	QNDFIDSYDVTMLLQDDN-G---KQYYEYH-KGLSLSDFEVLYGNTADEIIKLRLDKVL	56	98	
	†PZA	(56 aa)	3	QNDFLDSYDVTMLLQDDN-G---KQYYEYH-KGLSLSDFEVLYGNTVDEIIKLRVDKIS	56	89	
	B103	(56 aa)	3	QNDFIDSYTLCWLLRDDD-G---NEHWEVH-PGLSLSDFEVVYGNNPHQIVKLRLDKEV	56	64	
	NF	(56 aa)	3	QNDFIDSYTLSWLLRDDD-G---CEHWEVH-EGLSLSDFEVVYGNNPHQIVKLNLVKEI	56	61	
	vB_BsuP-Goe1	(56 aa)	3	QNDFIDSYTLSWLLRDD-VG---SEHWEVH-EGLSLSDFEVVYGNNPHQIVKLNLVKEI	56	59	
	Karezi	(130 aa)	36	QLGFEDSYMIQVQV-SSDQ----EEWVECH-ENMSLSDFEVMYGNISGEIKRMTVVKYE	88	38	
	BeachBum	(91 aa)	38	EERFVDSYTLIYIT-EDETG---KR-FEAILENQTIEETEIIYGNIIDKIIVWNVILTM	91	29	
	Harambe	(55 aa)	2	SERFIDSYTLIYIT-EDESG---KR-FDCILENQTQEDCEIIYGNIIDKIIVWNMILDM	55	27	
	VMY22	(56 aa)	1	MEGFKDSYTLIYVTRDEE-G---KM-FDIKLENQTKEECEIIYGMITDEILIWNMILEG	54	23	
	GA1	(130 aa)	29	HKGFTDSYLLVMIL-ENEVG---ETRLEVS-EGLTFDEVGYIVGSVSDNILHMHTYNYC	82		100
	SRT01hs	(124 aa)	29	HKEFTDSYLLVLIL-EDVVG---ETRVEVS-EGLTFDEASYIIGGTSDNILNMHMINYC	82		73
	vB_Bpu_PumA2	(74 aa)	1	MTQFNDSYWMVIVTKDDY-G---QHTVIGYTE-LDLNEVGYIVGMTVEEIVECQFVKEG	54		27
	vB_Bpu_PumA1	(68 aa)	-9	VTQFRDSYWMVLVTKDDF-G---ECTIMGS-KEMTMDDIGYVIGMTIEEIIECQFVKEG	45		25
	WhyPhy	(77 aa)	1	MTQFSDSYWMVIVTKDGF-G---EYTTIRYNE-VDLNEIGYIIGMTIEEIIECQFAKEG	54		25
p56 distantly related sequences	DK2	(196 aa)	138	TNDKKDTYTLSYSYLGSD-GVTIKN-Y--RQSGLLKEEYEEMYGMDSDNWLSHSLVKDR	195	23	21
	MG-B1	(270 aa)	157	EE-LEETWTLKYIY-NVD-GV-VKD-YE--QNGMIKEDAEELIGMDSDNWIHYSLTKEE	208		
	vB_BthP-Goe4	(217 aa)	111	EEKEEQLYTLKYIY-DVD-GV-VKE-YE--QNGMLKEDAEELIGMDSDNWNHWSLTKEE	163	25	20
	Juan	(242 aa)	136	EEKEEQLYTLKYIY-EAE-GV-VKE-YE--QNGMLKEDAEELIGMDSDNWNHWSLTKEE	188		
	Aurora	(206 aa)	100	SIKDEETWTLKYIY-DVD-GV-IKE-FE--QNGMIKEDAEELIGMDSDNWNHYSLTKEE	152		
	QCM11	(206 aa)	100	EDNNEELWTLKYIY-DVD-GV-VKE-YE--QNGMLKEDAEELIGMDSDNWNHWSLTKEE	152		
	‡Stitch	(165 aa)	59	SIKGEETWTLKYIY-DVD-GV-VKE-YE--QNGMLKEDAEELIGMDSDNWNHWSLTKEE	111		
	§Claudi	(205 aa)	99	SIKGEETWTLKYIY-DVD-GV-VKE-YE--QNGMFKEDAEELIGMDSDNWNHWSLTKEE	151		
	Baseball_field	(253 aa)	147	SIKGEETWTLKYIY-DVD-GV-VKE-YE--QNGMFREDAEELIGMDSDNWNHWSLTKEE	199		
	DK3	(200 aa)	148	EEK-EEFWTLRYNLVVNN-----KE-KEVVQYHMIKEDAEELIGMDSDNWNKYSLEKEV	199	18	16
	DK1	(255 aa)	202	EEK-EELWTLRYNLVVNNN----KE-KEVVQYHMIKEDAEELIGMDSDNWNKYSLEKEV	254		
	DLn1	(191 aa)	139	EEK-EELWSLEYVY-EKE-GI----RKNVILDPQPLEGIHELVGMDCDNWVSWTIEKEV	190		

* Identical sequence is encoded by *Bacillus* phage vB_BveP-Goe6 (ASR76788.1). † Identical sequences are encoded by *Bacillus* phage Whiting18 (QRD99282.1) and *Bacillus* phage Arbo1 (UIS65815.1). ‡ Identical sequences are encoded by *Bacillus* phages StevenHerd11 (AZF88314.1) and RadRaab (ASU04169.1). § Proteins with Identical sequences in the region shown in this alignment are encoded by *Bacillus* phages VioletteMad (QDH50288.1), KonjoTrouble (ASU04129.1), and SerPounce (ARQ95541.1).

Figure 3.5.1. MSA of p56 sequences and distantly related sequences. The top group of sequences shows the PHI-BLAST search generated p56 sequences using Phi29 p56 as an input. The secondary structure of Phi29 p56 is shown above its sequence. Percent identities with either Phi29-p56 or GA1-p56 are shown on the right. The bottom group includes sequences output from PSI-BLAST search only within *Salasmaviridae* genomes, using the VMY22 p56 sequence as input. These sequences share less sequence identity with the validated p56 sequences and their close homologs. Bold font in *Bacillus* phage name column indicates sequences tested via an UngIn assay in this study or previous studies.

3.5.2. Cloning, expression and Ung inhibition assay of putative p56 homologs

The *Bacillus* phage VMY22 gene “VMY22_4” (Gene ID: 26625151) was cloned into a pRSET-C expression vector by Ms Shelaine Fleck, a University of Alberta (Edmonton, Canada) Biochemistry 497 course unit summer intern, hosted by the Savva research lab. Any other laboratory work (i.e., apart from oligo design, synthetic gene sequence design, and experiment design) reported in this section was performed by Ms Rosalia Santangelo, formerly an undergraduate project student in the Savva research lab,

C-terminal truncated versions of the DK2 gene "DK2_00007" product (65aa out of 196aa; accession: AZU99760.1), the DK3 gene "DK3_00008" product (77aa out of 200aa; accession: AZU99806.1) and the vB_BthP-Goe4 gene "Goe4_c00070" product (114aa out of 217aa; accession: AYD87716.1) were cloned using OE-PCR/ligation (section 2.1.1.11), with primers P30-P35. The DK3 gene DK3_00008 contained a KEEKEEKEEKEEKEE motif, meaning 15 sequential codons (45 bases) that are solely formed by A and G bases. Synthetic gene block orders that contain such a high number of sequential A/G bases are rejected; therefore, this motif was not included in the synthetic gene sequence but was inserted later via iPCR/ligation to a cloned gene missing this motif, using primers P36 and P37. All the genes were expressed at small-scale (section 2.1.2.1). An UngIn assay validated for robustness of its readout in crude lysate environments (section 2.1.2.7.1), was used with these genes. The phage VMY22 “VMY22_4” p56 homolog inhibited Ung; however, the candidates cloned from DK2, DK3, and Goe4 phages did not appear to inhibit Ung activity (Figure 3.5.2).

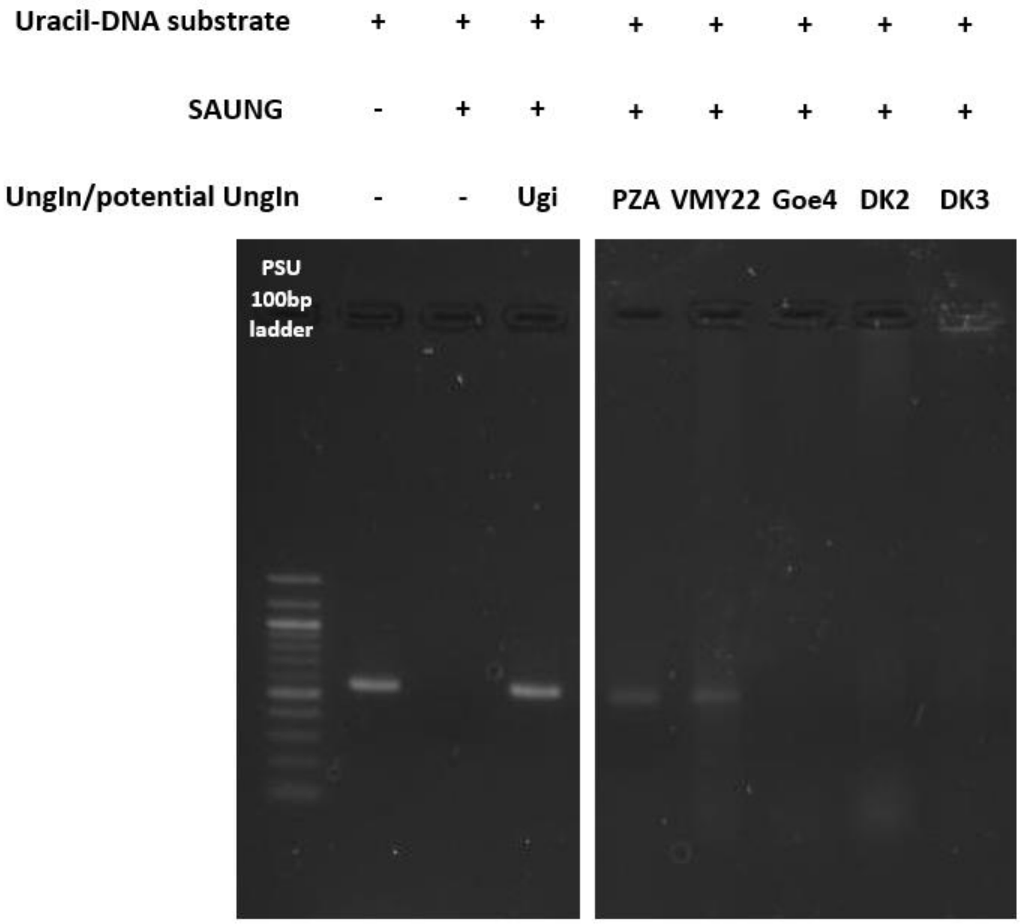

Uracil-DNA substrate + + + + + + + +
SAUNG - + + + + + + +
UngIn/potential UngIn - - Ugi PZA VMY22 Goe4 DK2 DK3
PSU
100bp
ladder

Figure 3.5.2. Visual U-DNA attrition assay of candidate p56-type Ung inhibitors. All lanes contained a uracil-DNA substrate treated with reaction buffer and conditions. SAUNG was added at [39.7 nM]. The UDG assay showed that Ugi, PZA p56, and VMY22 p56 are Ung inhibitors while the hits assayed from Goe4, DK2, and DK3 do not apparently inhibit Ung. This figure is formed of 2 cropped images. Original uncropped images are available and stored electronically. The data shown in this figure represents three independent replicates.

3.5.3. Genomic comparison of *Salasmaviridae* phages encoding sequences with homology to p56

A phylogenetic tree of *Salasmaviridae* Phi29-like phages splits this family into two distinctive branches (Figure 3.5.3a). Interestingly, one of the branches includes all the phages known to encode p56 (including VMY22), while the other includes sequences divergent from p56 in critical motifs involved in Ung inhibition (including DK2, DK3, and Goe4).

All the *Salasmaviridae* phages encode homologous DNA polymerase and homologous terminal protein sequences, and utilise a similar replication mechanism. There is a significant difference in the genome size of phages in different branches; p56 encoding phages have a genome size of 18,379-21,781bp, while the phages not known to encode p56 have a genome size range of 23,946-28,950bp (Table 3.5.3). The genomically larger phages have additional genes in the early gene region. All the p56 related sequences output from sequence similarity search (section 3.5.1) are found in the left early region. Genomic comparison of the Goe4 genome with the Phi29 genome reveals important differences (Figure 3.5.3b); Goe4, and all other phages within the same branch of *Salasmaviridae* encode a dUTPase gene, this gene has an important role in reducing the concentration of dUTP during phage replication and might provide an alternative strategy that these phages use to increase their replication efficiency by decreasing the possibility of uracil misincorporation into replication intermediates. It could be that these phages have lost the Ung inhibition function of their p56 related sequences after they have gained the alternative protective gene, dUTPase. Alternatively, phages known to encode p56 might have lost their dUTPase genes and then evolved via adaptation of p56 related sequences into functional UngIns.

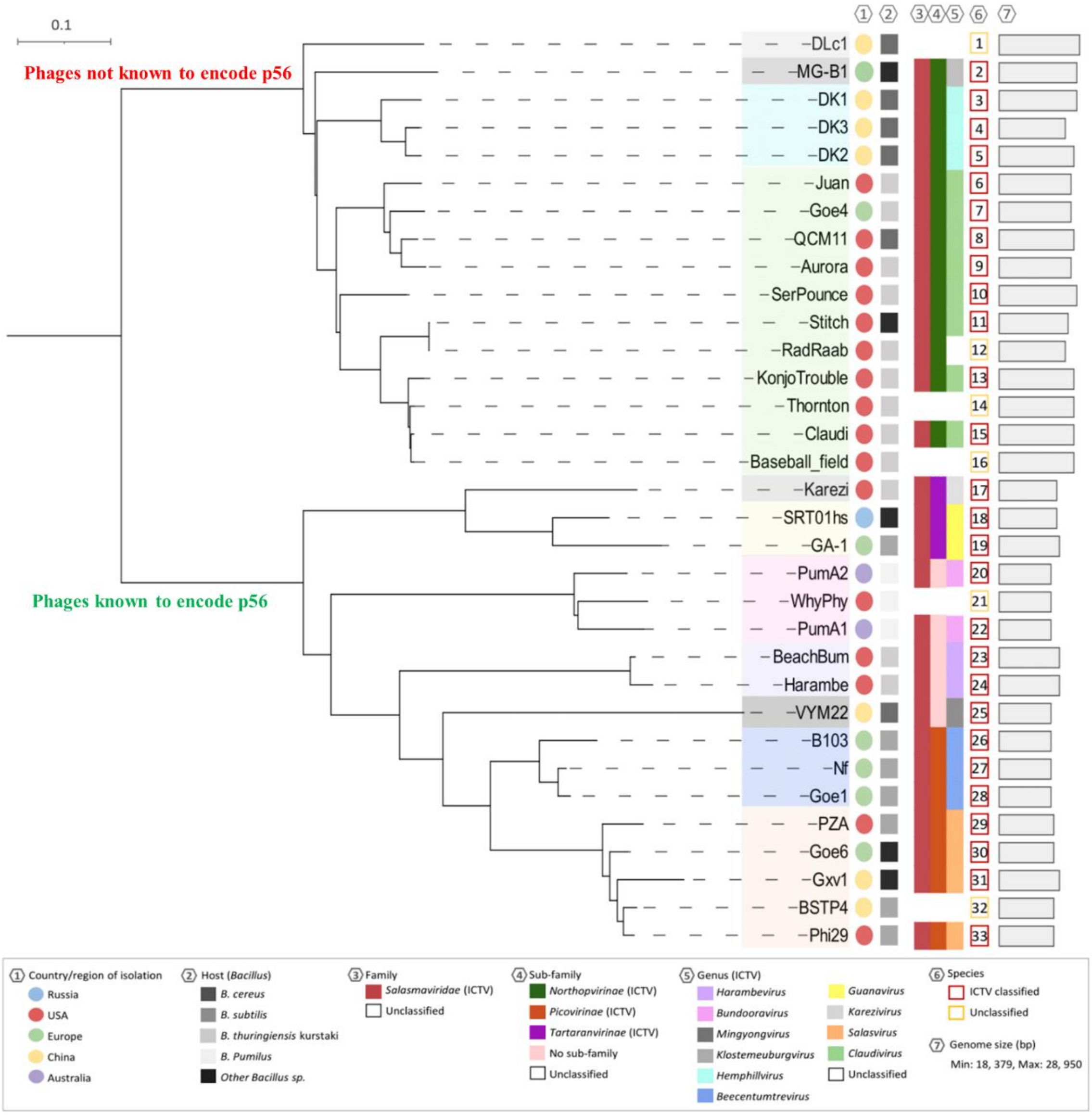


Figure 3.5.3a. Phylogenetic tree of *Salasmaviridae*. The clustering and colour coding correlate to the ICTV genus classifications. The main 2 branches separate phages known to encode p56 from phages not known to encode p56. (Modified from Stanton *et al*, 2021)[147].

Table 3.5.3. Genome sizes of *Salasmaviridae* phages infecting *Bacillus*.

	Species	Accession	Genome length (bp)
Phages known to encode p56	*Bacillus* virus Goe1	NC_049975	18379
	Bacillus virus PumA1	NC_049971	18466
	Bacillus virus VMY22	NC_028789	18609
	Bacillus virus B103	NC_004165	18630
	Bacillus phage WhyPhy	NC_055917	18642
	Bacillus phage Nf	NC_049976	18753
	Bacillus virus PumA2	NC_049972	18932
	Bacillus virus Goe6	NC_049965	19105
	Bacillus virus Goe6	MF407276	19105
	Bacillus phage BSTP4	MW354668	19145
	Bacillus virus phi29	NC_011048	19282
	Bacillus phage Arbo1	OL744111	19362
	Bacillus virus PZA	NC_001423	19366
	Bacillus phage BSTP6	MW354670	19367
	Bacillus phage Whiting18	MW477480	19548
	Bacillus virus Karezi	NC_049970	20083
	Bacillus virus SRT01hs	NC_049973	20784
	Bacillus virus BeachBum	NC_049961	21054
	Bacillus virus GA1	NC_002649	21129
	Bacillus virus Harambe	NC_049960	21684
	Bacillus virus Gxv1	NC_049974	21781
Phages not known to encode p56	*Bacillus* phage RadRaab	MF156580	23946
	Bacillus phage StevenHerd11	MK084630	23953
	Bacillus phage Ademby	OL744112	24162
	Bacillus virus Stitch	NC_031032	24320
	Bacillus virus Juan	NC_049963	25032
	Bacillus phage DLn1	MZ384014	25379
	Bacillus virus Goe4	NC_049966	25722
	Bacillus virus Aurora	NC_031121	25908
	Bacillus virus QCM11	NC_049959	26054
	Bacillus virus KonjoTrouble	NC_049964	26061
	Bacillus phage VioletteMad	MN082624	26061
	Bacillus phage Thornton	NC_055914	26319
	Bacillus virus DK2	NC_049968	26357
	Bacillus virus Claudi	NC_031015	26504
	Bacillus phage Baseball_field	NC_055905	26863
	Bacillus virus DK3	NC_049969	26865
	Bacillus virus DK1	NC_049967	27180
	Bacillus virus MGB1	NC_021336	27190
	Bacillus virus SerPounce	NC_049962	27206
	Bacillus phage DLc1	NC_055908	28950

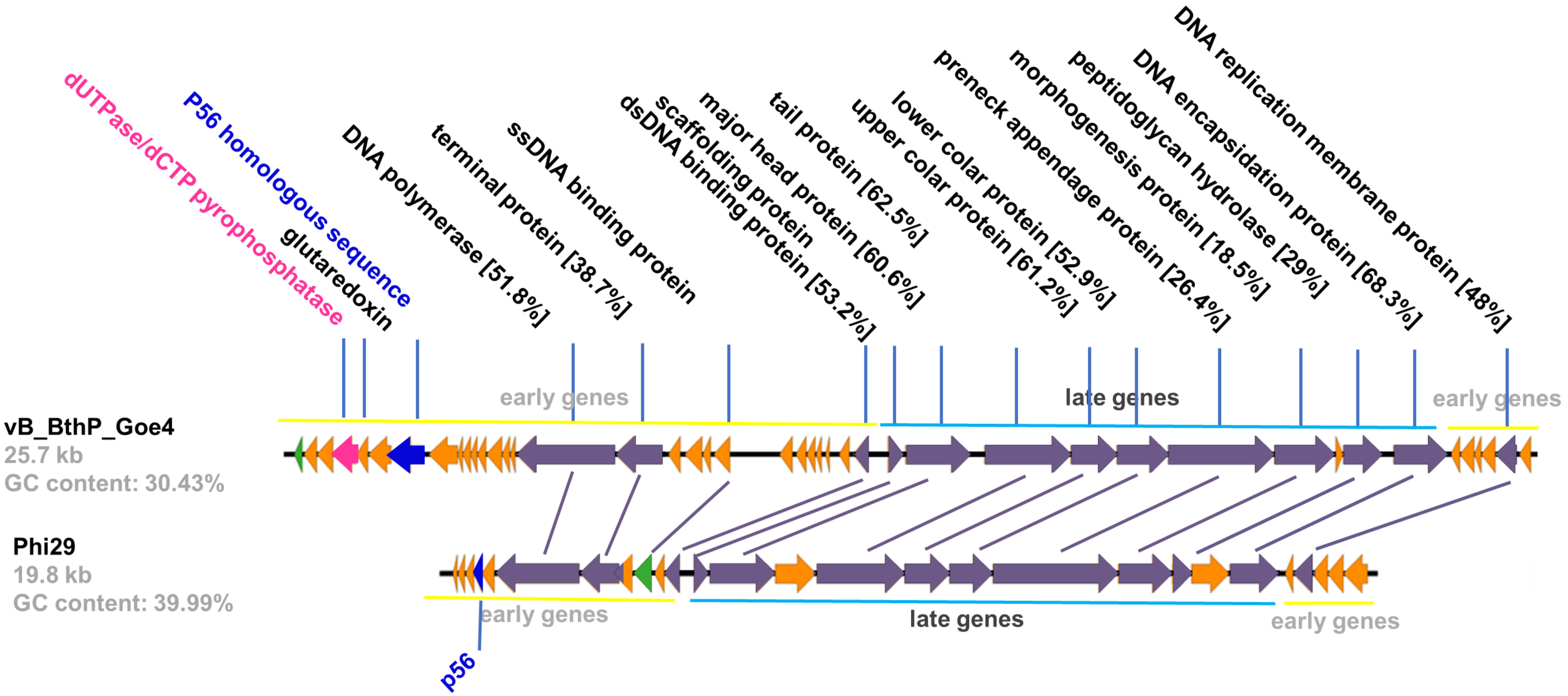

dUTPase/dCTP pyrophosphatase
P56 homologous sequence
glutaredoxin
DNA polymerase [51.8%]
terminal protein [38.7%]
ssDNA binding protein
dsDNA binding protein [53.2%]
scaffolding protein
major head protein [60.6%]
tail protein [62.5%]
upper colar protein [61.2%]
lower colar protein [52.9%]
preneck appendage protein [26.4%]
morphogenesis protein [18.5%]
peptidoglycan hydrolase [29%]
DNA encapsidation protein [68.3%]
DNA replication membrane protein [48%]
early genes
late genes
early genes
vB_BthP_Goe4
25.7 kb
GC content: 30.43%
Phi29
19.8 kb
GC content: 39.99%
early genes
late genes
early genes
p56

Figure 3.5.3b. A genomic comparison of vB_BthP_Goe4 and Phi29 phages. Genes are represented as arrows. Homologous proteins are depicted in purple. Both Phi29 p56 and the distant homologous sequence in Goe4 (blue arrows) are located in the left early gene region of their genomes. Interestingly, a dUTPase gene (shown in pink) is found in Goe4, DK2, DK3 and other *Salasmaviridae* genomes from that branch, but not in any phage encoding a p56 sequence.

3.6. Analysis of proteins with possible structural homology to UngIns

3.6.1. Ugi structural homologs

A structural homology search on the Dali server selecting a representative databank (PDB25) for similar structures to Ugi showed 47 hits with Z-score>4.0 (proteins with loosely similar structures to Ugi); nine of which have unidentified functions (Table 3.6.1). Of this shortened list, six proteins are <200 amino acids; 4 of them have RMSD values <3.5 Å (hits 2, 3, 15, and 28). Hit 3 (protein SSP0047) was identified previously as a *Staphylococcus aureus* Uracil-DNA glycosylase inhibitor[62]. Calculating the theoretical isoelectric point of the remaining hits using the ExPASy web server showed that hit 2 is a basic protein (pI: 9.24), while hits 15 (pI: 5.49) and 28 (pI: 3.96) are acidic proteins. Hit 15 is a protein called BACUNI_04723 encoded by *Bacteroides uniformis* (PDB code: 4MXT), Its exact function is unknown; however, it was defined as an outer-membrane lipoprotein carrier protein. A structural superposition for Ugi and protein BACUNI_04723 (Figure 3.6.1a) shows that despite some fold similarity that they share, protein BACUNI_04723 has 12 β-strands instead of five in Ugi, and four α-helices rather than two α-helices in Ugi. In addition, Ugi 1st β-strand (docks in Ung-DNA binding cleft) and 2nd α-helix (plays a main role in electrostatic binding to Ung) have no basic residues, while in the protein BACUNI_04723, the equivalent β-strand has 2 basic residues but no acidic ones (Figure 3.6.1a), and no equivalent structure to the 2nd α-helix of Ugi exists in protein BACUNI_04723. Therefore, protein BACUNI_04723 is unlikely to function as an Ugi and was not further investigated.

Table 3.6.1. Summary of Ugi hits from a Dali structural homology search. Dali server generated list of hits with z-score >4.0 are shown in descending order. Lali: length of alignment; nres: number of residues.

No:	Chain	Z	rmsd	lali	nres	%id	Description
1	1lqm-H	18	0.7	83	84	100	URACIL-DNA GLYCOSYLASE INHIBITOR
2	2yzy-A	7	2.6	62	163	15	PUTATIVE UNCHARACTERIZED PROT TTHA1012
3	2kcd-A	6	2.7	71	120	14	UNCHARACTERIZED PROTEIN SSP0047
4	4z48-A	6	3.3	65	240	8	UNCHARACTERIZED PROTEIN
5	6h6g-A	5	3	74	2139	11	TCDB2, TCCC3
6	5oa3-0	5	2.4	63	479	8	EUKARYOTIC TRANSLATION INITIATION FAC 2D
7	5mz9-A	5	3.3	62	887	10	MGP-OPERON PROTEIN 3
8	2mf7-A	5	3.9	72	127	11	MITOCHOND IMPORT INNER MEM TRANSLOCASE
9	4q9t-A	5	3.3	62	384	8	NUCLEOPORIN NUP133
10	5efv-B	5	3.7	66	635	9	PHI ETA ORF 56-LIKE PROTEIN
11	2fkj-A	5	6	55	361	2	OUTER SURFACE PROTEIN A
12	5kph-A	5	2.7	61	85	3	DE NOVO BETA SHEET DESIGN PROTEIN OR485
13	6p0x-A	5	2.8	67	320	13	SALMONELLA PLASMID VIRULENCE: SPVB
14	6hih-A	5	2.6	65	136	9	CYTOCHROME C
15	4mxt-A	5	2.8	66	187	9	UNCHARACTERIZED PROTEIN
16	6rh5-A	5	2.3	58	138	7	ADAPTIN EAR-BINDING COAT-ASSOCIA PROT 1
17	5f75-C	5	3.3	60	475	7	THIOCYANATE DEHYDROGENASE
18	6dlo-A	5	2.6	52	311	10	LEU-RICH REPEAT SER/THR-PROTEIN KINA
19	6eu4-A	5	3.7	62	587	10	TAIL SPIKE PROTEIN
20	2gu3-A	5	3.4	59	128	14	YPMB PROTEIN
21	2hye-A	5	3.1	61	1140	5	DNA DAMAGE-BINDING PROTEIN 1
22	4k90-B	5	2.4	57	207	2	EXTRACELLULAR METALLOPROTEINASE MEP
23	5yvq-A	5	4	53	358	8	TAIL FIBER PROTEIN S
24	4r1k-B	5	3.8	68	136	6	UNCHARACTERIZED PROTEIN
25	1bp1-A	5	3.2	68	456	6	BACTERICIDAL/PERMEABILITY-INCREAS PROTEI
26	4o9d-B	5	3.1	55	392	5	RIK1-ASSOCIATED FACTOR 1
27	5hal-A	4	3.5	55	103	4	UNCHARACTERIZED PROTEIN
28	1nnv-A	4	3.2	65	107	15	HYPOTHETICAL PROTEIN HI1450
29	1q7f-B	4	2.2	49	282	2	BRAIN TUMOR CG10719-PA
30	6mlt-A	4	3.7	60	603	12	HEMOLYSIN-RELATED PROTEIN
31	5nz7-A	4	2.9	58	984	9	CELLODEXTRIN PHOSPHORYLASE
32	4gl6-B	4	3.2	63	241	5	HYPOTHETICAL PROTEIN
33	2oaj-A	4	2.5	50	875	4	PROTEIN SNI1
34	1ei5-A	4	2.9	54	518	7	D-AMINOPEPTIDASE
35	1xs0-C	4	3.2	66	129	6	INHIBITOR OF VERTEBRATE LYSOZYME
36	6hiu-A	4	2.5	62	143	5	CYTOCHROME P460
37	3buu-A	4	2.7	62	224	2	HP LOLA SUPERFAMILY PROTEIN NE2245
38	3u1w-C	4	2.4	59	250	10	HYPOTHETICAL PERIPLASMIC PROTEIN
39	3f7w-A	4	3.1	58	288	5	PUTATIVE FRUCTOSAMINE-3-KINASE
40	2hq7-B	4	2.6	58	142	9	PROTEIN, RELATED TO GEN STRESS P 26
41	5ijn-E	4	3.2	61	1083	10	NUCLEAR PORE COMPLEX PROTEIN NUP155
42	4i0o-A	4	2.6	55	463	11	PROTEIN ELYS
43	5yx4-A	4	3.8	62	232	8	CHALCONE-FLAVONONE ISOMERASE FMLY
44	3bws-A	4	3	58	407	17	PROTEIN LP49
45	5cd6-A	4	2.8	64	575	8	TPR-DOMAIN CONTAINING PROTEIN
46	2qea-A	4	3.9	61	160	10	PUTATIVE GENERAL STRESS PROTEIN 26
47	5h3x-A	4	3.8	59	267	17	FIBRONECTIN/FIBRINOGEN BINDING PROTEIN

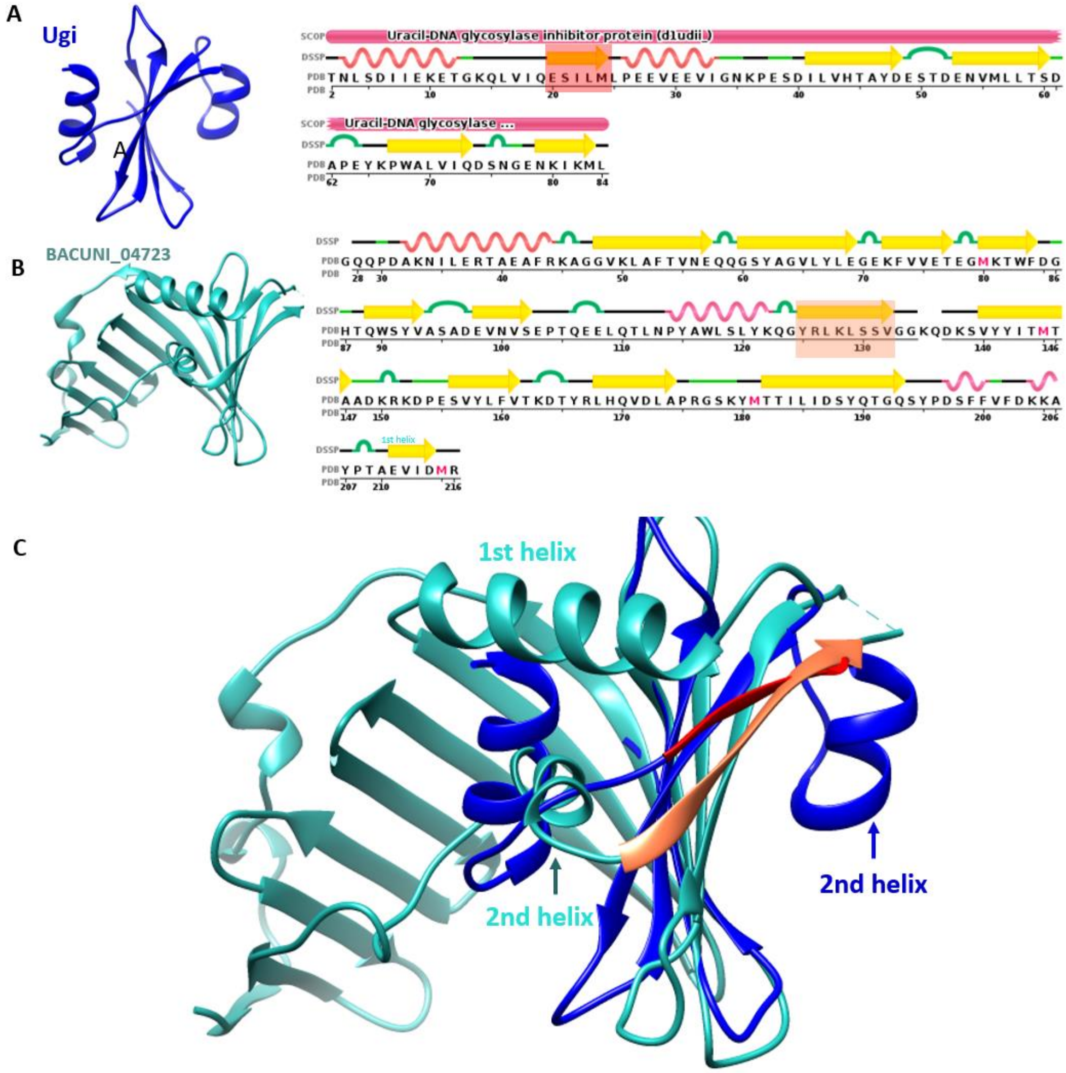

A
Ugi
SCOP
Uracil-DNA glycosylase inhibitor protein (d1udii_)
DSSP
PDB TNLSDIIEKETGKQLVIQESILMLPEEVEEVIGNKPESDILVHTAYDESTDENVMLLTSD
PDB 2 10 20 30 40 50 60
Uracil-DNA glycosylase ...
PDB APEYKPWALVIQDSNGENKIKML
PDB 62 70 80 84
B
BACUNI_04723
PDB GQQPDAKNILERTAEAFRKAGGVKLAFTVNEQQGSYAGVLYLEGEKFVVETEGMKTWFDG
PDB 28 30 40 50 60 70 80 86
PDB HTQWSYVASADEVNVSEPTQEELQTLNPYAWLSLYKQGYRLKLSSVGGKQDKSVYYITMT
PDB 87 90 100 110 120 130 140 146
PDB AADKRKDPESVYLFVTKDTYRLHQVDLAPRGSKYMTTILIDSYQTGQSYPDSFFVFDKKA
PDB 147 150 160 170 180 190 200 206
1st helix
PDB YPTAEVIDMR
PDB 207 210 216
C
1st helix
2nd helix
2nd helix

Figure 3.6.1a. Structural comparison between Ugi and protein BACUNI_04723. Rows A and B corresponds to 3D-structure and sequence aligned secondary structure of Ugi and BACUNI_04723, respectively. Row C shows superimposition of the 2 different structures. 1st β-strand of Ugi that docks in Ung-DNA binding cleft is coloured red in C and highlighted in A, the equivalent β-strand in BACUNI_04723 is coloured coral in C and highlighted in B. Difference in the acidic residues content of Ugi highlighted β-strand and both α-helices (8 acidic residues and 0 basic residues in these particular secondary structures) in comparison to BACUNI_04723 equivalent β-strand and 2 α-helices (2 acidic residues and 6 basic residues in these particular secondary structures) indicates poor similarity in the negative charge distribution, which means that BACUNI_04723 is unlikely to act as a DNA mimic protein and is unlikely to be an Ugi-like Ung inhibitor.

Hit 28 is a protein called HI1450 encoded by *Haemophilus influenzae* and several other bacterial species^{47}. HI1450 length (107 aa) is comparable to Ugi (84 aa) and its structural homolog SAUGI (112 aa). Calculated isoelectric point and molecular weight (Mw) of HI1450 (pI/Mw: 3.96/12524.93) are comparable to the ones for Ugi (pI/Mw: 4.13/9475.72) and SAUGI (4.51/13228.06). Interestingly, similar to the known Ung inhibitors, HI1450 is a dsDNA mimic protein with negative charge distribution pattern that mimics the dsDNA charge distribution[47]. Moreover, many crucial Ugi residues that perform binding to Ung located in the 1st β-strand (ESI motif) and 2nd α-helix (acidic rich motif EEVEE) are conserved in HI1450. Furthermore, an Ung residue (most commonly leucine, but rarely also phenylalanine or arginine) that has an important role in specifically stabilising the Ung-DNA pre-catalytic complex via its insertion into the minor groove is, except when it is an arginine, sequestered by the Ugi hydrophobic pocket. HI1450 has a hydrophobic pocket that aligns well structurally to the Ugi hydrophobic pocket, albeit smaller in size (Figure 3.6.1b). Sequence alignment of 58 aa residues in HI1450 (spanning from E48 to W105) with 54 aa residues in Ugi (spanning from E20 to I72) shows >50% conservation (Figure 3.6.1b). These multiple similar properties to Ugi lend greater rationale to further investigate HI1450 as a structural homolog to Ugi.

The *E. coli* version of HI1450 was amplified from NEB® 5-alpha competent *E. coli* cells using 1 µL of a chemically competent cell aliquot as a DNA template and using primers P38 and P39. The *E. coli* HI1450 was cloned into both the pRSET-C and pSDM4_U12_Ung vectors via OE-PCR/ligation (section 2.1.1.11). Protein expression yielded a weakly expressed protein with very low solubility (Figure 3.6.1c); assaying this soluble protein for Ung inhibitory character via agarose gel visualization of Ung treated uracil-DNA substrate (section 2.1.2.7.1) showed that HI1450 is apparently not able to inhibit Ung (Figure 3.6.1c).

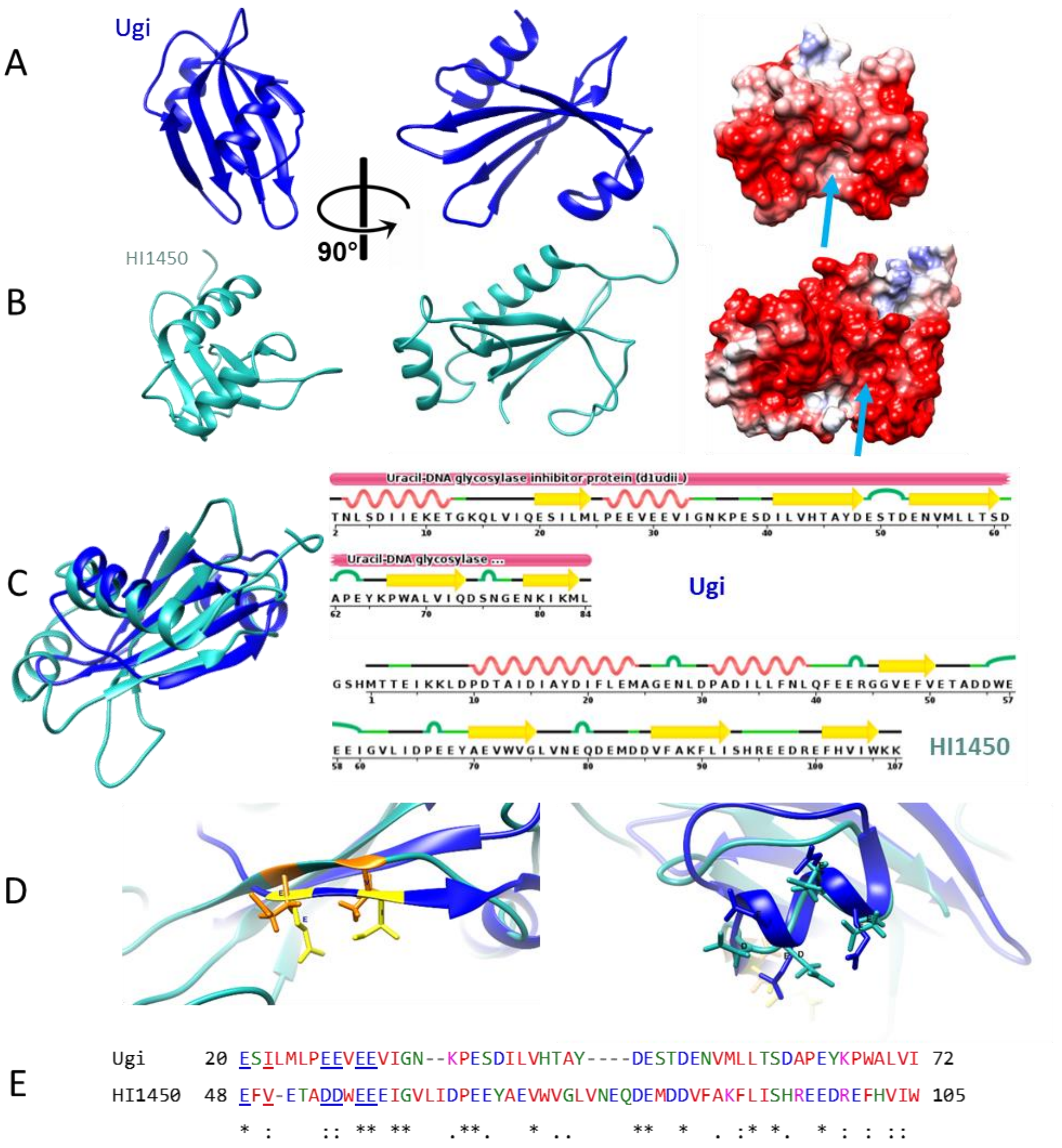

A
Ugi
90°
B
HI1450
C
Uracil-DNA glycosylase inhibitor protein (d1udii_)
TNLSDIIEKETGKQLVIQESILMLPEEVEEVIGNKPESDILVHTAYDESTDENVMLLTSD
2 10 20 30 40 50 60
Uracil-DNA glycosylase ...
APEYKPWALVIQDSNGENKIKML
62 70 80 84
Ugi
GSHMTTEIKKLDPDTAIDIAYDIFLEMAGENLDPADILLFNLQFEERGGVEFVETADDWE
1 10 20 30 40 50 57
EEIGVLIDPEEYAEVWVGLVNEQDEMDDVFAKFLISHREEDREFHVIWKK
58 60 70 80 90 100 107
HI1450
D
E
Ugi 20 ESILMLPEEVEEVIGN--KPESDILVHTAY----DESTDENVMLLTSDAPEYKPWALVI 72
HI1450 48 EFV-ETADDWEEEIGVLIDPEEYAEVWVGLVNEQDEMDDVFAKFLISHREEDREFHVIW 105
* : :: ** ** .**. * .. ** * . :* *. * : : ::

Figure 3.6.1b. **Structural comparison between Ugi and HI1450**. Rows A (Ugi) and B (HI1450) as ribbon representations on the left, and (rotated) in the middle to the front view (Ung binding interface) of Ugi and its equivalent pose in HI1450. The front view is repeated on the right with coulombic surface colouring (i.e., negative and positive charges are represented in red and blue, respectively); blue arrows are used to highlight the hydrophobic pocket of Ugi, and equivalent hydrophobic pocket of HI1450. Row C shows superimposition of the 2 different structures on the left, and sequence aligned secondary structure on the right. The 1st β-strand of Ugi that docks the Ung-DNA binding cleft includes a motif (ESI) that is conserved in Ugi/SAUGI. The E residue in this ESI motif is identical in all known Ugi/SAUGI homologues. The third residue of this ESI motif is a conserved hydrophobic residue in all Ugi/SAUGI homologues. Interestingly, the equivalent motif in HI1450 is EFV, the conserved residues of this motif are shown as sticks in row D in the left, and underlined in row E. The acidic residues of the Ugi 2nd α-helix and equivalent residues in the HI1450 loop are shown as sticks on the right in row D (underlined in row E). A pairwise sequence alignment of a 52 amino acid span of Ugi and its equivalent span in HI1450 shows 14/52 (27%) identity and 28/52 (54%) positivity between the 2 proteins. Combining all the above, HI1450 was considered a structural homolog of Ugi.

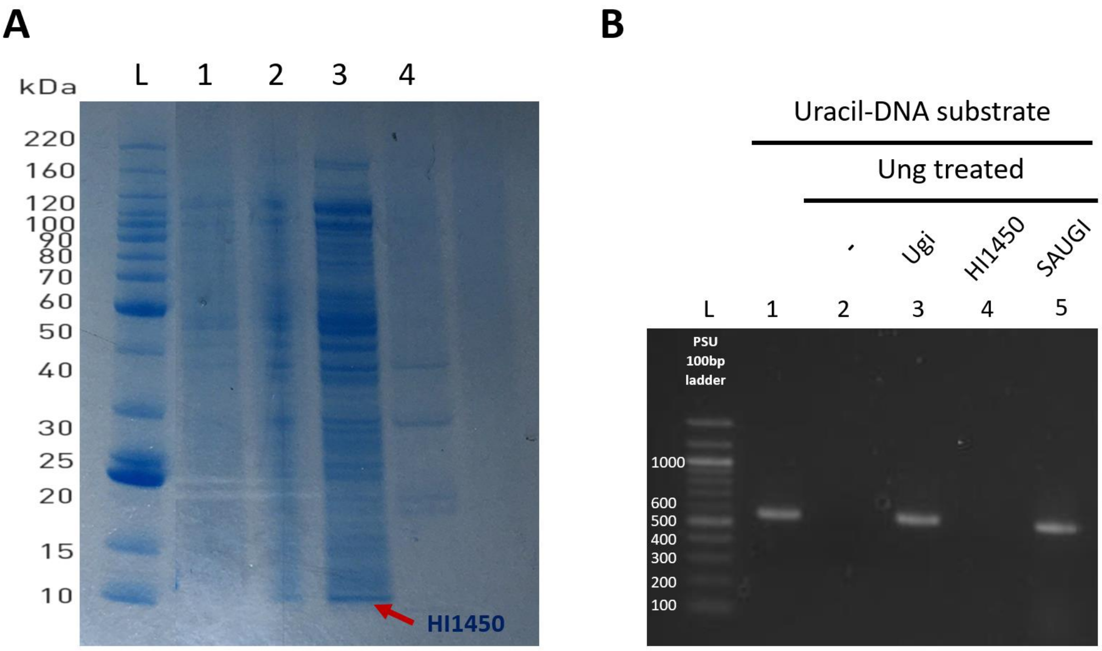


Figure 3.6.1c. HI1450 Ung inhibition assay. A) SDS-PAGE gel analysis for the HI1450. L: Benckmark^{TM} protein ladder (Invitrogen); 1: pre-induction sample; 2: cell harvest post induction; 3: clarified supernatant fraction; 4: insoluble fraction. B) Ung inhibition assay. L: DNA ladder; 1: untreated uracil-DNA substrate. 2: Ung-treated uracil-DNA substrate; 3: same as lane 2 including 1 µL of PBS1 Ugi [2.5 mg/mL]; 4: same as lane 2 with the addition of 1 µL HI1450-transformed expressed lysY/I^q cell lysate. 5: same as lane 2 with the addition of 1 µL SAUGI-transformed expressed *lysY/I*^{q} cell lysate. The agarose gel data represents three independent replicates.

Assaying HI1450 via the bacterial lethal assay (section 2.1.2.7.3) confirmed the *in vitro* assay results and HI1450 is unable to rescue the phenotype, thus appearing to lack any UngIn character. Superimposition of HI1450 and Ugi (Ugi-Ung complex) revealed steric hindrance that would prevent HI1450 binding Ung as Ugi does (Figure 3.6.1d).

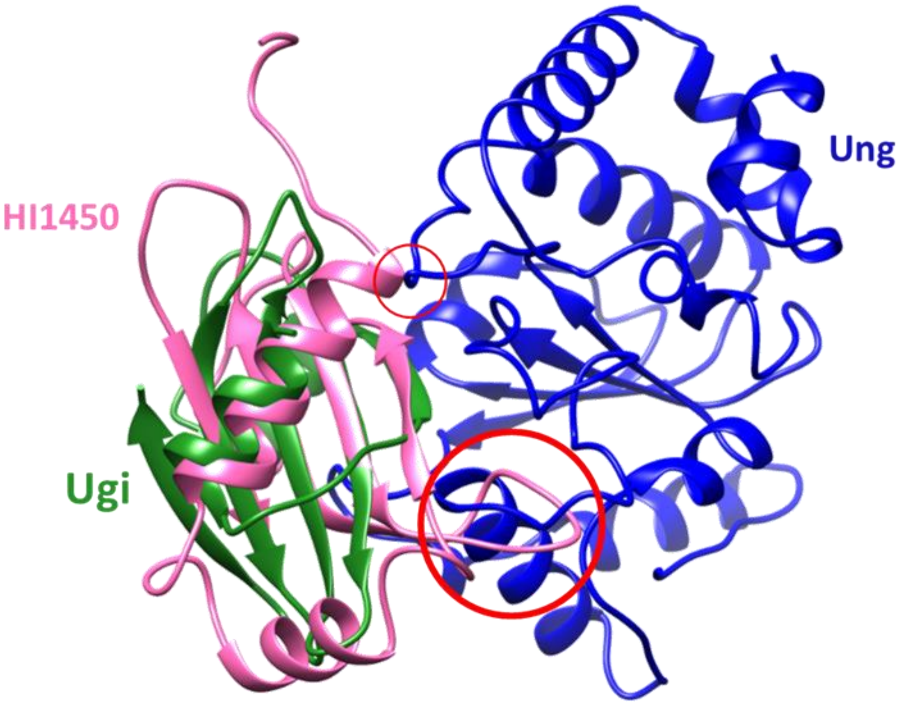


Figure 3.6.1d. Superposition of HI1450 (PDB: 1NNV) and Ugi in Ugi-Ung complex (PDB: 1UDI). Superposition shows that HI1450 would make clashes with Ung if overlaid with Ugi. One clash would happen at the N-terminus, and the other clash would happen at an extended 12aa loop of HI1450. Clashes are highlighted with red circles. Ung inhibition might be possible from a mutant lacking these features.

3.6.2. SAUGI structural homologs

A structural homology search for SAUGI showed 6 hits with Z-score>4.0, the first and second hits are an SAUGI variant, and Ugi. Only one of the other hits (hit 6) has no identified function (Table 3.6.2). This hit has RMSD value >3.5 Å; therefore, none of the hits met criteria for further investigation.

Table 3.6.2. Summary of SAUGI hits from a Dali structural homology search.

No	Chain	Z	rmsd	lali	nres	%id	Description
1	2kcd-A	15.9	1.9	110	120	94	UNCHARACTERIZED PROTEIN SSP0047
2	1lqm-H	6.4	3.1	74	84	15	URACIL-DNA GLYCOSYLASE
3	2fkj-A	5.0	4.7	65	361	11	OUTER SURFACE PROTEIN A
4	6kme-A	4.3	4.2	79	284	3	PHYTOCHROMOBILIN SYNTHASE
5	4wac-A	4.2	4.2	65	498	8	GLYCOSYL TRANSFERASE, GROUP 1 FAMILY PROTEIN
6	1yqf-A	4.2	4.4	63	182	6	HYPOTHETICAL PROTEIN LMAJ011689

3.6.3. Potential p56 structural homologs

A structural homology search for p56 showed 13 hits with Z-score>4.0; two of which (hit 4: protein TT1725 and hit 13: protein MTH889) are acidic proteins with <100 aa length and without an identified function (Table 3.6.3).

The dimeric p56 docks with the Ung DNA binding site via the alpha-helix of one subunit; its dimer interface constitutes a six-stranded beta-sheet on the opposite face.

Each monomer of TT1725 and MTH889 consists of 2 α-helices on one side and a β-sheet of 4 strands on the opposite side (Figure 3.6.3). Unlike p56, TT1725 protein has a positive charge on the equivalent face to the p56 Ung-binding interface. In addition, the hydrophobic core that performs hydrophobic sequestration of the Ung minor groove intercalation loop apical residue is absent in TT1725 protein (Figure 3.6.3). Therefore, TT1725 is unlikely to be a DNA mimic; it has no plausible similarity to p56. Protein MTH889 exists as a heptamer, its 3D-structure indicates that it is unlikely to act as a dsDNA mimic. Consequently, those potential structural homologs of p56 were not sufficiently convincing to be cloned and tested for Ung inhibition activity.

Table 3.6.3. p56 hits from a Dali structural homology search.

No:	Chain	Z	rmsd	lali	nres	%id	Description
1	2le2-A	7.2	1.5	48	56	90	P56
2	2pc6-A	5.6	2.0	48	164	8	PROBABLE ACETOLACTATE SYNTHASE ISOZYME III
3	1vg9-C	4.7	2.5	46	502	7	RAB GERANYLGERANYLTRANSFERASE
4	1j27-A	4.5	2.2	48	98	4	HYPOTHETICAL PROTEIN TT1725
5	5flg-A	4.4	3.2	48	253	6	6-CARBOXYHEXANOATE—COA LIG
6	5w0r-A	4.4	2.6	47	548	4	MBP FUSED ACTIVATION-INDUCED CYTIDINE DEAMINASE
7	3cj8-A	4.4	2.3	44	219	14	2,3,4,5-TETRAHYDROPYRIDINE-2,6-DICARBOXYLATE
8	4ney-B	4.4	2.9	49	173	8	ENGINEERED PROTEIN OR277
9	6c0f-S	4.4	2.2	48	171	19	SACCHAROMYCES CEREVISIAE S288C 35S PRE-RIBOSOMAL
10	2lmc-A	4.3	2.3	46	59	11	BACTERIAL RNA POL- INHIBITOR
11	1b33-N	4.3	2.1	46	67	7	ALLOPHYCOCYANIN, ALPHA CHAIN
12	1q5y-D	4.3	1.9	47	80	6	NICKEL RESPONSIVE REGULATOR
13	2raq-B	4.1	2.1	49	94	12	CONSERVED PROTEIN MTH889

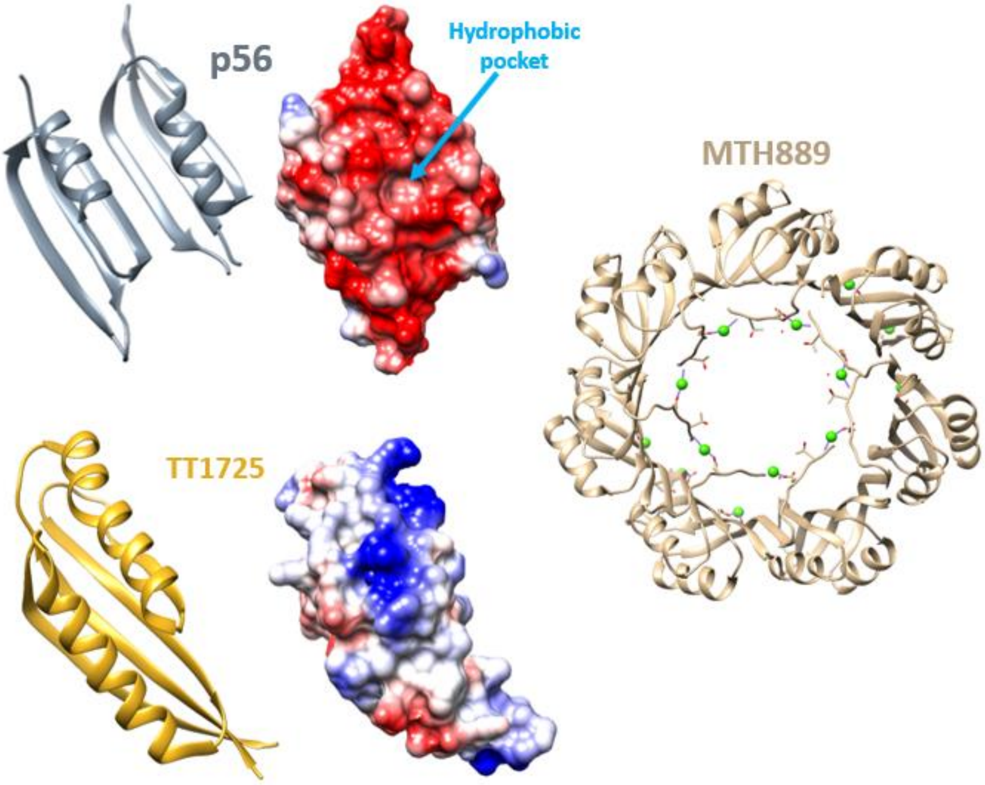


Figure 3.6.3. Left (top and bottom) Structural comparison of p56 with a structural homolog TT1725. The comparison reveals that TT1725 lacks both the negative charge and the hydrophobic pocket of p56. Protein MTH889 on the right side of the figure forms a multimeric structure of 7 monomers and is not a structural homologue of p56.

3.6.4. Vpr structural homologs

The Dali server's generated list of Vpr structural homologs with Z-score>4.0 contained 5 proteins with unidentified function (Table 3.6.4); two of which (hit 23: protein VNG1086C, and hit 39: protein LPG2271) are acidic proteins with <100 aa length and with RMSD <3.5 Å. The Vpr secondary structure consists of three α-helices connected by two loops. A hydrophobic pocket that Vpr uses to sequester the protruding residue of Ung minor groove intercalation loop is formed by 2 residues of the 2nd α-helix and 2 residues of the 3rd α-helix. Protein VNG1086C has 4 α-helices; superposition with Vpr shows that VNG1086C has a significantly shorter 2nd α-helix and does not have an equivalent hydrophobic core (Figure 3.6.4). Consequently, there is poor probability for VNG1086C to act as an Ung inhibitor.

Protein LPG2271 has 4 α-helices; superposition with Vpr shows that the LPG2271 surface that is equivalent to the Vpr Ung-binding interface shows very poor negative charge. Consequently, LPG2271 is unlikely to act as a dsDNA mimic protein, and it is unlikely that it would act as an Ung inhibitor. Interestingly, among all results generated by the Dali server upon searching for potential Ung inhibitor structural homologs, the 1st hit in Vpr search results has the highest Z-score (8.9), even though it shares <20% identity with Vpr; this hit protein is called Vpx. The Vpx structure is very similar to Vpr structure. Vpx Is known to bind the protein DCAF1 (a protein that Vpr is also known to bind)[149]. Vpr binds DCAF1 via one of its surfaces and binds human UNG2 via the opposite surface. Vpx binds DCAF1 via one of its surfaces, but binds sterile α motif (SAM) and histidine/aspartate (HD)-containing protein 1 (SAMHD1) via the opposite surface[150]. Intriguingly, Vpx and Vpr share considerable sequence similarity and are juxtaposed in the HIV-2/SIVsm genomes[149]. Vpx superposition with Vpr shows that Vpx has significantly weaker negative charge on the equivalent interface of Vpr Ung-binding interface; in addition, Vpx lacks the hydrophobic pocket that Vpr uses to inhibit UNG2 (Figure 3.6.4). The absence of Vpr Ung-binding characteristics in Vpx indicates a poor probability of Vpx to act as Ung inhibitor. Several previous studies demonstrated that Vpx does not bind UNG2[149,151].

Table 3.6.4. Summary of Vpr hits from a Dali structural homology search.

No:	Chain	Z	rmsd	lali	nres	%id	Description
1	4cc9-B	8.9	2.1	74	98	19	PROTEIN VPRBP;
2	6cgh-A	5.7	3.3	61	89	11	DNAJ HOMOLOG SUBFAMILY C MEMBER 2;
3	4bjm-A	5.6	6	70	226	10	AVRM;
4	6agb-B	5.5	12.7	57	784	12	CHROMOSOME V, COMPLETE SEQUENCE;
5	4bpm-A	5.3	9.6	69	162	7	PROSTAGLANDIN E SYNTHASE, FUSION PEPTIDE;
6	4noo-B	5.2	2.6	63	95	6	VGRG PROTEIN;
7	2x0s-A	5.2	6.4	72	899	8	PYRUVATE PHOSPHATE DIKINASE;
8	3ljc-A	5	2.7	57	239	7	ATP-DEPENDENT PROTEASE LA;
9	4heo-B	5	2.7	54	60	7	PHOSPHOPROTEIN;
10	6icz-A	4.8	7.7	66	2253	11	PROTEIN MAGO NASHI HOMOLOG 2;
11	6rjw-A	4.8	3.2	56	161	11	LYSM DOMAIN PROTEIN;
12	5f3o-A	4.7	3.1	57	194	2	EhRNaseIII229;
13	2dd4-H	4.6	9.2	63	156	5	THIOCYANATE HYDROLASE ALPHA SUBUNIT;
14	4wat-A	4.6	2.9	61	332	3	PFRH5;
15	6jpu-A	4.5	9.2	53	577	17	UNCH AAA DOMAIN-CONTAINING PROTEIN C31
16	5y6o-D	4.5	3.3	56	108	9	DEATH DOMAIN-ASSOCI PROTEIN 6,TRANSCRIPTIONAL
17	6uxe-B	4.5	2.7	57	85	7	CYSTEINE DESULFURASE, MITOCHONDRIAL;
18	1oks-A	4.5	2.6	52	53	6	RNA POLYMERASE ALPHA SUBUNIT;
19	6iz2-A	4.4	4	61	145	8	DINB/YFIT FAMILY PROTEIN;
20	4x8d-A	4.4	2.4	56	429	5	SULFOXIDE SYNTHASE EGTB;
21	2hh6-A	4.4	6.2	66	112	6	BH3980 PROTEIN;
22	1gvn-A	4.4	3.1	57	87	9	EPSILON;
23	2gf4-A	4.3	2.5	50	89	14	PROTEIN VNG1086C;
24	6fv7-A	4.3	3.2	62	421	10	AQ128;
25	5oqk-A	4.3	3.4	63	147	5	VOLTAGE-GATED HYDROGEN CHANNEL 1;
26	3gi7-A	4.3	3.2	63	103	10	SECRETED PROTEIN OF UNKNOWN FUNCTION DUF1311;
27	5d92-B	4.3	2.9	58	342	7	AF2299 PROTEIN,PHOSPHATIDYLINOSITOL SYNTHASE;
28	4akk-A	4.3	11.5	65	368	17	NITRATE REGULATORY PROTEIN;
29	4oe8-C	4.2	3.9	60	87	5	INTERLEUKIN-12 SUBUNIT BETA;
30	6m9k-D	4.2	2.4	51	67	20	EXONUCLEASE;
31	4w8f-B	4.2	8.6	67	2609	4	DYNEIN HEAVY CHAIN LYSOZYME CHIMERA;
32	6gdj-A	4.2	3.6	57	71	7	MTO2;
33	5yck-A	4.2	12.9	68	449	9	MULTI DRUG EFFLUX TRANSPORTER;
34	2efl-A	4.2	3.5	63	281	3	FORMIN-BINDING PROTEIN 1;
35	5vxa-A	4.2	2.6	54	178	11	G-3',5'-BIS(DIPHOSPHATE)3' PYROPHOSPHOHY
36	4myy-B	4.2	3.7	60	84	3	CURG, CURH FUSION PROTEIN;
37	6ny2-Y	4.2	3.9	71	915	8	DNA TARGET STRAND;
38	2es4-E	4.2	5.5	48	278	13	LIPASE;
39	5l1a-A	4.2	3.2	58	108	9	UNCHARACTERIZED PROTEIN;
40	1r5i-D	4.1	3.3	60	214	13	HLA CLASS II HISTOCOMPATIBILITY ANTIGEN, DR
41	3bbz-A	4.1	2.8	48	48	10	P PROTEIN;
42	5h79-D	4.1	2.9	52	142	4	IMMUNOGLOBULIN G-BINDING PROTEIN A;
43	1pc6-A	4.1	11	64	141	6	PROTEIN NINB;
44	5lfj-B	4.1	3	55	113	16	BACTERIAL PROTEASOME ACTIVATOR;

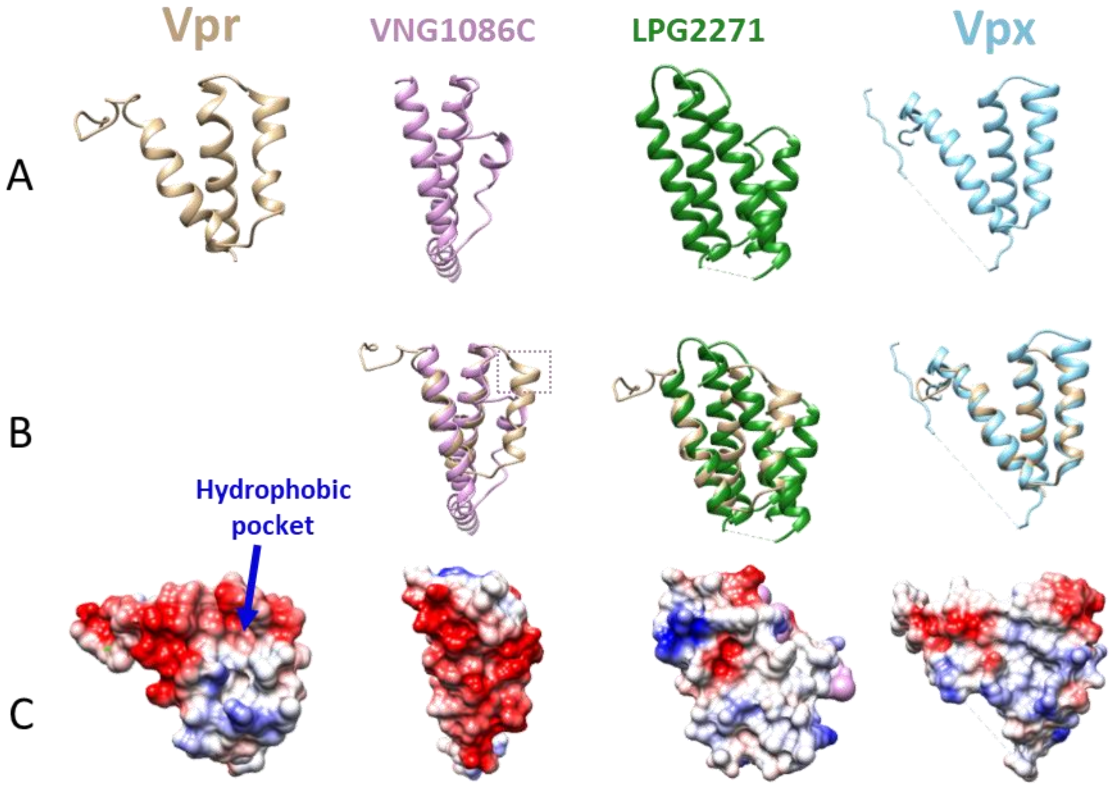


Figure 3.6.4. Structural comparison of Vpr with its potential homologs. A) The ribbon structure view of Vpr and its potential homologs. B) The superposition of Vpr with its potential homologs. C) The coulombic surface colouring view of Vpr and its potential homologs at same orientation shown in (A) and (B). Vpr sequesters the UNG2 essential minor groove intercalation loop protruding residue via a hydrophobic pocket that is formed by residues located in 2 α-helices. An essential part of one of Vpr α-helices that contributes in forming of that hydrophobic pocket does not have an equivalent structure in VNG1086C (dashed line square in row B); thus, an equivalent hydrophobic pocket is absent in VNG1086C. LPG2271 lacks the negative charge that Vpr exhibits on the equivalent (superimposable) interface. Vpx has significantly weaker negative charge on the superimposable interface; in addition, Vpx lacks the hydrophobic pocket that Vpr uses to inhibit UNG2

3.7. Conclusion

It is argued in this chapter that the sequence plasticity of Ung inhibitors has hindered unambiguous identification of novel proteins that are able to inhibit Ung.

In this chapter it is shown that for discovery of new UngIns, combining phylogeny-guided searches with genomic map analysis supported with insights of signature motif conservation seems to be a much more powerful approach than searching by sequence similarity alone.

There are highly divergent sequences annotated as uracil-DNA glycosylase inhibitors within sequence databases, which could encode an Ugi (*Bacillus* phage vB_BpuM-BpSp), p56 (*Bacillus* phage VMY22), or SAUGI. New UngIn homologs have been discovered during the present study by applying the approaches presented.

In this thesis, we have developed an UngIn selection conditional lethal assay (section 3.1). This assay allows scanning library fragments of genomes known or expected to encode UngIn sequences (chapter 5). Novel synthetic Ugi variants were identified in this thesis (section 3.2), which increased our knowledge of tolerated mutation and were used for PHI-BLAST searches in targeted organisms (chapter 5).

The protein Ac950 is annotated as SAUGI; however, our conditional lethal assay data suggests that it doesn't act as an UngIn (section 3.4.6): Thus, protein annotations might be misleading and empirical tests are required. Signature motif conservation is more important than sequence identity to retain the UngIn functionality; proteins MCUGI, MBUGI, SYUGI, and JMUGI share with SAUGI an ID% similar to the ID% that Ac950 shares with SAUGI, but importantly retain signature motif conservation, which probably underlies the reason that they are acting as UngIns (section 3.4.4).

Some phages in the *Salasmaviridae* family are known to encode UngIns, suggested to protect single strand intermediates formed during protein-primed DNA replication of these phages[58,60,84]. However, despite similarity in encoded sequences in the closely related genomes of *Northropvirinae* viruses, including *Bacillus* phages vB_BthP-Goe4, DK1, DK2, DK3, and DLc1[96,97,148], the present study suggests no Ung inhibitor is present (section 3.5). Differences observed in these sequences appear in motifs crucial to Ung inhibition. The genomes in question are also found to encode a dUTPase, which may compensate for the lack of an Ung inhibitor.

Chapter 4

4. Structural tolerance of sequence plasticity in UngIns

4.1. Introduction

All the known UngIns are DNA mimic proteins, they belong to three architecturally discrete families converging upon a universal mechanism of Ung inhibition via hydrophobic sequestration of an essential residue for Ung catalytic activity[31,58]; and even within a single architecture there is considerable sequence diversity (as verified in Chapter 3 of this thesis).

It is known that for most protein families, protein structures are more conserved than sequences[99]. It is possible, given their strategic utility to viruses, that UngIn sequence variants that have a known UngIn fold but sequences lacking sufficient homology for detection using current bioinformatics approaches may exist in other annotated genomes.

The aim of this chapter is to uncover a sequence syntax underlying structure conservation among UngIns. Structural analysis was performed on sequence divergent examples of Ugi, SAUGI and p56 to better understand the limits of tolerance in sequence variation supporting

UngIn function. Structural insights can be recruited as building blocks for UngIn-specific heuristic approaches to find novel naturally encoded UngIn variants.

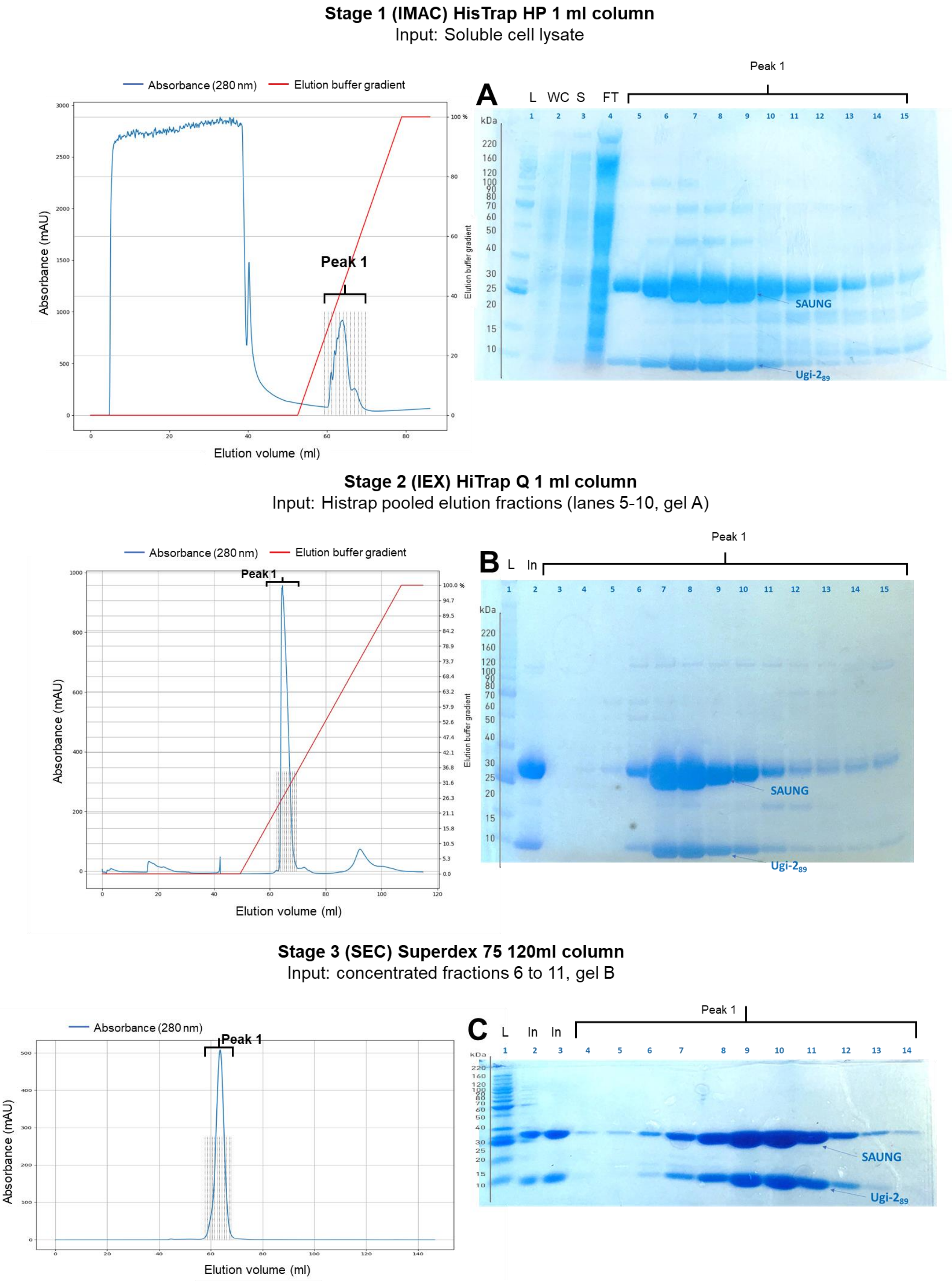
4.2. Structural analysis of Ugi-2

Two Ugi-2 variants, Ugi-2_{89} and Ugi-2_{108}, were validated as UngIns in this thesis (section 3.3.4). The variant Ugi-2_{89}, however, showed higher stability (section 3.3.4) and hence was chosen for further structural characterisation.

4.2.1. Large-scale expression and purification of Ugi-2_{89} in complex with SAUNG

Ugi-2_{89} and SAUNG were expressed from constructs pRSCUgi-2_{89} (section 3.3.2) and pRS-SAUNG6xH (A construct encoding a C-terminal hexahistidine-tagged SAUNG; previous work of Dr Claire Bagneris) separately at large-scale as described (Section 2.1.2.1). Cell pellets were resuspended in buffer A (Table 2.1.2.5); resuspended cells were mixed prior to performing cell lysis (Section 2.1.2.2). The mixed cell extract was sequentially purified by IMAC, anion IEX and finally SEC (Figure 4.2.1).

SEC elution fractions with peak absorbance, and high purity as validated by SDS-PAGE (Figure 4.2.1), were concentrated at 3200 x*g* using an Ultracel 3K centrifugal filter (Millipore) to reach a protein concentration of 15 mg/ml then stored at 4°C.



Stage 1 (IMAC) HisTrap HP 1 ml column
Input: Soluble cell lysate

Stage 2 (IEX) HiTrap Q 1 ml column
Input: Histrap pooled elution fractions (lanes 5-10, gel A)

Stage 3 (SEC) Superdex 75 120ml column
Input: concentrated fractions 6 to 11, gel B

Figure 4.2.1. Purification of Ugi-2$_{89}$ and C-terminal histidine-tagged-SAUNG proteins. Chromatograms and SDS-PAGE gels are shown for IMAC, IEX, and SEC stages. Gel lanes contain the following samples: **Stage 1**: L= Benchmark ladder; WC= Mixed whole cells; S= Mixed cell extract (soluble fraction); FT= HisTrap flow through; Peak 1 (5-15) = Elution fractions with peak absorbance. **Stage 2**: L= Benchmark ladder; In= injected sample (HisTrap pooled elution fraction lanes 5 to 10, gel A); Peak 1(3-15) = Elution fractions with peak absorbance. **Stage 3**: L= Benchmark Ladder; In (2-3) = Injected concentrate (concentrated fractions 6 to 11, gel B); Peak 1 (4-14) = Elution fractions with peak absorbance from SEC. Fractions labelled 8-11 on gel C were pooled and concentrated for crystal screening.

4.2.2. Crystallisation of Ugi-2_{89} in complex with SAUNG

Purified protein sample was used at 15 mg/ml in commercial crystallisation screens (JCSG-Plus and PACT premier; Molecular Dimensions)[152]. Sitting drops were prepared at both 1:1 and 2:1 ratios of protein to mother liquor with volumes of 75 nl and 100 nl protein, respectively; plates were incubated at 16°C. Four conditions were selected for the observed presence of potentially crystalline material (Figure 4.2.2). Putative protein crystals from JCSG-Plus condition D5 were flash-frozen in liquid nitrogen without adding cryoprotectant as the mother liquor of this condition includes MPD, which can act as a cryo-protectant; putative protein crystals from other conditions were cryoprotected in corresponding mother liquor solutions supplemented with 20% (v/v) ethylene glycol. Crystals from JCSG-Plus A5 and D5 conditions exhibited a flat rectangular plate morphology and crystals from PACT premier A2 and A3 conditions exhibited needle-like morphology (Figure 4.2.2).

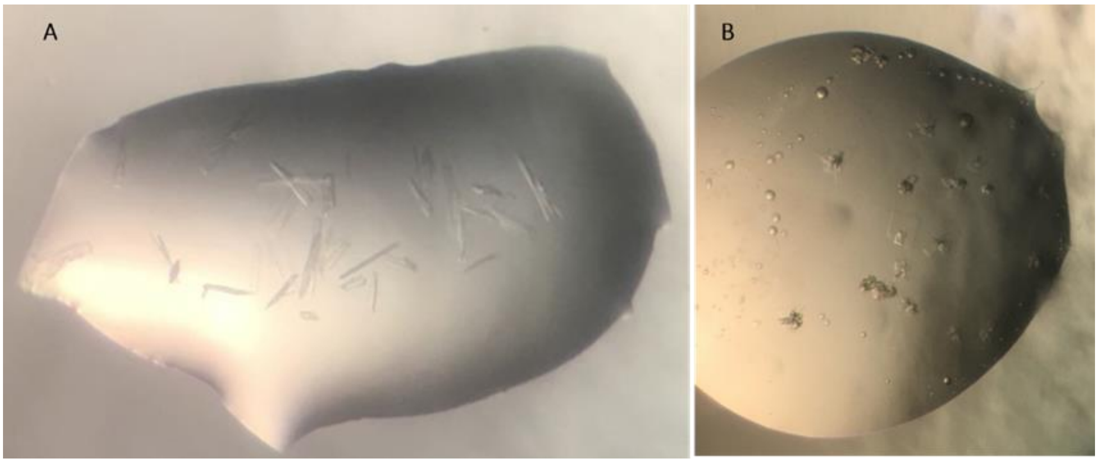


Figure 4.2.2. Putative protein crystals of Ugi-2_{89} in complex with SAUNG. (A) Putative protein crystals formed in drop A3 of PACT premier^{TM} (Molecular Dimensions), conditions are 0.1 M SPG buffer*, pH 6.0, 25% w/v PEG** 1500. Putative protein crystals with similar morphology were also observed in drop A2 of the same screen, whose conditions are similar to the conditions of A3 drop except that the pH in drop A2 is 5.0. (B) Putative protein crystals formed in drop D5 of JCSG-Plus^{TM} (Molecular Dimensions), conditions are: 0.1 M HEPES^{+}, pH 7.5, 70% v/v MPD^{++}. Putative protein crystals with similar morphology were also observed in drop A5 of the same screen, whose conditions are: 0.2 M magnesium formate dihydrate, 20% (w/v) PEG 3350. *SPG buffer: succinic Acid, sodium phosphate monobasic monohydrate, glycine. **PEG: polyethylene glycol. $^{+}$HEPES: 2-(4-(2-Hydroxyethyl)-1-piperazinyl)ethanesulfonic Acid. $^{++}$MPD: 2-Methyl-2,4-pentanediol.

4.2.3. Structure determination of Ugi-2_{89} in complex with SAUNG

X-ray diffraction data was collected on beamline IO4 at the Diamond light source facility with the help of Dr Claire Bagneris. The xia2 pipeline was used for automated data processing using XDS; AIMLESS was used to scale the structure factors. The structure was initially phased by molecular replacement with MOLREP[115] using Chain A from PDB accession 3WDG [the apo form of SAUNG][62] and chain I from PDB accession 1UDI [the apo form of PBS1 Ugi, 32% ID with Ugi-2_{89}][22] as search models. The structure was solved at a resolution of 2.6 Å. Structure refinement used REFMAC5[118], and model building was performed using the program COOT[117]. Statistics for data collection and refinement are summarised in Table 4.2.3. This model has been deposited in the protein data bank (PDB ID: 8AIM) and is currently on hold for release (for validation metrics please see appendix C).

Table 4.2.3. Crystallographic data collection and refinement statistics for Ugi-2:SAUNG complex crystals

Data collection	
Wave length (Å)	0.9795
Space group	C 2 2 21
Unit cell a, b, c (Å)	137.65, 142.82, 82.76
α, β, γ (˚)	90.00, 90.00, 90.00
Resolution (Å)	52.92-2.6 (2.72-2.6)
Unique reflections	25462 (3052)
Redundancy	1.9 (1.9)
Completeness (%)	100 (100)
I/σ(I)	1.64
R_{merge}	0.121 (0.588)
Refinement	
R_{work} (%)	0.202
R_{free} (%)	0.246
Bond RMSD (Å)	0.007
Angle RMSD (˚)	1.469
Mean B value/no of atom	46/9750
Ramachandran plot (%)	
Most favoured (%)	95.29
Allowed (%)	4.71
Outliers (%)	0.00

4.2.4. DNA mimicry in the Ugi-2$_{89}$ and Ugi structures

The inhibitor protein structure exhibits a high degree of structural homology (RMSD=1.343 Å; 32% ID) to other deposited PBS1 Ugi structures from the Protein Data Bank (PDB). The structure of Ugi-2$_{89}$ includes 5 antiparallel β-strands that form a highly twisted β-sheet sandwiched between 2 short α-helices (Figure 4.2.4).

PBS1 Ugi is known to mimic Ung-bound DNA. The twisted β-sheet of Ugi mimics the bent shape of the Ung-bound DNA, Ugi-2$_{89}$ β-sheet exhibits a similar twisting pattern. A structural comparison of the SAUNG-bound Ugi-2$_{89}$, HSV1UNG-bound PBS1 Ugi, and hUNG-bound DNA demonstrates that both Ugi variants possess striking overall similarities with DNA. The negative charge distribution on the surface of each Ugi variant mimics the negative charge distribution of phosphate groups on Ung-bound DNA. In addition, the grooves of DNA are mimicked by grooves on the surface of each Ugi variant (Figure 4.2.4).

As in PBS1 Ugi, Ugi-2$_{89}$ β-strand 1 has distortions along its edge that enable it to follow the path of the compressed DNA first strand backbone more closely. The second strand of DNA can also be traced within the Ugi-2$_{89}$ structure by the loop between 3rd and 4th β-strands, the edge of β-strand 4, and the loop between the 2nd and 3rd β-strand (Figure 4.2.4).

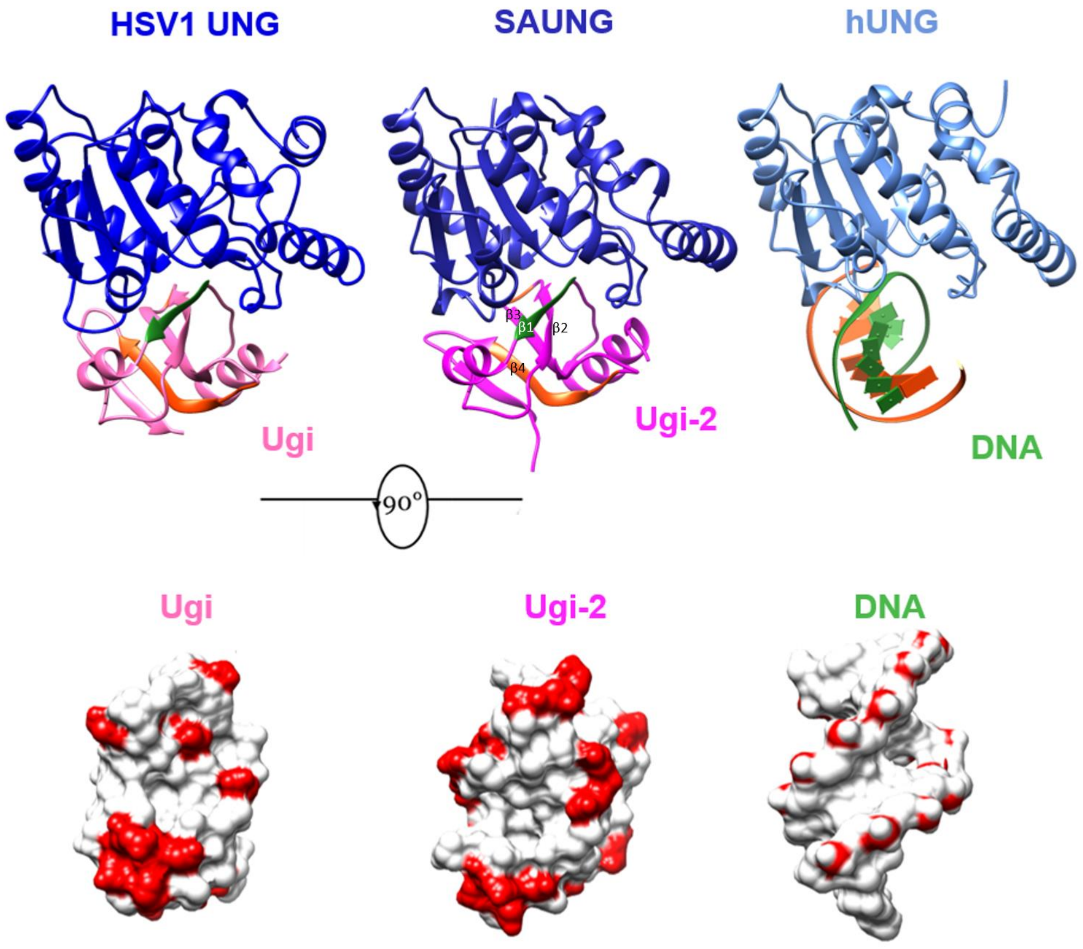


Figure 4.2.4. DNA mimicry of Ugi variants. Top row left to right: The structure of HSV1-UNG with PBS1 Ugi (PDB: 1UDI), the structure of SAUNG with Ugi-2$_{89}$ presented in this thesis, and the structure of hUNG with dsDNA (PDB: 1SSP). DNA backbone 1st strand (coloured green) is mimicked by the 1st β-strands (coloured green) of Ugi variants β-sheet. DNA backbone second strand (coloured orange) can also be traced within both Ugi variants by multiple secondary structures (see main text) coloured orange in both Ugi variants. Bottom row: the Ung-binding interfaces of PBS1 Ugi, Ugi-2$_{89}$, and dsDNA. Bottom row depictions were generated by rotating top row view by 90° degrees, removing Ung chains for clarity, then showing the surfaces of Ung-bound molecules. The red colour in bottom row corresponds to the acidic residues of Ugi variants and the phosphate groups of dsDNA. A similar pattern of negative charge distribution can be observed in the three different molecules.

4.2.5. Negative charge conservation in Ugi variants

Ugi-2$_{89}$ has 21 acidic residues in its primary sequence, while PBS1 Ugi has 18 acidic residues. There are only 8 positions of conserved negatively charged residues between the two Ugi sequence variants; of which, 5 are located in the Ung-binding β-strand 1 and α- helix 2 in a

12aa span (Figure 4.2.6, panel A). The PROSITE pattern of these acidic residues, E-X(6)-[ED]-[ED]-X-[ED]-[ED], could be employed in sequence similarity search using PROSITE patterns to uncover potential novel naturally occurring Ugi variants.

4.2.6. Sequence plasticity of the Ugi hydrophobic pocket

The Ung-sequestration hydrophobic pocket of PBS1 Ugi is formed by 8 residues (Figure 4.2.6); of which, only four are identical in Ugi-2_{89}. This provides important data on tolerated variations including the essential residues that perform Ung inhibition, which could be used in combination with acidic residues conserved PROSITE pattern to explore sequence space more specifically for Ugi variants.

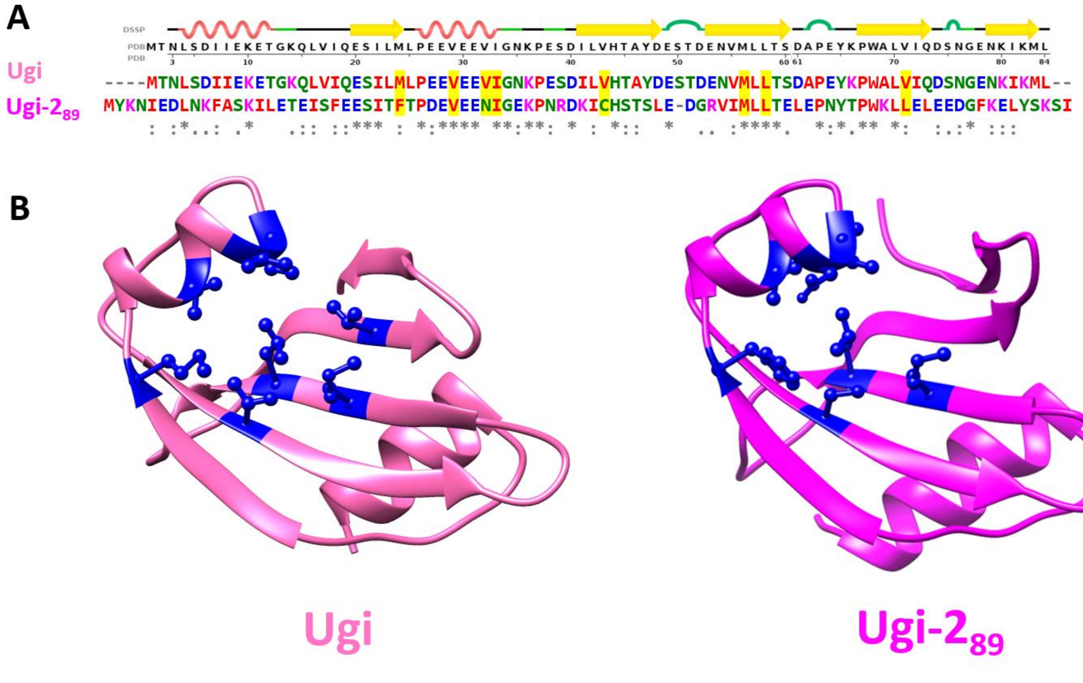

A
DSSP
PDB MTNLSDIIEKETGKQLVIQESILMLPEEVEEVIGNKPESDILVHTAYDESTDENVMLLTSDAPEYKPWALVIQDSNGENKIKML
PDB 3 10 20 30 40 50 60 61 70 80 84
Ugi
Ugi-2₈₉
----MTNLSDIIEKETGKQLVIQESILMLPEEVEEVIGNKPESDILVHTAYDESTDENVMLLTSDAPEYKPWALVIQDSNGENKIKML--
MYKNIEDLNKFASKILETEISFEESITFTPDEVEENIGEKPNRDKICHSTSLE-DGRVIMLLTELEPNYTPWKLLELEEDGFKELYSKSI
B
Ugi
Ugi-2₈₉

Figure 4.2.6. Ugi variant hydrophobic pockets. Both Ugi and Ugi-2_{89} exhibit a hydrophobic pocket serving to sequester a conserved catalytically critical residue (when it is a leucine or phenylalanine), employed by Ungs to intercalate DNA via the minor groove and stabilise the pre-catalytic complex. The residues comprising the hydrophobic pocket are highlighted yellow in the sequence alignment (panel A) and are shown in dark blue in the ribbon representation of both Ugi variants (panel B). The residue positions are structurally conserved and preserve function; however, the residue types are altered as evident in panel A.

4.2.7. Tolerated mutations in the core of the Ugi protein

The impressive structural conservation in Ugi variants (RMSD=1.343 Å, Figure 4.2.7) at low levels of sequence identity (32%) suggests the possibility of finding other Ugi variants that are not readily detectable by simple sequence similarity tools. This is especially because the sequence heterogeneity is not limited to the loops and other flexible secondary structures. It can be observed in some core-forming secondary structures such as the 2nd β-strand (41-ILVHTAYD-48 in PBS1 Ugi, and 45-KICHSTSL-52 in Ugi-2_{89}) in which only a histidine residue is identical among these 2 Ugi variants (Figure 4.2.7). Within the remaining residues that form this β-strand, there are 4 out of 7 residues that are non-conservative mutations (I41→K45, V43→C47, Y47→S51, and D48→L52).

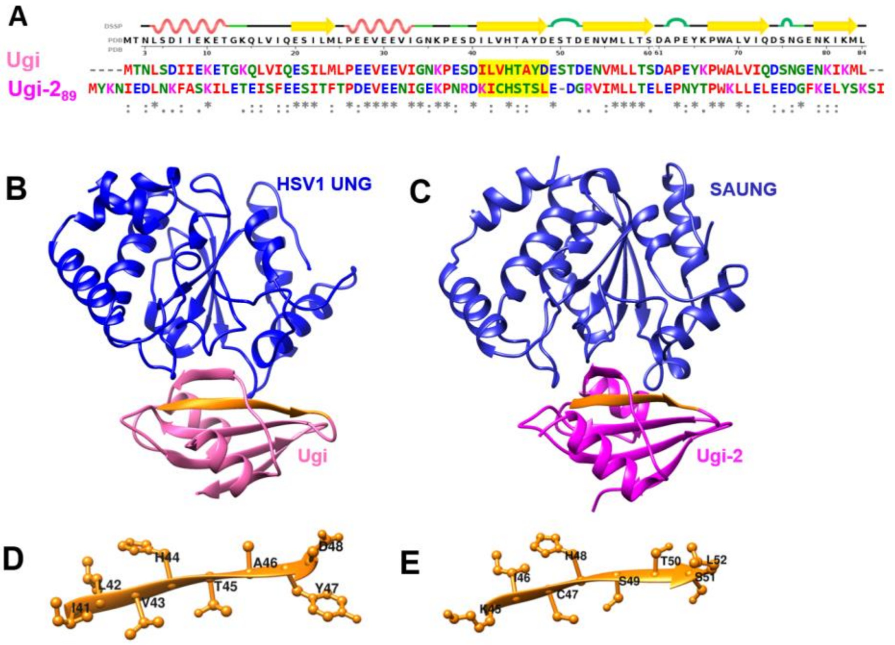

A
DSSP
PDB MTNLSDIIEKETGKQLVIQESILMLPEEVEEVIGNKPESDILVHTAYDESTDENVMLLTSDAPEYKPWALVIQDSNGENKIKML
PDB 3 10 20 30 40 50 60 61 70 80 84
Ugi ----MTNLSDIIEKETGKQLVIQESILMLPEEVEEVIGNKPESDILVHTAYDESTDENVMLLTSDAPEYKPWALVIQDSNGENKIKML--
Ugi-289 MYKNIEDLNKFASKILETEISFEESITFTPDEVEENIGEKPNRDKICHSTSLE-DGRVIMLLTELEPNYTPWKLLELEEDGFKELYSKSI
B
HSV1 UNG
Ugi
C
SAUNG
Ugi-2
D
I41
L42
V43
H44
T45
A46
Y47
D48
E
K45
I46
C47
H48
S49
T50
S51
L52

Figure 4.2.7. A comparison of Ugi variants. (A) structure-based sequence alignment of both Ugi variants with DSSP secondary structure of PBS1 Ugi on the top row. (B) HSV1-UNG:PBS1 Ugi complex crystal structure (PDB: 1UDI). (C) SAUNG:Ugi-2_{89} complex crystal structure (new data presented in this thesis). Even though Ugi variants share relatively low sequence identity (32%), their structures share the same fold, and are superimposable, with an RMSD value of 1.343 Å. (D) The 2nd β-strand of Ugi. (E) The 2nd β-strand of Ugi-2_{89}. Low sequence identity between Ugi variants can be observed in some core-forming secondary structures such as the 2nd β-strand (coloured orange in B, C, D, and E) in which only 1 out of 8 residues (see main text) is identical among these 2 Ugi variants and 4 out of 7 residues are non-conservative mutations.

4.3. Structural analysis of MCUGI1

4.3.1. Large-scale expression and purification of MCUGI1 in complex with SAUNG

MCUGI1 was expressed and co-purified with SAUNG in the same manner as described for Ugi-2 (section 4.2.1) with the difference that clarified lysates were mixed rather than resuspended cell pellets. The purification scheme and results are shown in Figure 4.3.1. SEC elution fractions with peak absorbance, and high purity as validated by SDS-PAGE were concentrated at 3200 x*g* using an Ultracel 3K centrifugal filter (Millipore) to reach a protein concentration of 25 mg/ml then stored at 4°C.

4.3.2. Initial crystallisation of MCUGI1 in complex with SAUNG

Purified protein sample was used at 25 mg/ml in commercial crystallisation screens (JCSG-Plus and PACT premier; Molecular Dimensions)[152]. Sitting drops were prepared at both 1:1 and 2:1 ratios of protein to mother liquor with volumes of 75 nl and 100 nl protein, respectively; plates were incubated at 16°C. Multiple conditions were selected for the observed presence of potentially crystalline material (Figure 4.3.2). Putative protein crystals grew on JCSG-Plus condition E2, whose conditions are: 2.0 M Ammonium sulphate, 0.2 M Sodium chloride, 0.1 M MES, pH 6.5. These crystals exhibited a sea urchin morphology, optimisation trials for the crystallisation conditions were performed to get crystals more suitable for data collection.

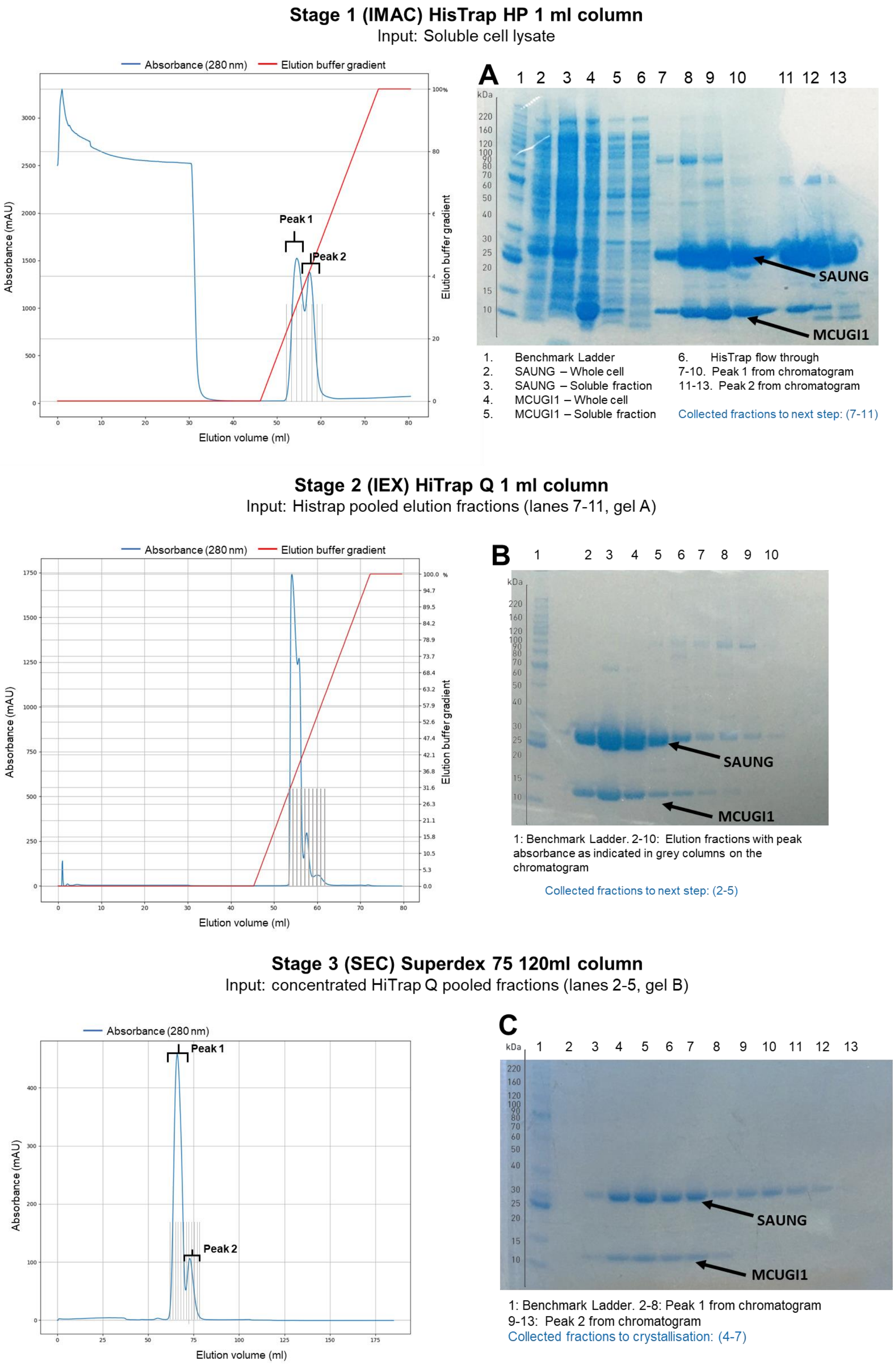


Figure 4.3.1. MCUGI1-SAUNG purification. Chromatograms and SDS-PAGE gels are shown for IMAC, IEX, and SEC stages.

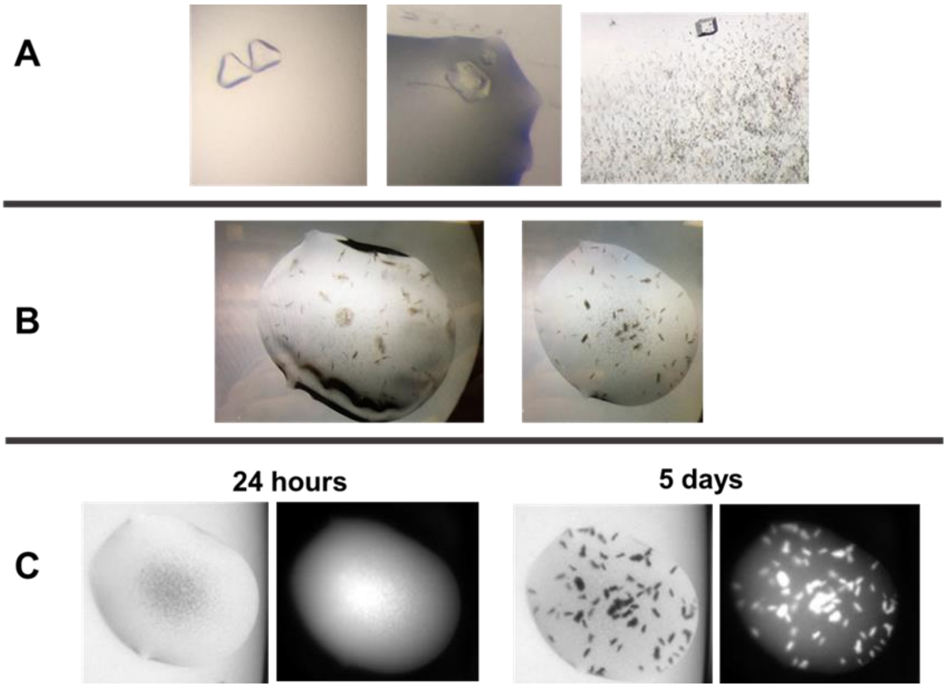


Figure 4.3.2. Putative protein crystals for MCUGI1 in complex with SAUNG. A) Crystals with no fluorescence under UV light, and with poor diffraction; these crystals grew on drops E1 (0.2 M sodium fluoride, PEG 3350 20%) and H1 (0.2 M sodium fluoride, PEG 3350 20%, 0.1 M Bis-Tris propane, pH 8.5) of PACT-premier crystallisation screen. **B)** Putative protein crystals formed in both sub-wells (both ratios of protein to mother liquor) of drop E2 of JCSG-Plus^{TM} screen; conditions are 2.0 M Ammonium sulphate, 0.2 M Sodium chloride, 0.1 M MES, pH 6.5. **C)** Comparison of drop E2 after 24 hours and after 5 days of setting the crystallisation screens showing the growth of fluorescent putative protein crystal under UV light.

4.3.3. Crystallisation optimisation of MCUGI1 in complex with SAUNG

The crystallisation condition of JCSG-Plus drop E2 was optimised. The same ratios of protein and mother liquor as the original crystallisation trial were used. Sitting drop crystallisation screens were set up at varying ammonium sulphate concentrations (1.1-2.2 M at 0.1 intervals), pH (5.5 to 7.5 at 0.2 intervals), or via changing precipitant composition by lowering the ammonium sulphate concentration and adding different types of PEG (1,500; 4,000; 6,000; or 20,000) at different concentrations (Figure 4.3.3). Solely varying salt concentration or changing the pH didn't form any better crystals. However, several drops produced crystals with different shapes by varying precipitant composition (Figure 4.3.3). An optimised condition of

1 M ammonium sulphate, 0.2 M sodium chloride, 9.55%(w/v) PEG 20.000, 0.1 M MES, at pH 6.5 led to formation of better-diffracting needle-shaped crystals. Those crystals were cryoprotected using the mother liquor solution supplemented with 25% (v/v) ethylene glycol.

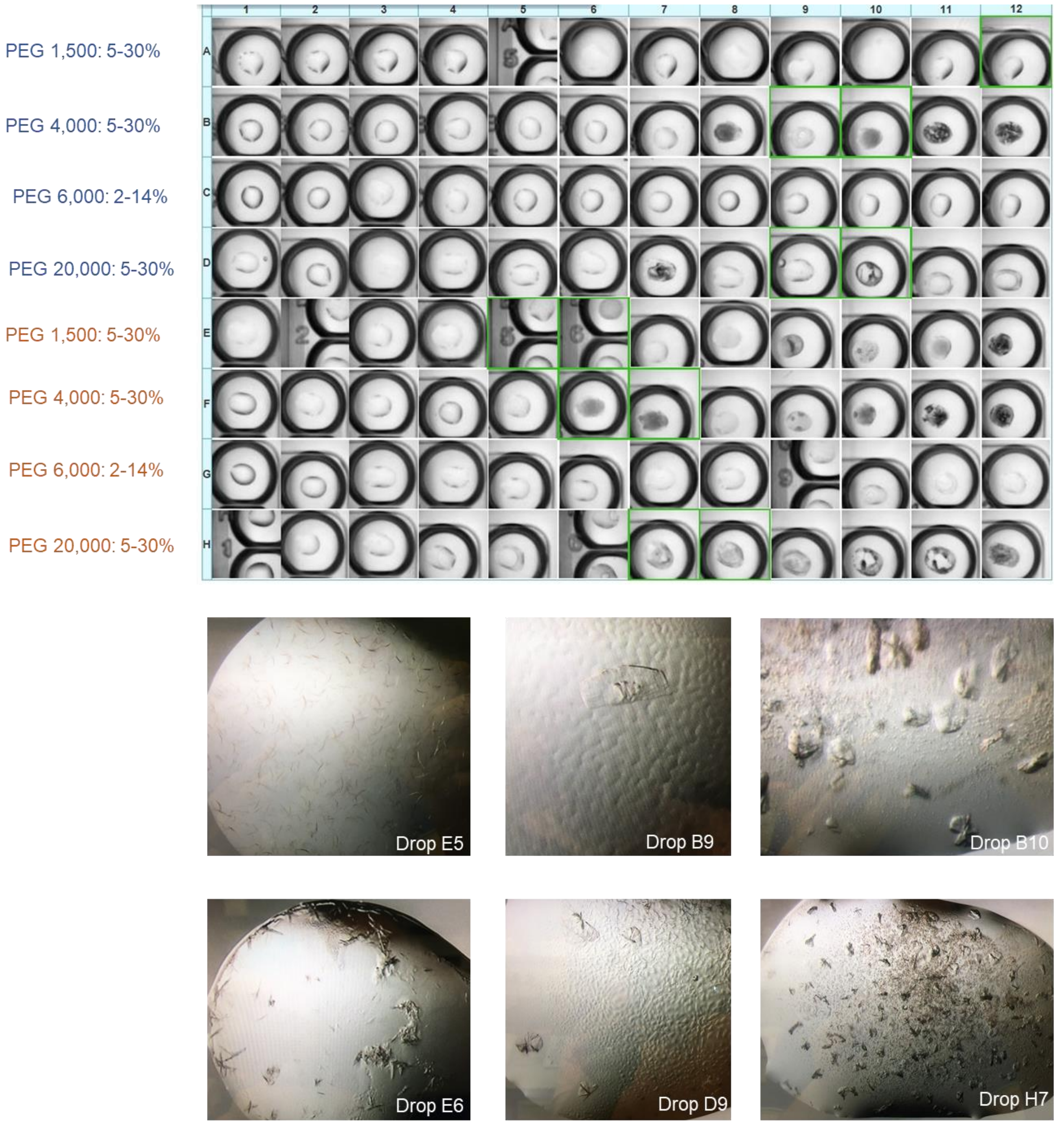


Figure 4.3.3. Crystallisation optimisation of MCUGI1-SAUNG complex crystals. Concentration of NaCl and MES, and the pH were set to be the same as in the original crystallisation conditions. Ammonium sulphate was added at 0.5 M into the entire rows A, B, C, and D and at 1 M into the entire rows E, F, G, and H. Composition of different PEG types is indicated next to each row. Concentrations were varied from the lowest to the highest value of the range (column 1 to 12, respectively) at equal intervals. Squares with green borders indicate drops in which crystals were formed. Crystals with a variety of shapes were formed with PEG 1,500; PEG 4,000; and PEG 20,000 both at 0.5 M and 1 M ammonium sulphate concentrations.

4.3.4. Structure determination of MCUGI1 in complex with SAUNG

X-ray diffraction data was collected on beamline ID30B at the European Synchrotron Radiation Facility (ESRF) with the help of Dr Nikos Pinotsis. The xia2 pipeline was used for automated data processing using XDS; AIMLESS was used to scale the structure factors.

The structure was initially phased by molecular replacement with MOLREP[115] using 3WDG [The complex of SAUGI-SAUNG][62] as the search model. The structure was solved at a resolution of 2.70 Å. Structure refinement used REFMAC5[118], and model building was performed using the program COOT[117]. Statistics for data collection and refinement are summarised in Table 4.3.4. This model has been deposited in the protein data bank (PDB ID: 8AIN) and is currently on hold for release (for validation metrics please see appendix C).

Table 4.3.4. Crystallographic data collection and refinement statistics for MCUGI1:SAUNG complex crystals

Data collection	
Wave length (Å)	0.919763
Space group	P 63 2 2
Unit cell a, b, c (Å)	91.52, 91.52, 158.62
α, β, γ (˚)	90.00, 90.00, 120.00
Resolution (Å)	45.80-2.70 (2.83-2.70)
Unique reflections	11391 (1470)
Redundancy	11.9 (12.6)
Completeness (%)	100 (100)
I/σ(I)	2.12
R_{merge}	0.251 (1.682)
Refinement	
R_{work} (%)	0.208
R_{free} (%)	0.256
Bond RMSD (Å)	0.0078
Angle RMSD (˚)	1.586
Mean B value/no of atom	49/5286
Ramachandran plot (%)	
Most favoured (%)	92.06
Allowed (%)	7.94
Outliers (%)	0.00

4.3.5. Tolerated mutations in the Ung-binding β-strand of SAUGI

As previously described for Ugi vs Ugi-2, there is also an impressive structural conservation between SAUGI and MCUGI1 variants (RMSD=1.005 Å, Figure 4.3.5) at low levels of sequence identity (29%). Sequence plasticity can even be observed in the Ung-binding β-strand. Out of eight residues of the Ung-binding β-strand (24-ECESIEEI-31 in SAUGI, and 24-LTEFVQLG-31 in MCUGI1; Figure 4.3.5), only one glutamic acid residue is identical between SAUGI and MCUGI1. Within the remaining residues that form this β-strand, 5 out of 7 residues are non-conservative mutations. This glutamic acid residue in the Ung-binding β-strand is found to be absolutely conserved in all synthetic and natural Ugi variants identified in this thesis (section 3.2), and all the 900+ SAUGI natural occurring sequences output by PSI-BLAST search of SAUGI (section 3.4.1).

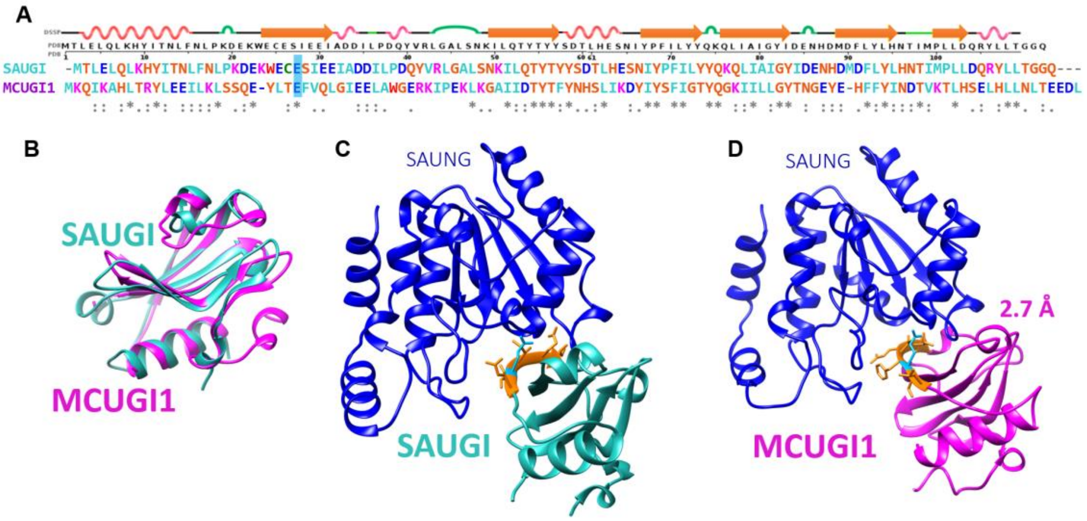

A
DSSP
PDB MTLELQLKHYITNLFNLPKDEKWECESIEEIADDILPDQYVRLGALSNKILQTYTYYSDTLHESNIYPFILYYQKQLIAIGYIDENHDMDFLYLHNTIMPLLDQRYLLTGGQ
PDB 1 10 20 30 40 50 60 61 70 80 90 100
SAUGI -MTLELQLKHYITNLFNLPKDEKWECESIEEIADDILPDQYVRLGALSNKILQTYTYYSDTLHESNIYPFILYYQKQLIAIGYIDENHDMDFLYLHNTIMPLLDQRYLLTGGQ---
MCUGI1 MKQIKAHLTRYLEEILKLSSQE-YLTEFVQLGIEELAWGERKIPEKLKGAIIDTYTFYNHSLIKDYIYSFIGTYQGKIILLGYTNGEYE-HFFYINDTVKTLHSELHLLNLTEEDL
B
SAUGI
MCUGI1
C
SAUNG
SAUGI
D
SAUNG
2.7 Å
MCUGI1

Figure 4.3.5. A comparison of SAUGI and MCUGI1 structures. (A) structure-based sequence alignment of SAUGI and MCUGI1 with DSSP secondary structure of SAUGI on the top row. (B) Superposition of SAUGI (PDB:3WDG, coloured cyan) and MCUGI1 (new data presented in this thesis, coloured magenta); RMSD between 73 pruned atom pairs (no pair is longer than 2 Å) is 1.005 Å, and across all 93 pairs is 3.425 Å. (C) SAUNG:SAUGI complex structure (PDB:3WDG). (D) SAUNG:MCUGI1 complex crystal structure (determined in this thesis). Even though MCUGI1 shares low sequence identity (29%) with SAUGI, their structures are superimposable. The 1st β-strand of SAUGI and MCUGI1 are coloured orange in panels C and D. In this β-strand, only 1 out of 8 residues (a glutamic acid residue, highlighted sky blue in panel A and coloured sky blue in panels C and D) is identical between both variants.

4.3.6. Sequence plasticity of the SAUGI hydrophobic pocket

The Ung-sequestering hydrophobic pocket of SAUGI is formed by 6 residues (Figure 4.3.6); of which, only three are identical in MCUGI1. The other three residues forming the hydrophobic pocket in MCUGI1 do not align either at the sequence or structure level with those forming the hydrophobic pocket in the wild-type SAUGI. These variations in the hydrophobic pockets provide important data that could be used to explore sequence space for SAUGI homologues.

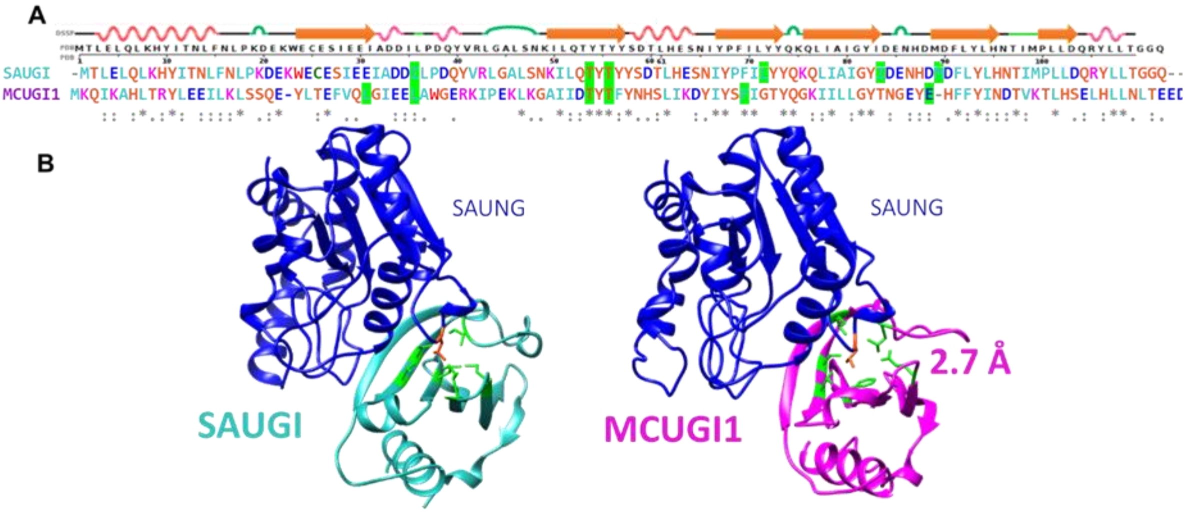

A
DSSP
PDB MTLELQLKHYITNLFNLPKDEKWECESIEEIADDILPDQYVRLGALSNKILQTYTYYSDTLHESNIYPFILYYQKQLIAIGYIDENHDMDFLYLHNTIMPLLDQRYLLTGGQ
PDB
SAUGI -MTLELQLKHYITNLFNLPKDEKWECESIEEIADDILPDQYVRLGALSNKILQTYTYYSDTLHESNIYPFILYYQKQLIAIGYIDENHDMDFLYLHNTIMPLLDQRYLLTGGQ--
MCUGI1 MKQIKAHLTRYLEEILKLSSQE-YLTEFVQLGIEELAWGERKIPEKLKGAIIDTYTFYNHSLIKDYIYSLIGTYQGKIILLGYTNGEYE-HFFYINDTVKTLHSELHLLNLTEED
B
SAUNG
SAUGI
SAUNG
MCUGI1
2.7 Å

Figure 4.3.6. SAUGI and MCUGI1 hydrophobic pockets. SAUGI and MCUGI1 hydrophobic pockets serve to sequester a conserved catalytically critical residue (a leucine in SAUNG, coloured red orange in panel B) employed by Ungs to intercalate DNA via the minor groove and stabilise the pre-catalytic complex. The residues comprising the hydrophobic pocket are highlighted green in the sequence alignment (panel A) and are shown in green in the ribbon representation of both SAUGI and MCUGI1 (panel B). The residue positions are neither sequence nor structure conserved, nevertheless preserving Ung inhibition function

4.3.7. Negative charge conservation in Ugi/SAUGI homologues

SAUGI has 17 acidic residues in its primary sequence, while MCUGI1 has 18 acidic residues. There are only 5 positions of conserved negatively charged residues between the two sequences; of which, 3 are located in 10aa span in the Ung-binding β-strand 1 and α-helix 2.

The PROSITE pattern of these acidic residues, E-X(6)-[ED]-[ED], is also conserved in Ugi and Ugi-2 sequences (section 4.2.5) and hence could be employed in sequence similarity search and heuristics-based approaches (section 5.4.1) to uncover potential novel naturally occurring Ugi/SAUGI homologues.

MCUGI1 surface negative charge distribution does not mimic DNA as precisely as observed in Ugi, Ugi-2 and SAUGI. However, a conserved hydrophobic pocket, an identical glutamic acid residue, and a conserved acidic motif are all common properties among all these four Ugi/SAUGI homologues (Figure 4.3.7).

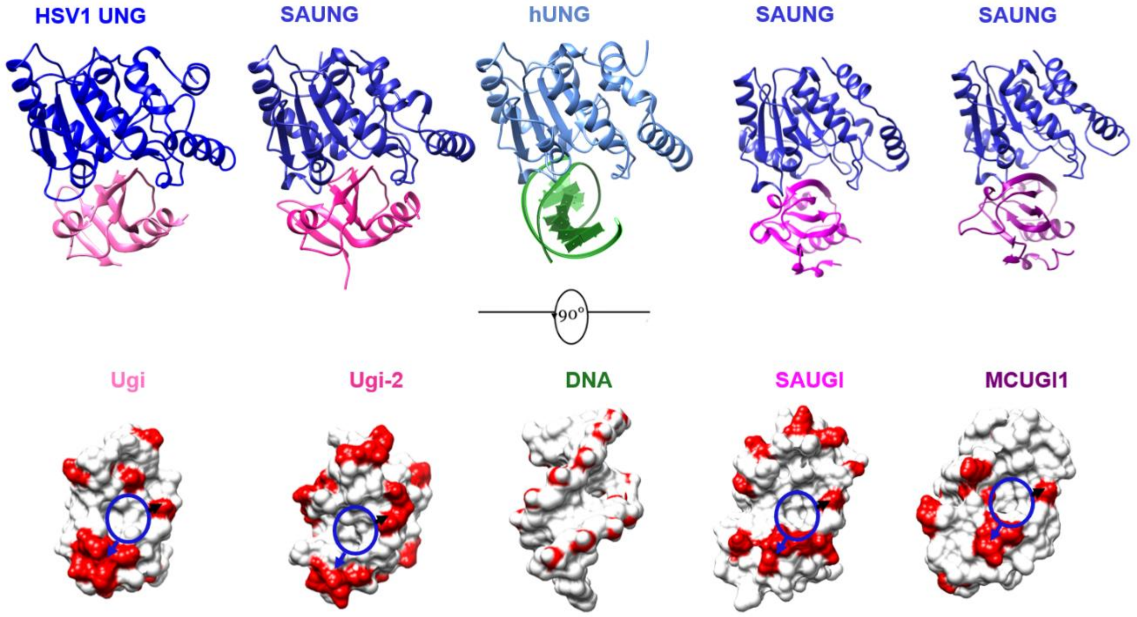


Figure 4.3.7. Mutual structural properties of Ugi/SAUGI homologues. Top row left to right: The structure of HSV1-UNG with PBS1 UGI (PDB: 1UDI), the structure of SAUNG with Ugi-2_{89} presented in this thesis, and the structure of hUNG with dsDNA (PDB: 1SSP), the structure of SAUNG with SAUGI (PDB: 3WDG), and the structure of SAUNG with MCUGI1 (presented in this thesis). Bottom row left to right: the Ung-binding interfaces of Ugi, Ugi-2, dsDNA, SAUGI, and MCUGI1. Bottom row depictions were generated by rotating each top row view by 90° degrees, removing Ung chains for clarity, then showing the surfaces of Ung-bound molecules. The red colour in the bottom row corresponds to the acidic residues of Ugi/SAUGI homologues and the phosphate groups of dsDNA. An identical glutamic acid residue in the 1st β-strand of Ugi/SAUGI homologues (black arrow), an acidic motif (blue arrow), and a hydrophobic pocket (blue circle) are structurally conserved properties among all these 4 Ugi/SAUGI homologues.

4.4. Structural analysis of VMY22 p56

The VMY22 p56 and *Bacillus weidmannii* Ung (BwUng) were cloned, expressed, co-purified, and co-crystallised by Ms Shelaine Fleck, a University of Alberta (Edmonton, Canada) Biochemistry 497 course unit summer intern, hosted by the Savva research lab. X-ray data was collected on beamline IO4 at the Diamond light source facility by Dr. Claire Bagneris (Rosalind Franklin Lab manager, Department of Biological Sciences, Birkbeck College). Data processing, structure refinement, and structure analysis were performed in this thesis.

4.4.1. Structure determination of VMY22 p56 in complex with BwUng

The xia2 pipeline was used for automated data processing using XDS; AIMLESS was used to scale the structure factors. The structure was phased by molecular replacement with MOLREP[115] using 4L5N [The complex of PZA-p56 and HSV1UNG][58] as search model. The structure was solved at a resolution of 2.45 Å. Structure refinement used REFMAC5[118], and model building was performed using the program COOT[117]. Statistics for data collection and refinement are summarised in Table 4.4.1. This model has been deposited in the protein data bank (PDB ID: 8AIL) and is currently on hold for release (for validation metrics please see appendix C).

Table 4.4.1 Crystallographic data collection and refinement statistics for VMY22 p56:BwUng complex crystals

Data collection	
Wave length (Å)	0.976250
Space group	P 1 21 1
Unit cell a, b, c (Å)	85.327, 97.495, 100.555
α, β, γ (˚)	90.000, 111.362, 90.000
Resolution (Å)	48.87-2.45 (2.52-2.45)
Unique reflections	56337 (4587)
Redundancy	4.4 (4.3)
Completeness (%)	99.8 (99.7)
I/σ(I)	2.53
R_{merge}	0.109 (0.548)
Refinement	
R_{work} (%)	0.204
R_{free} (%)	0.237
Bond RMSD (Å)	0.0074
Angle RMSD (˚)	1.413
Mean B value/no of atom	36/21921
Ramachandran plot (%)	
Most favoured (%)	97.22
Allowed (%)	2.78
Outliers (%)	0.00

4.4.2. Structural similarities and differences in p56 variants

The structure of VMY22 p56 is found to be strongly conserved when compared to the structure of PZA p56 (RMSD: 0.883 Å) at just 24% sequence identity. A dimer of p56 subunits inhibits Ung, and both monomers are involved in binding to Ung. Two residues, a glutamic acid residue and a tyrosine residue, located in the α-helix of p56 are crucial both for dimerization of p56 (E from the 1st monomer binds Y of the second monomer and vice versa) and form the walls of an Ung-sequestering hydrophobic pocket[58]. These 2 residues are conserved in VMY22 p56 (Figure 4.4.2). The 6-membered Ung-sequestering hydrophobic pocket of PZA p56 is formed by mirror residues within the helices of the dimer (Figure 4.4.2). Interestingly, the residues forming the 6-membered Ung-sequestering hydrophobic pocket of VMY22 p56 are not solely located within the helices of the dimer. Two tryptophan mirror residues participating in hydrophobic pocket forming are located in the 3rd β-strands of the dimer subunits (Figure 4.4.2). These differences provide additional insights that can be used to build new PROSITE patterns to explore sequence space for p56 homologous sequences.

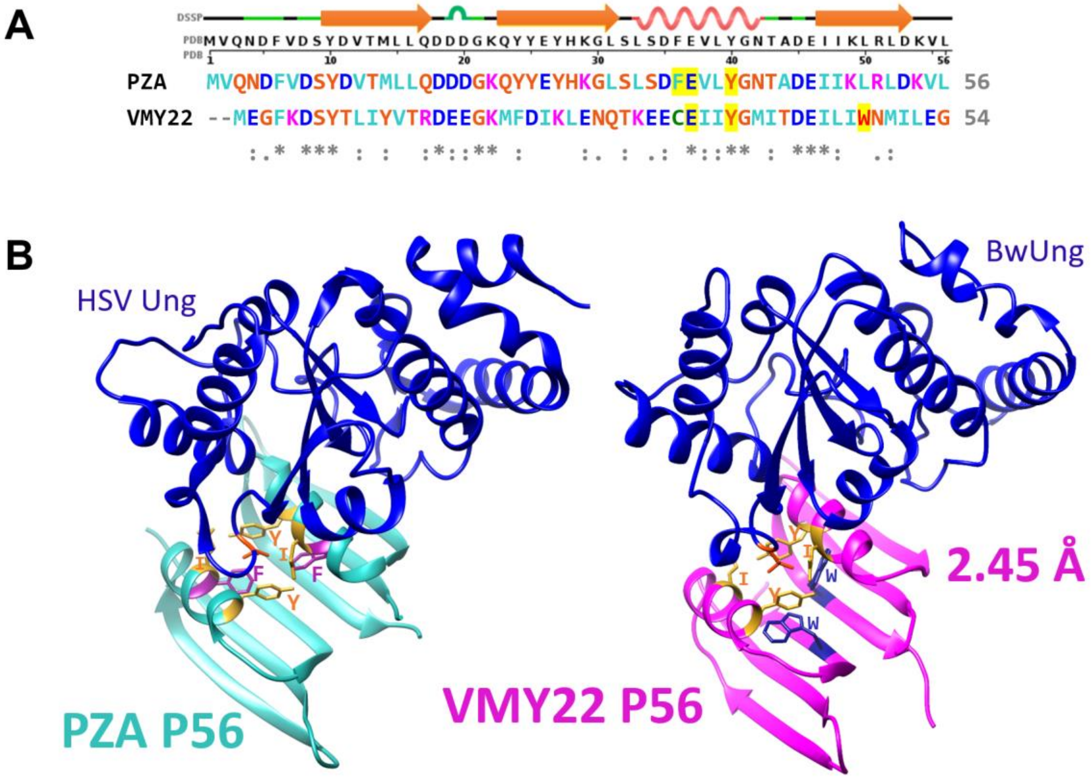

A
DSSP
PDB MVQNDFVDSYDVTMLLQDDDGKQYYEYHKGLSLSDFEVLYGNTADEIIKLRLDKVL
PDB 1 10 20 30 40 50 56
PZA MVQNDFVDSYDVTMLLQDDDGKQYYEYHKGLSLSDFEVLYGNTADEIIKLRLDKVL 56
VMY22 --MEGFKDSYTLIYVTRDEEGKMFDIKLENQTKEECEIIYGMITDEILIWNMILEG 54
:.* *** : : :*::** : :. : .: *::** : ***: .:
B
HSV Ung
BwUng
Y
I
F
F
Y
I
Y
Y
I
W
W
I
Y
2.45 Å
PZA P56
VMY22 P56

Figure 4.4.2. A comparison of p56 variant structures. (A) structure-based sequence alignment of PZA p56 and VMY22 p56 with DSSP secondary structure of PZA p56 (PDB: 4L5N) on the top row. (B) HSV-Ung:PZA p56 complex structure (PDB:4L5N) and BwUng:VMY22 p56 complex structure (new data presented in this thesis). The p56 hydrophobic pocket serves to sequester an Ung catalytic residue (a leucine in both HSV-Ung and BwUng, coloured red orange in panel B). The mirror residues comprising the hydrophobic pocket are highlighted yellow in the sequence alignment (panel A) and are shown as sticks and labelled in the ribbon representation of both p56 variants (panel B). A crucial phenylalanine residue at the helix of PZA p56 (coloured and labelled in dark magenta) is absent in VMY22 p56; the function of this residue is compensated by a tryptophan residue (coloured and labelled in navy blue) in the 3rd β-strand of VMY22 p56 subunits. The hydrophobic pocket is notably deeper in VMY22 p56.

4.5. Conclusion

Ung inhibitory proteins (UngIns) exhibit marked sequence plasticity. Such heterogeneity has hindered the unambiguous identification of known or expected UngIns in the genomes of closely related bacteriophages or those employing replication strategies antagonised by Ung. In this chapter, molecular structures were obtained for three UngIn variants in complex with Ungs: (1) A sequence variant Ugi protein encoded by *Bacillus* phage vB_BpuM-BpSp (Ugi-2, 32% sequence identity with phage PBS1 Ugi), (2) a novel SAUGI sequence variant encoded by *Macrococcus* rather than *Staphylococcus* species (MCUGI1, 29% sequence identity with SAUGI), and (3) a highly sequence variant p56 protein from *Bacillus* phage VMY22 (23% sequence identity with phage Phi29 p56). Structural conservation is shown to be high at such low levels of sequence conservation, and structural insights from Ung inhibition achieved by these variants provide useful knowledge for extending the search for other naturally encoded Ung inhibitors, which are to date conspicuously lacking from uracil containing genomes that would be presumed to require antagonism of Ung.

Chapter 5

5. Searching for UngIn in target organisms

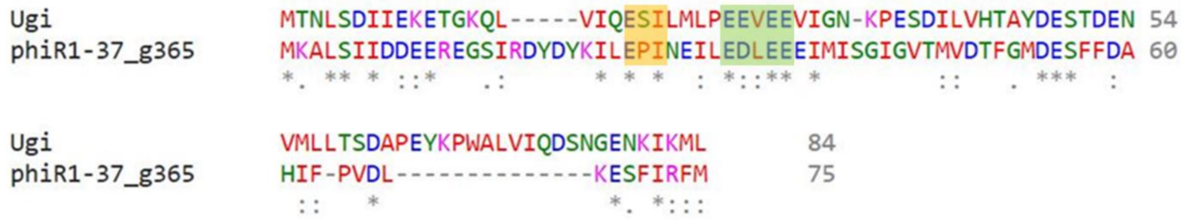
5.1. Introduction

A variety of organisms are known or expected to have UngIn sequences in their genomes. Some *Myoviridae* phages are known to have Ung-sensitive uracil-DNA genomes. These phages include *Yersinia* PhiR1-37[37], *Listeria* phage LPJP1[95], and *Staphylococcus* phages [93,94]. *Escherichia* phage T5, which is not a uracil-DNA phage, was reported to inhibit Ung activity[98]. Uracil-DNA genomes are also utilised by some roseophages[100]. However, no obvious UngIn sequence has been reported in any of these genomes. In this chapter, genome sequences that could encode known UngIn homologs were interrogated using sequence similarity search tools, and any UngIns with conserved biophysical attributes were searched for using heuristic-driven approaches in which mutual properties of specific UngIn types known sequences were deduced and applied to write algorisms to search for potential UngIn sequences that are undetectable by the available bioinformatics tools; finally, empirical screening of target genomes was also employed.

The full sequences of primers utilised in this chapter are listed in Table A.1, Appendix A.

5.2. Exploring the PhiR1-37 genome for UngIn homolog sequences

Performing BLASTP and PSI-BLAST searches for Ugi, SAUGI, or p56 homolog sequences in PhiR1-37 genome returned no sequences. Performing a PHI-BLAST search with Ugi as the template and the ESI motif PROSITE pattern: E-X-[ILVMF] returned no sequences. However, using the PROSITE pattern [ED]-[ED]-[ILV]-[ED]-[ED], which represents the acidic motif of Ugi α-helix 2 that contributes significantly to the Ung-Ugi electrostatic interaction[22], and changing the scoring matrix from BLOSUM62 into PAM30 (a more sensitive scoring matrix for identifying distant variants of protein homologs) returned 3 hits. One of these three hits (gp365) was interesting considering its molecular weight (8.75 kDa vs 9.5 kDa for Ugi), isoelectric point (3.884 vs 4.02 for Ugi), sequence length (75 aa vs 84 aa for Ugi), and sequence alignment with Ugi (good homology in acidic residue positions and in the ESI motif, Figure 5.2a).



```
Ugi             MTNLSDIIEKETGKQL-----VIQESILMLPEEVEEVIGN-KPESDILVHTAYDESTDEN 54
phiR1-37_g365   MKALSIIDDEEREGSIRDYDYKILEPINEILEDLEEEIMISGIGVTMVDTFGMDESFFDA 60
                *. ** * ::*    .:      * * *  : *::** *        ::   . ***  :

Ugi             VMLLTSDAPEYKPWALVIQDSNGENKIKML       84
phiR1-37_g365   HIF-PVDL--------------KESFIRFM       75
                 ::   *               *. *:::
```

Figure 5.2a. Sequence alignment of Ugi and g365 from phage PhiR1-37. The protein g365 has an EPI motif aligned with ESI (highlighted orange; EPI is a variant of the conserved Ugi/SAUGI motif that is observed in some naturally occurring SAUGI sequences, an example is the SAUGI variant with accession number: WP_053015655.1). Additionally, the protein g365 has a strong homology to the Ugi α-helix 2 motif (highlighted green).

The synthetic gene sequence for the potential Ugi homolog gene (locus_tag="phiR1-37_gp365") with accession number: YP_004934599 was designed *in silico* (Appendix A) as described (section 2.1.1.4). Synthetic DNA was obtained (IDT), amplification of g365 was performed via PCR using Q5 DNA polymerase and the primers P44 and P45. The g365 amplicon was isolated from 1.5% agarose and purified (section 2.1.1.8) and was cloned to pRSET-C using OE-PCR/ligation (section 2.1.1.11). The construct was designated pRS-g365.

The protein g365 was expressed at small-scale (section 2.1.2.1). Protein expression was induced by the addition 0.5 mM IPTG. Small-scale expression was tested at a variety of temperatures to decide upon the most suitable conditions (Figure 5.2b).

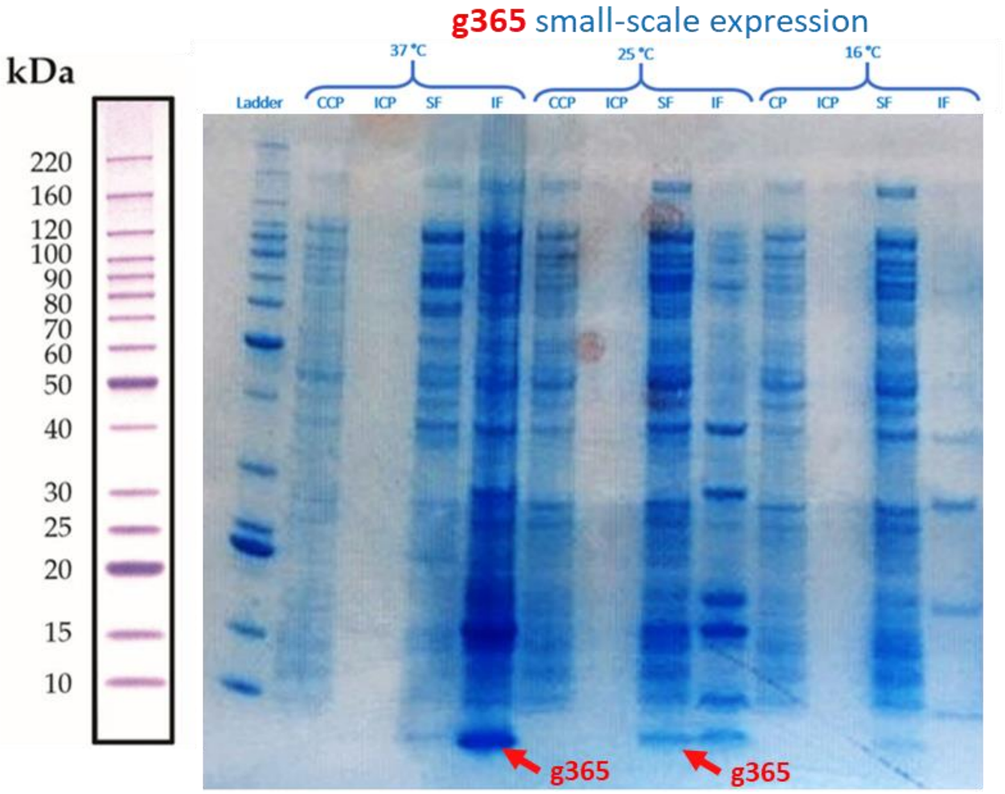


Figure 5.2b. SDS-PAGE showing recombinantly expressed g365 protein at a variety of incubation temperatures post induction at 0.5 mM IPTG. The optimal temperature for induced expression is observed to occur at 25 °C. CCP: control cell pellet, ICP: induced cell pellet, SF: Soluble fraction, IF: insoluble fraction.

Partial purification of the soluble fraction following initial cell lysis was performed according to the step-wise fractionation protocol (section 2.1.2.4). Eluates at 500 mM NaCl (Figure 5.2c) contained the peak band of target protein and hence were used in the UDG assay.

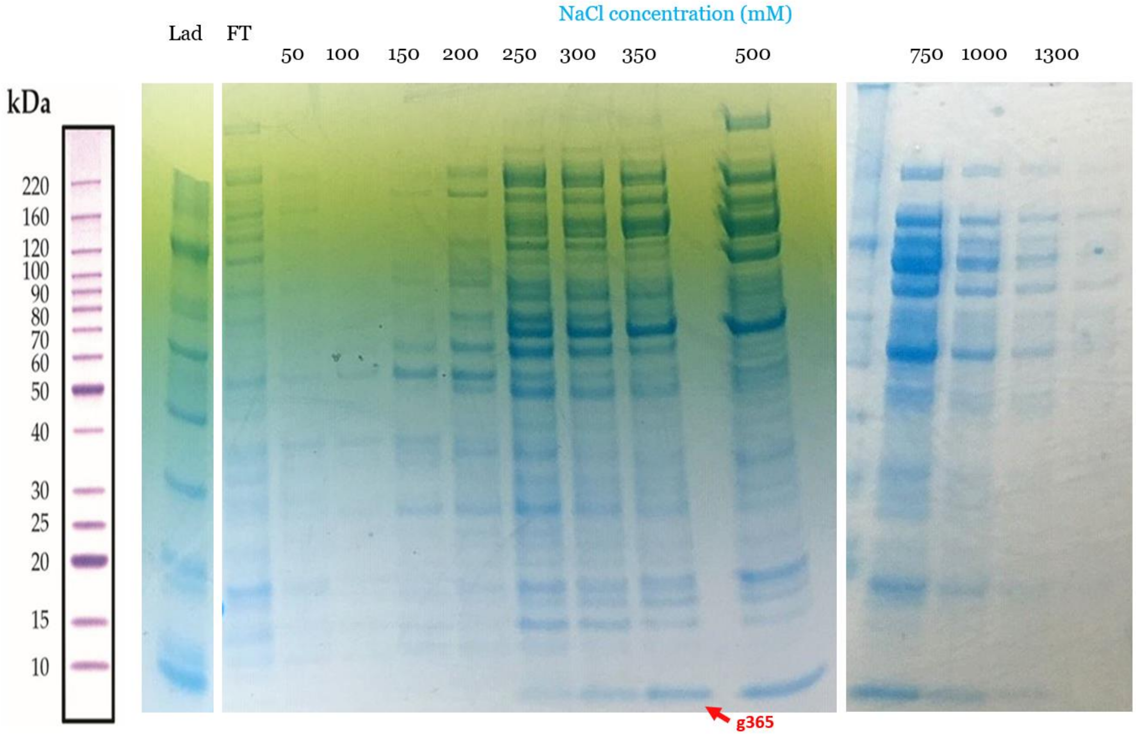


Figure 5.2c. SDS-PAGE analysis of Resin-bound g365 filtration on centrifugal filters (Ultrafree - MC) using a gradient of NaCl concentration buffer (Tris 20mM, pH 8). Most of g365 was eluted at 500 mM NaCl concentration. Eluates at 500 mM were selected for use in the UDG assay. Lad: BenchMark Ladder; FT: Flow through sample. This figure is formed of 3 cropped images. Original uncropped images are available and stored electronically.

In vitro UDG assay was used (section 2.1.2.7.1). *Staphylococcus aureus* UNG (SAUNG) activity on DNA substrates was tested when assayed alone or with either PBS1 Ugi or the g365 protein (Figure 5.2d). The protein g365 was not found to inhibit Ung.

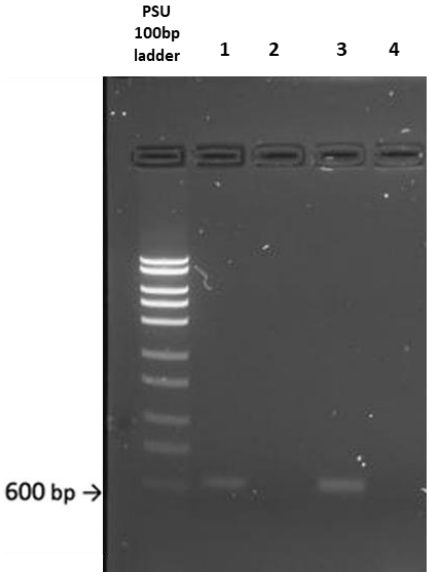


Figure 5.2d. UDG assay of the g365 protein. All reactions contained substrate DNA (~600 bp) in which all thymidine is supplanted by deoxyuridine. All reactions, excluding the leftmost control lane were incubated at 37 °C for 30 minutes. Lanes: (1) U-DNA substrate only; all other lanes included SAUNG at 39.7 nM, and the control Ugi (lane 3) or the tested g365 protein (lane 4). The protein g365 was not found to inhibit Ung. This gel represents three independent replicates.

5.3. Using insights from synthetic Ugi variants to search for novel UngIns

Insights from the tolerated mutations at the essential ESI motif of Ugi/SAUGI UngIn fold (section 3.2) were utilised to search for novel UngIn sequences. Using each synthetic Ugi variant uncovered in this thesis (section 3.2), along with a PROSITE pattern representing the Ung-binding β-strand motif of each synthetic Ugi, a PHI-BLAST search in the *Myoviridae* family (in this case from a synthetic variant that has an "EAM" motif in place of ESI) returned no Ugi sequences (as they have an ESI rather than an EAM motif); however, the PHI-BLAST search returned a hit encoded by the uracil-DNA *Staphylococcus* phage MarsHill that shares 31% identity with the synthetic "EAM" Ugi variants (Figure 5.3). This hit was undetectable using the known Ugi sequences in any PSI-BLAST or PHI-BLAST search. Suggested future work would involve assay of this hit for possible UngIn functionality.

```
EAM-Ugi         MTNLSDIIEKETGKQLV--IQEAMLMLPEEVEEVIGNKPES-----------DILVHTAY
MarsHill-hit    MM-------KKFNQEMINKIQEAMLPL---LKEELGNEIETVEFKVEDRLAINSLFVTAQ
                *        *: .::::  ****** *   ::* :**: *:           : *. **

EAM-Ugi         DE----------STDENVMLLTSDAPEYKPWALVIQDSNGENKIKML-------------
MarsHill-hit    QEMAAKLTKGGFNLDKTVFIFSDKAPEYKPVVSIYQSINKVTSLGSKRRFVKSILETIAH
                :*          . *:.*::::..****** . : *. *
```

Figure 5.3. MSA of synthetic Ugi variant (EAM-Ugi) and a phage MarsHill hit (accession: QQM14549.1) generated by PHI-BLAST search using EAM-Ugi and the PROSITE pattern I-Q-E-A-M (highlighted yellow) as inputs. The two proteins share 31% identity, the long insertions (highlighted green) in the MarsHill hit sequence take place between (rather than within) Ugi β-strands and α-helices.

5.4. Heuristics-driven approach for UngIn discovery

UngIn sequence/motif analyses and protein filter parameter design for both the Ugi/SAUGI fold and p56 fold were performed as part of this thesis. Naail Kashif-Khan, a fellow PhD student in the Savva research lab, wrote and ran a python script to perform the designed methodology according to my guidance and insights, and generated attributes and filter pipeline figures. Analyses of the candidate hits generated by the heuristics-driven approach was my own work.

5.4.1. Ugi/SAUGI Heuristics-driven search

5.4.1.1. A conserved motif in all Ugi/SAUGI sequences

A structure-based sequence alignment of Ugi and SAUGI (Figure 3.2) and a MSA of all the known and discovered Ugi/SAUGI sequences (in sections 3.3.1 and 3.4.1) reveal the presence of a conserved motif (designated the ESI motif, section 3.2) near the N-terminus of the protein (Figure 5.4.1.1). The ESI motif was defined as E-[ASVFHTNI]-[LVIFMT] (i.e., a strict ESI motif) or alternatively as E-[X]-[LVIFMT] (i.e., a lenient ESI motif). Taking into consideration ESI motifs of discovered synthetic Ugi variants (section 3.2), the third position of the ESI motif was re-defined as [LVIFMTPWC]. In both Ugi and SAUGI this motif is found in the β-strand that docks to the Ung DNA binding cleft. Ugi and SAUGI sequences not only share the conserved ESI motif, but also characteristic acidity and a distinct range of glycine/proline residues surrounding it. The presence of this motif and the sequence composition surrounding it was used to search for new UngIns, along with a suite of other sequence-based filters.

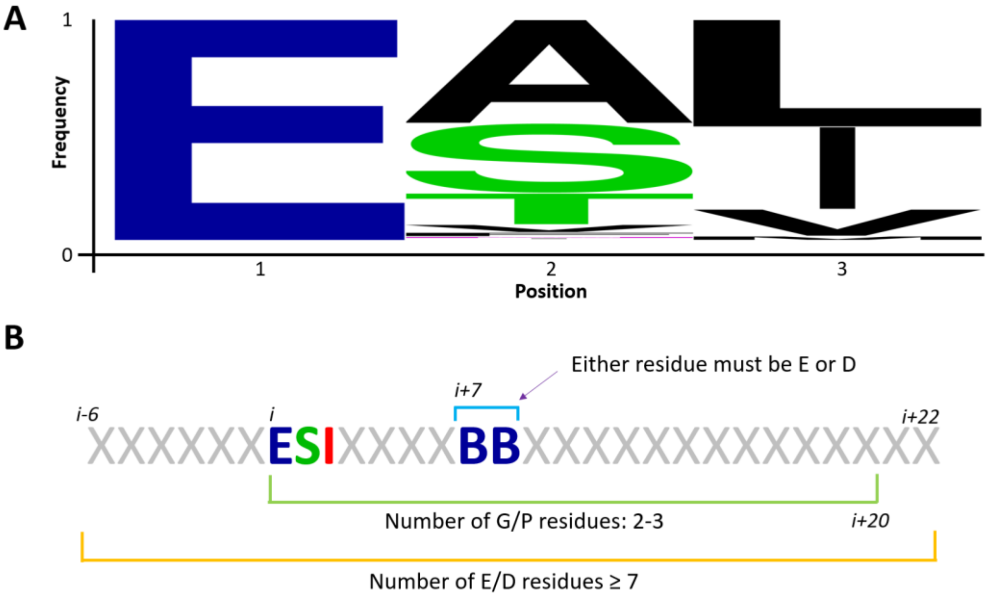

A
1
Frequency
0
1
2
3
Position
B
Either residue must be E or D
i+7
i-6
i
i+22
XXXXXXESIXXXXBBXXXXXXXXXXXXXX
Number of G/P residues: 2-3
i+20
Number of E/D residues ≥ 7

Figure 5.4.1.1. The ESI motif analysis. A) The ESI motif conservation shown as a sequence logo generated from an MSA of Ugi/SAUGI ESI motifs. The first position of the ESI motif is totally conserved as a glutamic acid (E), but the second and third positions show some sequence plasticity with the 3rd position being always hydrophobic. B) Conserved properties surrounding the ESI motif in Ugi/SAUGI sequences. If the first position of the ESI motif is designated residue i, then residue i+7 or i+8 is always acidic among all Ugi/SAUGI sequences. There are 2-3 glycine/proline residues and ≥7 acidic residues surrounding the ESI motif (adapted from a figure designed by Naail Kashif-Khan, a fellow PhD student in the Savva research lab).

5.4.1.2. Attributes of Ugi/SAUGI sequences

Analysis of hydrophobicity, acidity, molecular weight, glycine/proline content and glutamic acid/aspartic acid content within 991 Ugi/SAUGI sequences showed specific ranges for each analysed attribute (Figure 5.4.1.2). These attributes were utilised to build a set of filters for the discovery of novel UngIn sequences in target phage genomes.

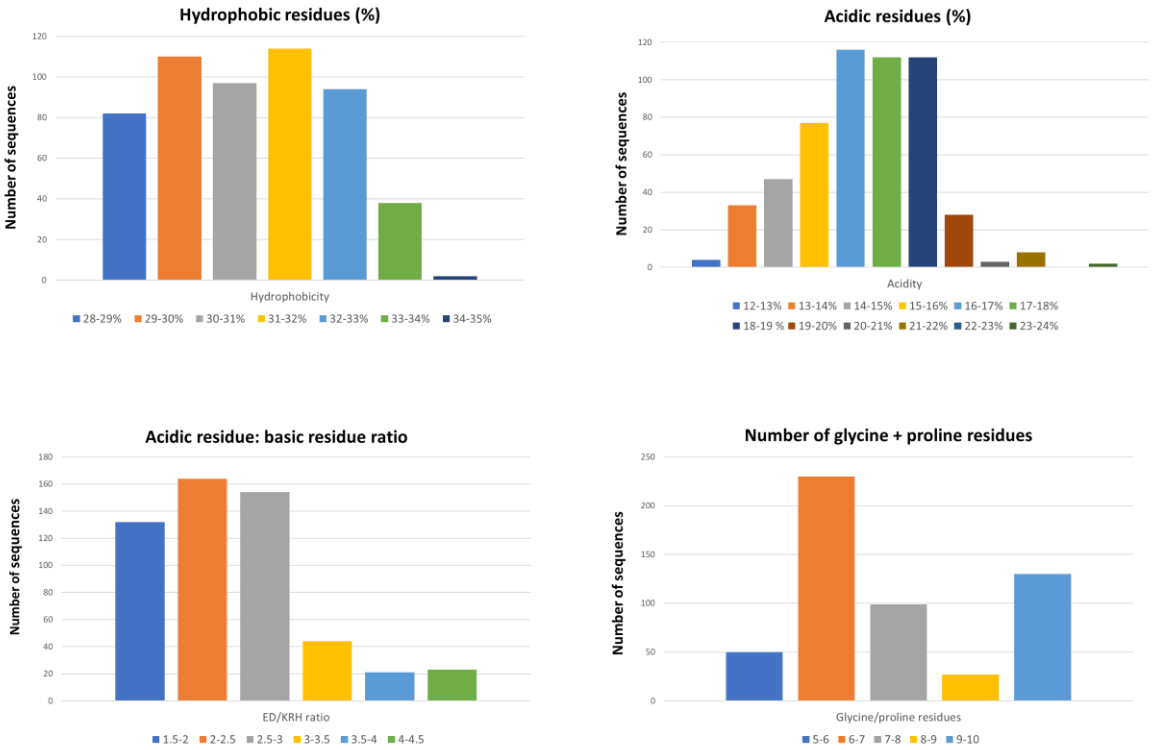


Figure 5.4.1.2. Analysis of specific Ugi/SAUGI properties. Histograms show the distribution of hydrophobicity, acidity, acidic/basic residue ratio, and glycine/proline content in Ugi/SAUGI sequences. Hydrophobic residues were defined as [IVLFMTWA], and acidic residues were defined as [ED].

5.4.1.3. Design and optimisation of the Ugi/SAUGI heuristics-based filter pipeline

Nucleic acid sequences underwent a six-frame translation, then a protein length filter, an acidity filter, a hydrophobicity filter, a glycine/proline residues content filter, and an acidic to basic residue ratio filter were applied (Table 5.4.1.3). Additional filters were used to remove sequences without an ESI motif satisfying the surrounding residue attributes described (Figure 5.4.1.1). Both strict and lenient ESI filters were imposed according to previous definitions (section 5.4.1.1). Filters and parameters used were optimised using known uracil-DNA phage genomes encoding annotated Ugi sequences (Figure 5.4.1.3) to validate the final set of optimised filters.

Table 5.4.1.3. Filter set and parameters used for single genome Ugi-like UngIn searches

Filter	Parameters
Minimum translation length	40 amino acids
Percentage of hydrophobic residues	36-48%
Percentage of acidic residues	12-24%
Number of glycine and proline residues	5-10
Ratio of acidic to basic residues	[E+D] : [K+R+H] ratio > 1.0
ESI motif filter	Lenient ESI filter: E-X-[IVLFMWPC]
ESI distance from C-terminus	50-120 residues from stop codon
ESI surrounding aspartates/glutamates	E+D count ≥ 7 in 28-residue window spanning from residue E-6 to residue E+22
ESI surrounding glycine/proline residues	G+P count= 2 or 3 in 20-residue window spanning from residue E-6 to residue E+20
ESI acidic residue position	7th or 8th residue from start of ESI motif must be E or D

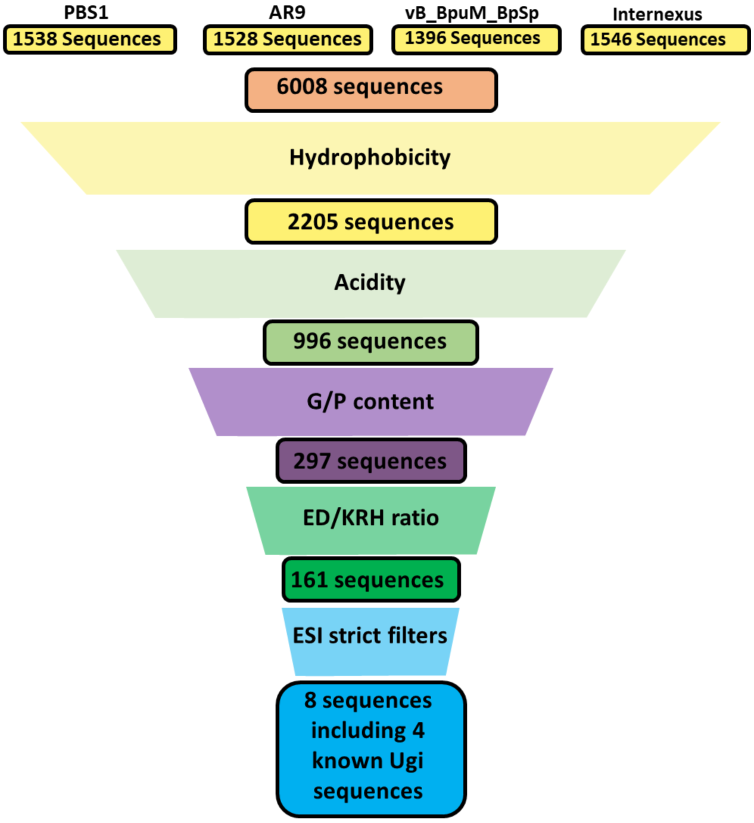


Figure 5.4.1.3. Validation filters using known Ugi-encoding genomes. Phage genomes PBS1, AR9, vB_BpuM-BpSp, and vB_BspM_Internexus were used as inputs to the pipeline, and parameters were refined until as few sequences as possible were retrieved at final output. Several filtration steps were applied. Out of 6008 input sequences, only 8 sequences including the 4 expected Ugi sequences passed the funnel of filtration.

5.4.1.4. Ugi-heuristic matches from *Yersinia* phage PhiR1-37

The filtration pipeline was run using the genome of *Yersinia* phage PhiR1-37 as an input. Filters returned two sequences matching the heuristic demands (Figure 5.4.1.4). These sequences pass all the filter checks and contain an ESI motif as expected. These 2 heuristic matches were assayed (section 5.5.1) but were found not to inhibit Ung.

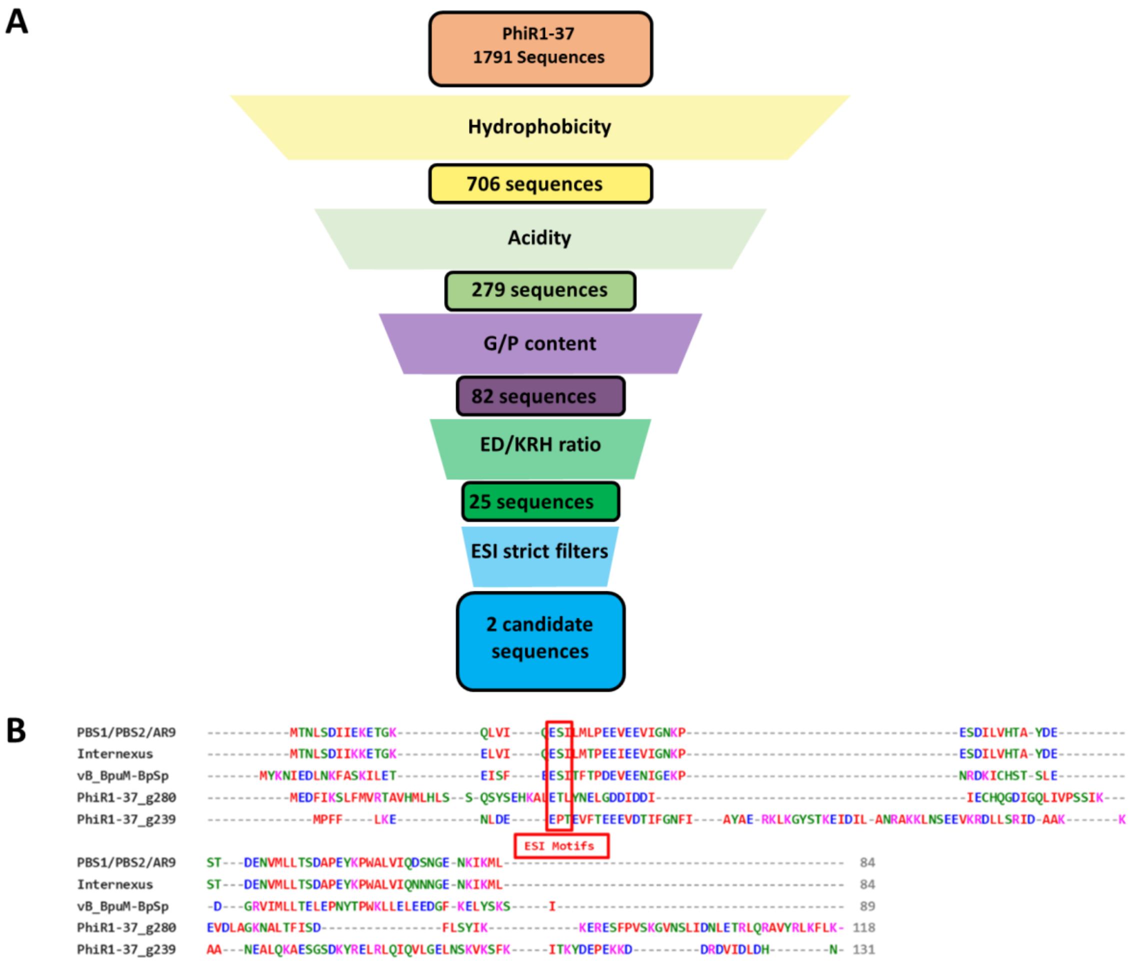

A
PhiR1-37
1791 Sequences
Hydrophobicity
706 sequences
Acidity
279 sequences
G/P content
82 sequences
ED/KRH ratio
25 sequences
ESI strict filters
2 candidate sequences
B
PBS1/PBS2/AR9 ----------MTNLSDIIEKETGK----------QLVI---QESILMLPEEVEEVIGNKP------------------------------ESDILVHTA-YDE--------
Internexus ----------MTNLSDIIKKETGK----------ELVI---QESILMTPEEIEEVIGNKP------------------------------ESDILVHTA-YDE--------
vB_BpuM-BpSp ------MYKNIEDLNKFASKILET----------EISF---EESITFTPDEVEENIGEKP------------------------------NRDKICHST-SLE--------
PhiR1-37_g280 ----------MEDFIKSLFMVRTAVHMLHLS--S-QSYSEHKALETLYNELGDDIDDI-----------------------------------IECHQGDIGQLIVPSSIK---
PhiR1-37_g239 -------------MPFF----LKE----------NLDE----EPTEVFTEEEVDTIFGNFI---AYAE-RKLKGYSTKEIDIL-ANRAKKLNSEEVKRDLLSRID-AAK------K
ESI Motifs
PBS1/PBS2/AR9 ST---DENVMLLTSDAPEYKPWALVIQDSNGE-NKIKML------------------------------------ 84
Internexus ST---DENVMLLTSDAPEYKPWALVIQNNNGE-NKIKML------------------------------------ 84
vB_BpuM-BpSp -D---GRVIMLLTELEPNYTPWKLLELEEDGF-KELYSKS-----I------------------------------ 89
PhiR1-37_g280 EVDLAGKNALTFISD---------------FLSYIK-----------KERESFPVSKGVNSLIDNLETRLQRAVYRLKFLK- 118
PhiR1-37_g239 AA---NEALQKAESGSDKYRELRLQIQVLGELNSKVKSFK-----ITKYDEPEKKD---------DRDVIDLDH--------N- 131

Figure 5.4.1.4. Candidate Ugi-heuristic matches from *Yersinia* phage PhiR1-37. A) The filter pipeline returned two candidate sequences out of 1791 possible protein-coding sequences. B) MSA including PhiR1-37 heuristic triaged sequences, aligned against known Ugi sequences from phages PBS1/PBS2/AR9, Internexus, and vB_BpuM-BpSp. The two candidate heuristic matches have ESI motifs (boxed in red in the MSA) that satisfy all heuristic parameters.

5.4.1.5. Ugi-heuristic matches from *Myoviridae*

All the *Myoviridae* family genomes were used as input for the filter pipeline with strictest parameters, including strict ESI filtering.

Genomes were also binned by GC content, our results demonstrate that all known Ugi variants were found in the second lowest GC bin; phage genomes encoding UngIns are therefore expected to have lower GC content relative to other phages: This most likely results from spontaneous deamination of cytosine-guanine base pairs to uracil-adenine not being corrected via Ung-induced BER in these phages. Indeed, the highest heuristic matches rate was found in GC bins from 25-35% (Figure 5.4.1.5a). Stricter sequence ESI motif filtering was applied to mass genome data, to limit the size of outputs and return a manageable number of heuristic matches for each GC bin. More lenient filters were applied when filtering single genomes, but when examining large files with thousands of genomes, stricter parameters were required. Heuristic matches generated from searching within *Myoviridae* genomes are listed in Table D.1 (Appendix D).

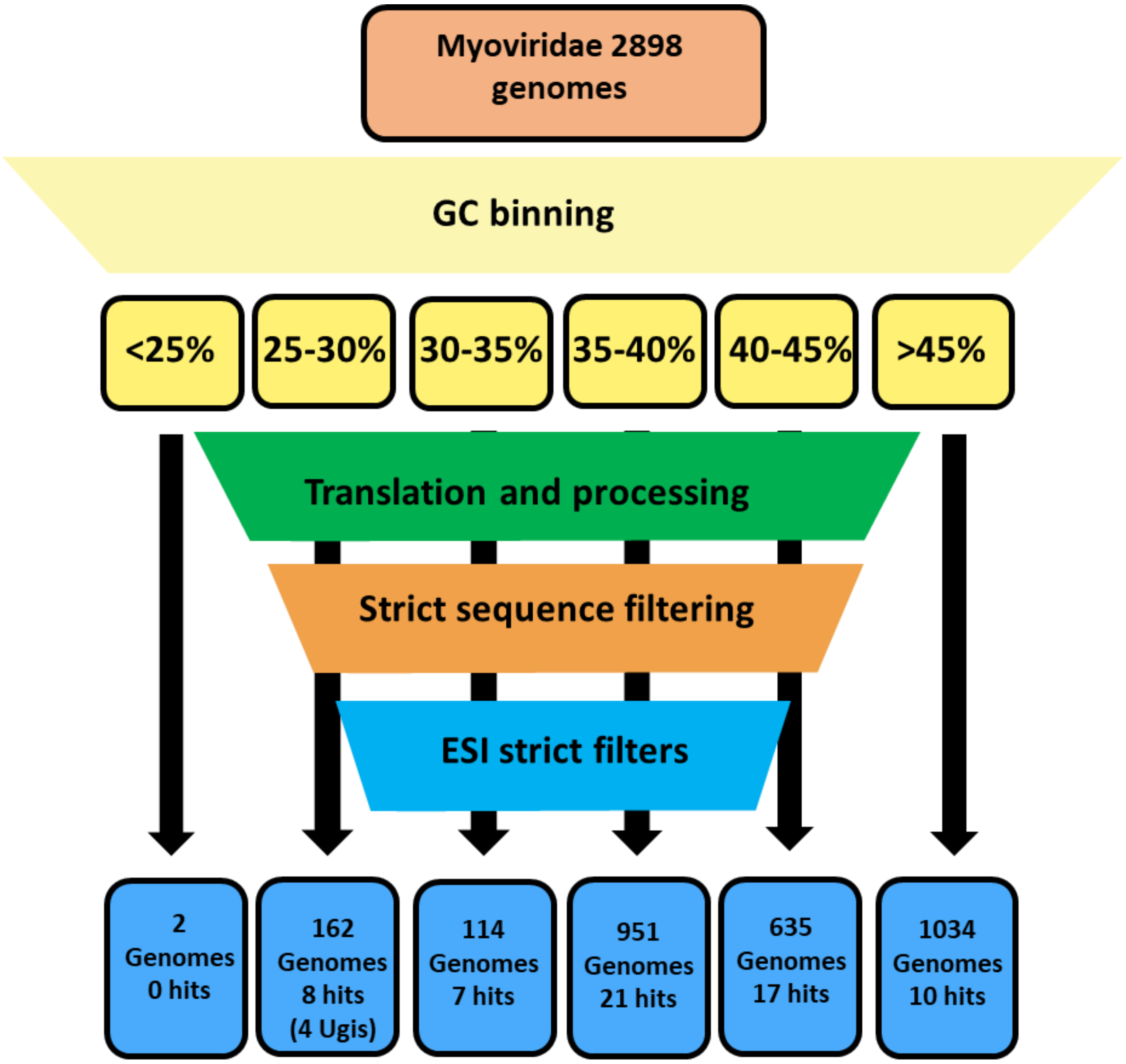
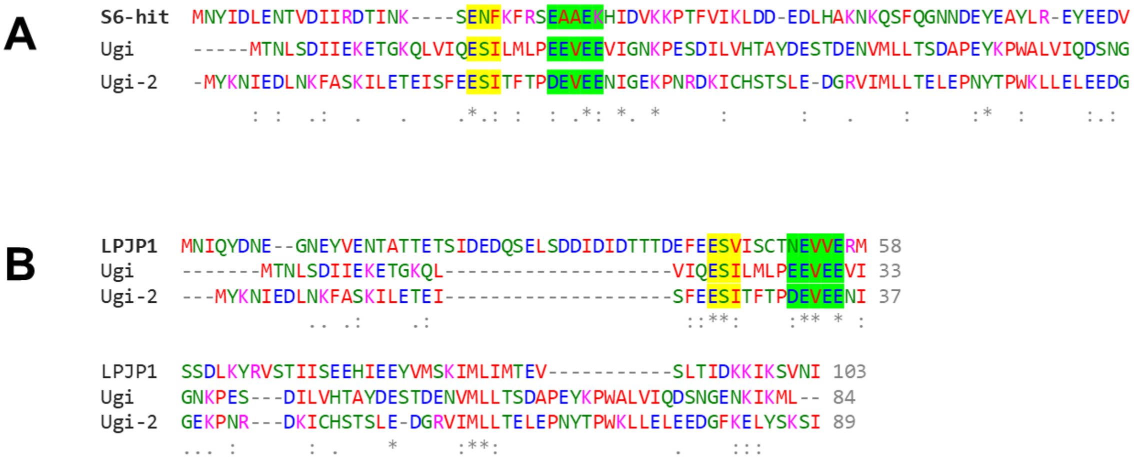


Figure 5.4.1.5a. Schematic of the filter pipeline for the *Myoviridae* Genomes. All known Ugi sequences were found in the second lowest GC content bin (25-30%). The highest heuristic matches rate (hits per 100 genomes) is found in the genomes with GC content of 25-35%.

None of the heuristic pipeline matches generated, other than Ugi sequences, is encoded by a known uracil-DNA genome (i.e., *Yersinia* phage PhiR1-37[37], *Listeria* phage LPJP1[95], or *Staphylococcus* phages S6[94], Machias, MarsHill, and Madawaska[93]). However, dropping one filter related to the glycine/proline residue count in a protein sequence led to identification of 2 heuristic pipeline matches in the uracil-DNA *Staphylococcus* phage S6 and *Listeria* phage LPJP1 (Figure 5.4.1.5b).



```
S6-hit    MNYIDLENTVDIIRDTINK----SENFKFRSEAAEKHIDVKKPTFVIKLDD-EDLHAKNKQSFQGNNDEYEAYLR-EYEEDV
Ugi       -----MTNLSDIIEKETGKQLVIQESILMLPEEVEEVIGNKPESDILVHTAYDESTDENVMLLTSDAPEYKPWALVIQDSNG
Ugi-2     -MYKNIEDLNKFASKILETEISFEESITFTPDEVEENIGEKPNRDKICHSTSLE-DGRVIMLLTELEPNYTPWKLLELEEDG
               : :  .:  .   .    .*.: :  : .*: *. *     :      :   .    :     :*  :     :.:
```

```
LPJP1     MNIQYDNE--GNEYVENTATTETSIDEDQSELSDDIDIDTTTDEFEESVISCTNEVVERM 58
Ugi       -------MTNLSDIIEKETGKQL--------------------VIQESILMLPEEVEEVI 33
Ugi-2     ---MYKNIEDLNKFASKILETEI--------------------SFEESITFTPDEVEENI 37
                     .. .:    .:                      ::**:    :** * :

LPJP1     SSDLKYRVSTIISEEHIEEYVMSKIMLIMTEV-----------SLTIDKKIKSVNI 103
Ugi       GNKPES---DILVHTAYDESTDENVMLLTSDAPEYKPWALVIQDSNGENKIKML-- 84
Ugi-2     GEKPNR---DKICHSTSLE-DGRVIMLLTELEPNYTPWKLLELEEDGFKELYSKSI 89
          ... :      : .    *     :**:               .    :::
```

Figure 5.4.1.5b. MSA of Ugi variants with heuristic matches from uracil-DNA phages. **A)** MSA of Ugi, Ugi-2 and a heuristic match from the *Staphylococcus* phage S6. **B)** MSA of Ugi, Ugi-2 and a heuristic match from the *Listeria* phage LPJP1. Both S6 and LPJP1 heuristic matches exhibit a plausible sequence conservation in the ESI motif (highlighted yellow) and the acidic motif located on the Ung-binding α-helix of Ugi (highlighted green).

The shared sequence identity between the identified heuristic matches and Ugi variants is very low; however, the similarity of the match alignment in the Ung-binding interface motifs is noticeable. It is suggested that further investigation of these heuristic pipeline matches could be considered as a part of future research.

5.4.2. p56 Heuristics-driven search

5.4.2.1. Conserved motifs in all p56 sequences

Multiple sequence alignment of p56 sequences (Figure 3.5.1) reveals the presence of two key motifs: the FXDSY (FX) motif and the EXXY (Ex) motif (Figure 5.4.2.1). These motifs are conserved in all known p56 sequences, and the EX motif involves in the dimerization and the Ung hydrophobic sequestration (as articulated in section 4.4.2). The FX motif was defined as F-X-D-S-Y while the EX-motif was defined as [EGS]-X-[LMVI]-[YVI]-G. As with the Ugi/SAUGI ESI-motif, sequence filters were applied to ascertain the presence of these two motifs, and the distance between them, to scan for p56-type UngIns.

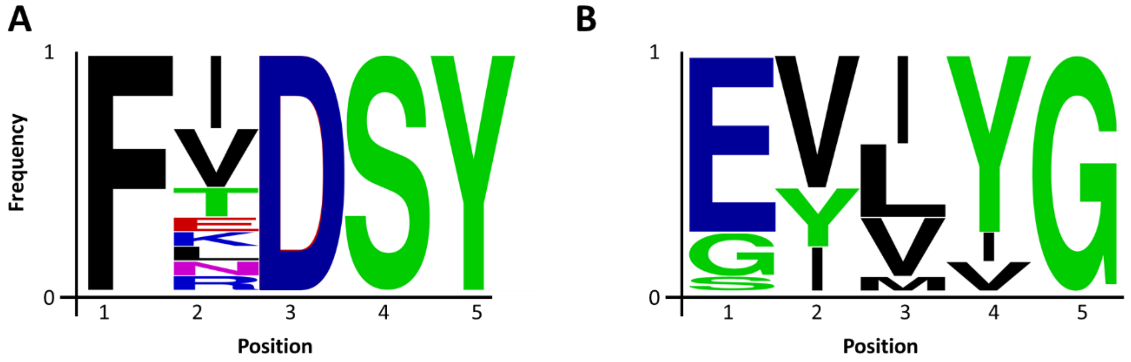


Figure 5.4.2.1. The p56 conserved motifs. A) The FXDSY motif Sequence logo generated from a p56 MSA. **B)** The EXXY motifs sequence logo generated from a p56 MSA.

5.4.2.2. p56 heuristics-based filter pipeline optimisation

The motif filters (for FX and EX) were verified by running them on phage genomes encoding known p56 sequences. Using only the two motif filters and the distance filter, all 19 p56 known sequences were returned, with zero additional sequences. Additional filters for protein size, acidity, and hydrophobicity were applied (Table 5.4.2.2).

Table 5.4.2.2. Filter set and parameters used for p56-like UngIn searches.

Filter	Parameters
Minimum translation length	40 amino acids
FX motif	F-X-D-S-Y
EX motif	[EGS]-X-[LMVI]-[YVI]-G
Distance between EX and FX motifs	23-28 Residues
Percentage of hydrophobic residues	36-48%
Percentage of acidic residues	12-24%

5.4.2.3. p56-heuristic matches from *Myoviridae*

The optimised filters were applied to the *Myoviridae* family genomes to scan for p56-like sequences. The search returned zero sequences. However, omitting the distance between Fx and Ex filters generated eight candidate sequence heuristic matches (Table D.2; Appendix D). None of these heuristic matches has obvious sequence similarity to any known p56 sequence.

5.5. Empirical Screening of PhiR1-37 for UngIns

5.5.1. Targeting ORFs satisfying heuristic UngIn properties

Yersiniophage PhiR1-37 is a jumbophage that has, similar to other Ugi-encoding phages (PBS1, PBS2, AR9), a uracil-DNA genome. This phage should encode an UngIn in order to survive, otherwise its genome would be degraded into smaller fragments upon host infection due to Ung and subsequent endonuclease activities in the host cell. No ORF in PhiR1-37 genome has a detectable sequence similarity to any known UngIn. The four types of UngIns have mutual properties including that they are small (less than 150 aa), acidic (with the exception of Vpr, which has an alkaline pI, this anomaly could be the result of the fact that it is a multifunctional protein with various protein-binding interfaces), and mimics DNA in the negative charge distribution on their surfaces. ORFs of PhiR1-37 were filtered based on these properties. There are 367 annotated proteins in PhiR1-37 genome, a shortened list of candidate-ORFs was created by applying specific filtration steps to the whole PhiR1-37 proteome. The filtration scheme was partly extracted from mutual properties of known UngIns; and it included 3 filters: (i) Acidity filter: all proteins that have an isoelectric point (pI) less than 6.00 (calculated on the ProtParam service on the EXPASY webserver[110]) passes this filter successfully, (ii) size filter: proteins that are composed of 35-150 aa passes this filter, (iii) and functionality filter: only proteins with an unidentified function passes through this filter. Out of 367 proteins, only 50 proteins passed through the 3-step filtration process (Table 5.5.1a). Oligos were designed (P48-P125; Table A.1) to amplify 39 genomic fragments (Table 5.5.1b) covering 56 genes including all these 50 candidate ORFs.

Yersiniophage PhiR1-37 (obtained commercially as a suspension, from vendor DSMZ), was used as a DNA template at 1 x 10^2 – 1 x 10^5 PFU (plaque forming units) in the genomic fragment amplification PCRs (section 2.1.1.1). As high-fidelity Q5 polymerase is not capable of amplifying fragments from uracil-DNA, amplicons were amplified firstly by a lower fidelity polymerase (Taq polymerase) for 8 cycles, then 1 µL of 25 µL volume Taq PCR reaction was used as a template to re-amplify with Q5 polymerase for additional 18 cycles.

Construct pSDM4_U12_Ung was linearised using primers P1 and P2. The amplified genomic fragments were cloned into pSDM4_U12_Ung via iPCR/overlap extension strategy (section 2.1.1.11). Sequence-verified constructs were transformed into CJ236 (Section 2.1.2.7.3). All 39 fragments were amplified successfully by Taq, then Q5, polymerases. Out of 39 designed fragments, OE-PCR showed bands with the expected size for 25 fragments; ligation, transformation and sequencing led to 22 constructs verified by sequencing. None of the 22 verified constructs covering 25 candidate ORFs was able to rescue the phenotype upon transformation into CJ236 cells (Table 5.5.1b).

Table 5.5.1a. PhiR1-37 triaged list of ORFs based on protein pI and length filtration

Gene	Gene name	Protein length (aa)	Protein parameters	
			Isoelectric point (pI)	Molecular weight (kD)
1	g032	93	5.8	10,621
2	g034	136	5.4	16,141
3	g038	100	5.7	11,679
4	g050	113	4.4	13,398
5	g051	117	4.8	14,016
6	g052	147	4.6	17,206
7	g065	54	4.1	6,035
8	g066	118	5.5	13,257
9	g071	58	5.0	6,498
10	g073	119	5.0	14,197
11	g088	86	5.0	9,958
12	g098	132	5.9	14,783
13	g110	79	4.7	9,253
14	g111	144	5.1	16,120
15	g112	95	3.9	10,727
16	g113	68	5.8	7,906
17	g117	51	4.0	5,629
18	g118	142	5.0	16,086
19	g119	109	5.4	12,991
20	g127	56	4.5	6,056
21	g132	94	5.9	10,560
22	g137	140	5.2	16,337
23	g142	95	6.0	9,440
24	g172	61	3.9	6,759
25	g182	122	5.3	14,489
26	g183	129	5.7	15,394
27	g186	123	5.9	14,583
28	g189	59	5.0	7,003
29	g214	106	4.2	12,151
30	g218	133	5.0	15,093
31	g219	119	5.8	13,696
32	g228	100	4.8	11,618
33	g236	140	4.6	15,304
34	g239	131	5.1	15,170
35	g268	69	5.1	7,710
36	g277	121	4.8	13,943
37	g280	118	5.5	13,508
38	g284	98	4.0	10,579
39	g291	90	5.0	10,036
40	g318	88	5.7	10,207
41	g319	115	4.9	13,184
42	g320	116	5.8	13,523
43	g323	106	5.8	12,244
44	g343	107	5.0	12,769
45	g344	90	4.4	10,902
46	g346	88	4.6	9,889
47	g351	139	4.9	16,037
48	g353	90	4.9	10,851
49	g361	91	5.5	10,419
50	g365	75	4.0	8,754

Table 5.5.1b. Work flow of cloning targeted PhiR1-37 genomic fragments into pSDM4_U12_Ung construct in place of U12. (✓) indicates that the step was performed successfully. (-) indicates that the step did not work successfully.

Fragment	gene	Taq PCR	Q5 PCR	OE-PCR	Correct Construct (verified by sequencing)	Survival colonies upon transformation into CJ236
1	g032	✓	✓	✓	-	
2	g034	✓	✓	-	-	
3	g038	✓	✓	-	-	
4	g050	✓	✓	-	-	
5	g051	✓	✓	✓	✓	No
6	g052	✓	✓	-	-	
7	g065/g066	✓	✓	-	-	
8	g071	✓	✓	-	-	
9	g073	✓	✓	-	-	
10	g088	✓	✓	-	-	
11	g098	✓	✓	✓	✓	No
12	g110/g111	✓	✓	-	-	
13	g112/g113	✓	✓	✓	✓	No
14	g117/g118	✓	✓	✓	✓	No
15	g119	✓	✓	✓	✓	No
16	g127	✓	✓	✓	✓	No
17	g132	✓	✓	✓	✓	No
18	g137 to g142	✓	✓	-	-	
19	g172	✓	✓	✓	-	
20	g182/g183	✓	✓	✓	✓	No
21	g186	✓	✓	✓	✓	No
22	g189	✓	✓	-	-	
23	g214	✓	✓	✓	✓	No
24	g218/g219	✓	✓	-	-	
25	g228	✓	✓	✓	✓	No
26	g236	✓	✓	✓	✓	No
27	g239	✓	✓	✓	✓	No
28	g268	✓	✓	✓	✓	No
29	g277	✓	✓	-	-	
30	g280	✓	✓	✓	✓	No
31	g284	✓	✓	✓	✓	No
32	g291	✓	✓	✓	✓	No
33	g318 to g323	✓	✓	✓	-	
34	g343/g344	✓	✓	-	-	
35	g346	✓	✓	✓	✓	No
36	g351	✓	✓	✓	✓	No
37	g353	✓	✓	✓	✓	No
38	g361	✓	✓	✓	✓	No
39	g365	✓	✓	✓	✓	No

5.5.2. Shotgun library screening of PhiR1-37 genome

The strategy of modifying DNA and using a uracil-DNA may render bacteriophages resistant to some restriction-modification systems utilised by their hosts[79,153]; however, there are several restriction endonucleases that are able to digest uracil-DNA. EcoRI, BamHI, and BclI are amongst these enzymes that can digest uracil-DNA efficiently[79,80,153,154]. The count of restriction sites that can be cut by each of these enzymes in PhiR1-37 genome and the average length of fragments generated by each enzyme were calculated (Table 5.5.2). Based on this analysis, EcoRI enzyme showed a suitable number of restriction sites within PhiR1-37 genome (144 sites) and an average fragment length (1.82 Kb) that is suitable for downstream cloning, hence was selected to perform digestion of PhiR1-37 DNA. An EcoRI site was introduced to pSDM4_U12_Ung construct between U12 and Ung genes by iPCR/Ligation strategy (section 2.1.1.10) using the primers P126 and P127 (Table A.1). The resulting construct (pU12_EcoRI_Ung) doesn't contain any other EcoRI restriction site. PhiR1-37 DNA (400 ng) was digested with EcoRI (30 U) in a 50 µL reaction mixture for 4 hours at 37°C. The reaction was stopped by heating at 65°C for 15 min. The construct pU12_EcoRI_Ung (100 ng) was linearised with EcoRI in the presence of CIP and was gel purified. EcoRI digested PhiR1-37 DNA (400 ng) and linearised pU12_EcoRI_Ung (80 ng) were ligated (section 2.1.1.9) at 16°C for 14 hours. Transformation of the ligation product into NEB 5-alpha cells yielded ~100 colonies. Transformation of plasmid DNA isolated from the lawn of NEB 5-alpha cells into CJ236 cells returned no surviving colonies, however.

Table 5.5.2. Restriction sites analysis of PhiR1-37 genome

Restriction enzyme	Restriction site count in PhiR1-37	Average fragment length
EcoRI	144	1.82 Kb
BamHI	13	20.17 Kb
BclI	85	3.09 Kb

5.5.3. Targeting hypothetical proteins exclusive to uracil-DNA phages

In addition to *Yersinia* phage PhiR1-37[37], *Staphylococcus* phages S6, MarsHill, Madawaska, Machias[93,94], and *Listeria* phage LPJP1[95] are *Myoviridae* jumbophages known to have uracil-DNA with no identified UngIn in their genomes. We hypothesised that if these phages encode a novel type UngIn, it could be one of the hypothetical proteins that are unique to these genomes. A TBLASTN search with the PhiR1-37 genome against the genomes of *Staphylococcus* phages MarsHill, Madawaska, Machias, and *Listeria* phage LPJP1 showed that 7 of the PhiR1-37 hypothetical proteins are encoded by all of these phages. A PSI-BLAST search was performed for each of these 7 hits to see if they are exclusive to uracil-DNA phages. It was found that 6 of these hits were indeed limited to uracil-DNA phages, including the ones known to encode Ugi, and 2 closely related *staphylococcus* phages (PALS2, and vB_StaM_SA1) that show genomic similarity to known uracil-DNA phages[155] and that, we would speculate, could be uracil-DNA genomes (Table 5.5.3). Two hits among these 6 hits are homologous to each other (genes g278 and g282 of PhiR1-37) and also exist as pairs of homologous sequences in all other uracil-DNA genomes. Primers P128-P139 were used to amplify the six target genes from PhiR1-37 DNA (Table A.1). The six identified hits were cloned into pSDM4_U12_Ung via OE-PCR (section 2.1.1.11). It was found that none of the hits indicated Ung inhibitory activity when their corresponding constructs were tested via transformation into CJ236.

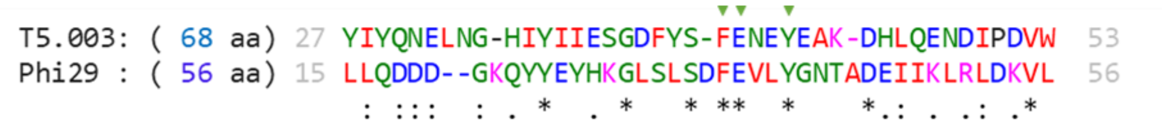
Table 5.5.3. Hypothetical proteins exclusively encoded by Uracil-DNA phages

PhiR1-37 encoded gene (length in aa, pI)	Homologous gene Accessions	Phage encoding the homologous gene	ID %	Accession length (aa)
g207 (863, 4.3)	YP_004934441.1	phiR1-37	100%	863
	YP_009283116.1	AR9	30%	904
	QXN70134.1	vB_BspM_Internexus	30%	904
	ALN97947.1	vB_BpuM-BpSp	30%	903
	QQO92841.1	Madawaska	28%	995
	QPI17170.1	vB_StaM_SA1	27%	995
	QQM14726.1	MarsHill	28%	995
	QDJ97626.1	PALS_2	28%	995
	QQO92559.1	Machias	27%	996
	YP_009664305.1	PBS1	30%	904
	QXN67971.1	LPJP1	27%	995
g278 (295, 5.0) and its homologous g282 (225, 4.8)	YP_004934512.1	phiR1-37	100%	295
	YP_009283146.1	AR9	44%	234
	QXN70164.1	vB_BspM_Internexus	43%	234
	ALN97977.1	vB_BpuM-BpSp	39%	240
	YP_009283165.1	AR9	37%	221
	QXN70182.1	vB_BspM_Internexus	36%	230
	ALN97995.1	vB_BpuM-BpSp	41%	214
	QQO92450.1	Machias	41%	211
	QDJ97768.1	PALS_2	39%	214
	QQM14590.1	MarsHill	39%	214
	QQO92712.1	Madawaska	39%	214
	QPI17045.1	vB_StaM_SA1	39%	210
	YP_004934516.1	phiR1-37	35%	225
	QPI17143.1	vB_StaM_SA1	39%	227
	QQO92553.1	Machias	37%	230
	QDJ97877.1	MarsHill	37%	227
	QDJ97877.1	PALS_2	37%	227
	QQO92814.1	Madawaska	37%	227
	YP_009664335.1	PBS1	44%	234
	QXN67771.1	LPJP1	40%	249
	QXN67837.1	LPJP1	45%	174
	YP_009664352.1	PBS1	37%	221
g234 (846, 4.9)	YP_004934468.1	phiR1-37	100%	864
	YP_009664354.1	PBS1	28%	738
	YP_009283164.1	AR9	28%	738
	QXN70181.1	vB_BspM_Internexus	28%	738
	ALN97996.1	vB_BpuM-BpSp	27%	734
	QQO92451.1	Machias	28%	848
	QPI17046.1	vB_StaM_SA1	28%	847
	QQM14591.1	MarsHill	27%	847
	QQO92713.1	Madawaska	27%	847
	QXN67836.1	LPJP1	28%	823
	QDJ97769.1	PALS_2	27%	847
g244 (290, 9.6)	YP_004934478.1	phiR1-37	100%	290
	YP_009283131.1	AR9	30%	274
	QXN70149.1	vB_BspM_Internexus	29%	269
	QQM14714.1	MarsHill	29%	257
	QQO92829.1	Madawaska	29%	257
	QQO92570.1	Machias	28%	255
	QPI17158.1	vB_StaM_SA1	35%	254
	QDJ97892.1	PALS_2	31%	227
	ALN97961.1	vB_BpuM-BpSp	29%	254
	QXN67980.1	LPJP1	33%	173
	YP_009664320.1	PBS1	30%	274
g196 (850, 7.0)	YP_004934430.1	phiR1-37	100%	850
	ALN97953.1	vB_BpuM-BpSp	26%	408
	QQO92567.1	Machias	23%	430
	QQO92832.1	Madawaska	20%	424
	QQM14717.1	MarsHill	20%	424
	QDJ97895.1	PALS_2	21%	420
	QPI17161.1	vB_StaM_SA1	20%	427
	QXN70140.1	vB_BspM_Internexus	24%	396
	YP_009283122.1	AR9	24%	396
	YP_009664311.1	PBS1	24%	396

5.6. Exploring *Escherichia* phage T5 genome for UngIn homolog sequences

Escherichia phage T5 is not a uracil-DNA genome. However, previous reports showed that single-stranded DNA regions are generated during its replication[156]. It was reported that phage T5 inhibits Ung activity early after injecting its genomic material into *E. coli*. The report utilised a T5-infected *E. coli* extract to test *in vitro* Ung activity; a mutant *Escherichia* phage T5 that is able to inject only 7.9% of its genomic material was able to inhibit Ung activity, suggesting that the ORF in charge for Ung activity modulation is encoded by the first 7.9% of T5 genome[98]. However, another report showed that a mutant phage T5 that lacks the dUTPase gene, hence has a uracilated genome, is unable to replicate in an Ung+ *E. coli*[157], indicating that Ung inhibition by a gene encoded by T5 is unlikely, while an effective phage-encoded dUTPase is potentially preventing incorporation of dUTPs into replication intermediates and hence protecting these intermediates from Ung-initiated restriction.

A manual alignment of each known UngIn sequence with each of the 17 ORFs of that 7.9% T5 genomic fragment showed poor similarity. However, the 3rd ORF of T5 (T5.003) encodes a 68aa acidic protein that has the PROSITE pattern: F-E-X-X-Y, which includes the 3 residues that participate in forming the Ung-sequestering hydrophobic pocket of Phi29 and PZA p56 variants[58,85] (Figure 5.6a).



```
                                               ▼▼  ▼
T5.003: ( 68 aa) 27 YIYQNELNG-HIYIIESGDFYS-FENEYEAK-DHLQENDIPDVW  53
Phi29 : ( 56 aa) 15 LLQDDD--GKQYYEYHKGLSLSDFEVLYGNTADEIIKLRLDKVL  56
                     : :::  : . *  . *   * **  *    *.: . .: .*
```

Figure 5.6a. Sequence alignment of T5.003 and Phi29 p56. The three residues that form the hydrophobic pocket of Phi29 p56 are conserved between both proteins (green triangles).

The T5.003 gene was amplified from T5 genomic DNA (SIGMA-Aldrich) using Q5 DNA polymerase and the primers P46 and P47. The construct pSDM4_U12_Ung (section 3.1.2) was linearised using primers P1-2. The gene T5.003 was cloned into pSDM4_U12_Ung construct via OE-PCR (section 2.1.1.11). The protein T5.003 was tested for Ung inhibition using the bacterial conditional lethal assay (section 2.1.2.7.3) and was unable to rescue the phenotype, implying that the protein is unable to inhibit Ung.

Another protein encoded in the 1st 7.9% of T5 genome (~10 kb genomic fragment that was reported to encode an UngIn sequence)[98], T5.015, was recently reported to have the ability to bind Ung and act as an endonuclease to selectively cleave uracil-containing DNA[158]. Interestingly, T5.015 does not inhibit Ung but rather requires a functional Ung to mediate its toxicity[158]. It could be that a complex of Ung, T5.015, and bacterial DNA limits the availability of Ung to act efficiently on an *in vitro* added uracil-DNA substrate, thus appearing as if Ung is inhibited in the original report[98].

The report, that suggested Ung activity inhibition by T5 phage[98], referred to pH dependent Ung inhibition by an *Escherichia* phage T5 encoded factor and that the Ung bound phage-encoded protein has a molecular weight of 10-15 kDa (compatible with T5.015 MW of 12.96 kDa)[98]. Interestingly, an AlphaFold predicted structure of T5.015 protein exhibits a negative charge distribution on one of its interfaces that resembles the Ugi negative charge distribution on its Ung binding interface (Figure 5.6b). In addition, the opposite interface of T5.015 is significantly positively charged, suggesting its role in binding to the DNA backbone (Figure 5.6b).

A prediction of the T5.015 structure in complex with Ung (structure modelled with AlphaFold-Multimer[159]) shows that T5.015 could dock to the Ung-DNA binding cleft (Figure 5.6c). Interestingly, in the AlphaFold-Multimer predicted complex structure, T5.015 binds to Ung using a β-strand and an α-helix whose sequences have similarity to the Ugi/SAUGI Ung-binding β-strand and α-helix (Figure 5.6c). The T5.015 protein is also predicted to sequester the apical residue of the Ung DNA minor groove intercalation loop. Intriguingly, the proposed Ung-binding interface of T5.015 has an (ESI) motif located in a β-strand and an acid sidechain-rich α-helix. Those 2 characteristics are ones that the heuristics approach, presented in this thesis (section 5.4.1) was built upon. The sequence similarity of these Ung binding motifs highlights the potential benefits of future applications of a heuristics-based approach to find potential Ung-binding proteins in targeted organisms.

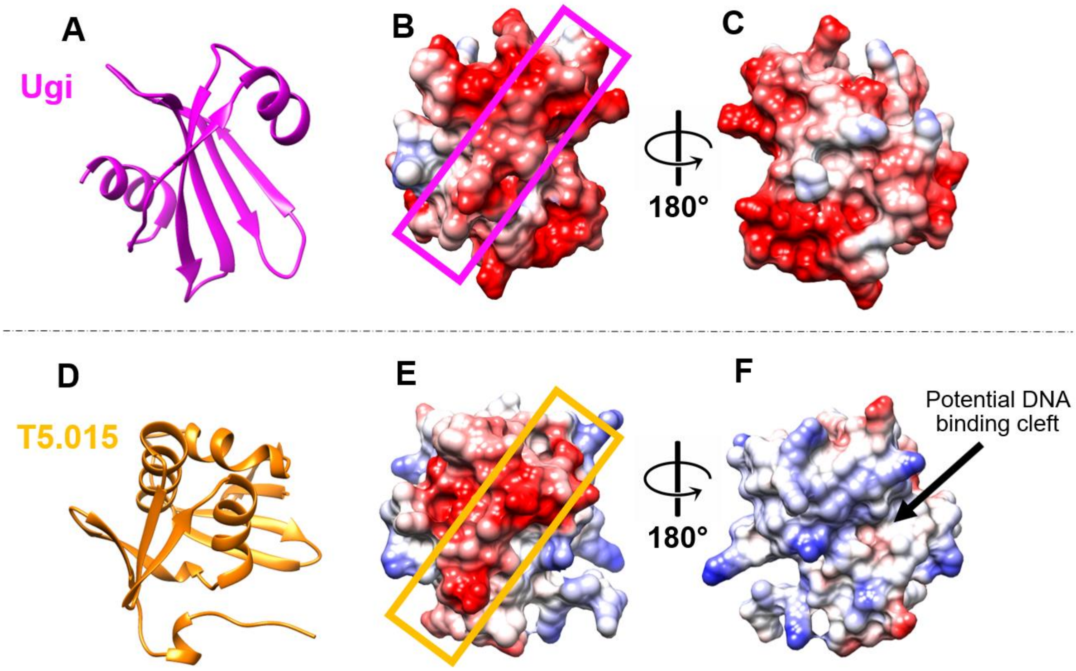


Figure 5.6b. T5.015 vs Ugi structure comparison. A) The crystal structure of Ugi (PDB: 1UDI) showing a front view of Ung-binding interface. B) Surface view of (A) with coulombic surface colouring, purple rectangle surrounds the DNA strand-mimicking structure of Ugi. C) A rotated view of (B) to show the opposite interface of Ugi. D) The AlphaFold model of the T5.015 structure (pLDDT score: 95.92). E) Surface view of (D) with coulombic surface colouring, orange rectangle surrounds a potential DNA strand-mimicking structure of T5.015 that is effectively equivalent to structure focused upon via the purple rectangle of panel B. F) A rotated view of (E) to show the opposite interface of T5.015 with an arrow indicating a potential DNA binding cleft.

The T5.015 model and Ugi structure do not structurally align. In addition, the potential Ung-binding β-strand of T5.015 is located in the C-terminal part of the protein rather than the N-terminal as seen in Ugi/SAUGI sequences. This points to possible optimisation that could be applied to the filter pipeline used in section 5.4 to increase the leniency of finding new Ugi-heuristic matches in targeted organisms. The suggested leniency was accordingly then applied during the uracil-DNA genome filtration process; however, no new heuristic matches were output from the thus modified filter pipeline.

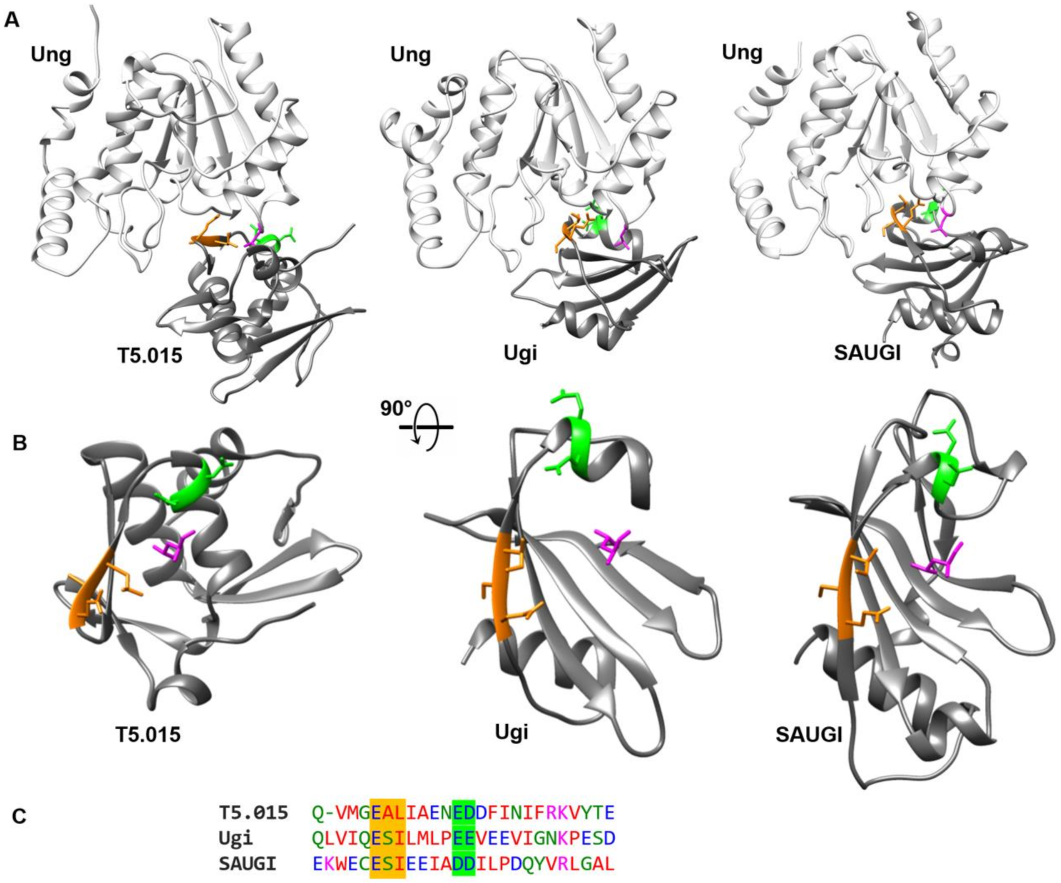

A
Ung
Ung
Ung
T5.015
Ugi
SAUGI
B
90°
T5.015
Ugi
SAUGI
C
T5.015 Q-VMGEALIAENEDDFINIFRKVYTE
Ugi QLVIQESILMLPEEVEEVIGNKPESD
SAUGI EKWECESIEEIADDILPDQYVRLGAL

Figure 5.6c. T5.015 Ugi-type heuristic conservation. (A) Ung structures (light grey) in complex with the molecules (dark grey) from left to right: T5.015 (AlphaFold-Multimer prediction), Ugi (PDB: 1UDI), or SAUGI (PDB: 3WDG). (B) Front view of Ung-binding interface of molecules shown in (A). (C) MSA of T5.015/Ugi/SAUGI Ung-binding β-strand and α-helix. Ugi/SAUGI conserved ESI motif and a potentially equivalent motif in T5.015 sequence are coloured orange in panels A-B and highlighted orange in panel C. A conserved acidic motif at the Ugi/SAUGI Ung-binding α-helix and potentially equivalent motif in T5.015 are coloured green in panels A-B and highlighted green in panel C. Ung DNA minor groove intercalation loop apical residue that Ugi/SAUGI sequester is coloured magenta in panels A-B. Interestingly, T5.015 passes the ESI filter and the ESI surrounding acidity filter that were used in the Ugi-triaged heuristics-driven search (section 5.4.1).

5.7. Conclusion

A variety of different sequence similarity tools can be used to interrogate UngIn homology at the sequence level. These tools might be not powerful in detecting distantly related homologs of proteins that have very few known sequences in the database such as Ugi (as discussed in chapter 6). In this vein, the use of general heuristics based on motifs and biophysical properties other than sequence, could also be employed. Limiting the search to targeted organisms, known or expected to have UngIn in their genomes, adds focus.

Phages that are reported or expected to encode Ung inhibitor sequences were explored in this chapter using: (1) Sequence similarity searches utilising naturally occurring (sections 5.2 & 5.6) or synthetic (section 5.3) UngIn sequences as queries; (2) Heuristics driven approaches for UngIn discovery (section 5.4); and (3) Empirical screening of targeted genomes (section 5.5). While some of the candidate genes were tested for Ung inhibition activity, others will still need to be further investigated as part of future research.

In this chapter, we were able to develop a novel method that potentially has the ability to uncover novel uracil-DNA phages in the *Myoviridae* family (section 5.5.3). Further analyses of the hits exclusively encoded by uracil-DNA phages is an interesting future avenue for investigation that might uncover novel proteins crucial for uracil-DNA synthesis or protection.

Chapter 6

6. General discussion

Ung-specific DNA mimic proteins, UngIns, have arisen in viruses at least on three independent occasions and are employed by viruses and transposable elements to modulate Ung activity as a restriction factor. Given the independent evolutionary pressure to create UngIns, these proteins might be expected to be more widespread. However, it is argued in this thesis that the sequence plasticity of UngIns has hindered unambiguous identification of novel proteins that are able to inhibit Ung.

There are relatively few deposited UngIn sequences, and assigning sequences as UngIns is challenging using simple bioinformatics sequence-based searches of genome accessions. It was demonstrated in this thesis that phylogeny-guided searches supported with insights of signature motif conservation in combination with genomic map analysis seems to be a much more powerful approach to find new UngIns (sections 3.4 & 3.5). Implementation of a variety of sequence and structural similarity search tools in combination with structure prediction tools also support phylogeny-guided searches (sections 3.4.6 & 3.6). The sequence plasticity of Vpr is less significant than the other UngIn types hence a sequence similarity searches for Vpr were not performed in this thesis.

Previous mutagenesis studies of UngIns provided limited data in which single amino acid variations were investigated in Ugi[82,160], p56[58] and SAUGI[88]. The work in this thesis shows that an empirical library-based mutagenesis approach, targeting the Ung-binding motif, was able to provide insights about tolerated mutations and uncovered novel synthetic sequences from position-specific residue sampling libraries derived from the PBS1 Ugi sequence to enrich the repertoire of sequences supporting UngIn function (section 3.2).

Work in this thesis has included solving the crystal structure of an Ugi sequence variant from phage vB_BpuM_BpSp (Ugi-2), a p56 variant from phage VMY22, and a previously unknown SAUGI variant encoded by *Macrococcus* species (MCUGI1). Analysis of these structures increases our knowledge of signature motif conservation in UngIns (chapter 4).

The expanded validated sequence repertoire from work presented in this thesis, was used, along with heuristic signatures that essentially define the contextual properties of currently known UngIns, to search in sequence space within genomes expected to encode Ung modulating sequences, in an attempt to uncover UngIn signatures (chapter 5).

A summary of the results generated from assaying variety of potential UngIns in this thesis is listed in Table 6.

The *in vivo* conditional lethal UDG assay has limitations that might led to some of the negative results. The conditional lethal UDG assay shows ability to inhibit the specific Ung carried by the used construct but not other variants of Ung. Additionally, if a transformation does not work efficiently, a false negative result might be generated. To increase the assay sensitivity, a much more efficient transformation methods, such as electroporation, could be applied.

For proteins like HI1450 and PhiR1-37 g365, protein expression/solubility could be enhanced by trying different expression strains and lysis buffers. Multiple concentrations/titrations could be done for these proteins to be assayed. In addition, alternative assays or biophysical characterisation methods could be employed to detect Ung-binding ability such as isothermal titration microcalorimetry or co-purification of candidate proteins with Ung.

Table 6. A summary of UngIn hits analyses in this thesis.

Potential UngIn	Identification method	Analyses	Analyses result
Ugi	Previously identified	Library mutagenesis and *in vivo* (conditional lethal) UDG assay	11 novel synthetic Ugi variants were identified
Ugi-2	Ugi BLAST	*In vitro* UDG assay and Crystal structure	Proved as UngIn
MCUGI1	SAUGI PSI-BLAST	*In vitro* UDG assay and Crystal structure	Proved as UngIn
MCUGI2	SAUGI PSI-BLAST	*In vitro* UDG assay	Proved as UngIn
MBUGI1	SAUGI PSI-BLAST	*In vitro* UDG assay	Proved as UngIn
SYUGI	SAUGI PSI-BLAST (limiting the search to *Staphylococcaceae*)	*In vitro* UDG assay	Proved as UngIn – novel SCC*mec* permutations
JMUGI	SAUGI PSI-BLAST (limiting the search to *Staphylococcaceae*)	*In vitro* UDG assay	Proved as UngIn – novel SCC*mec* permutations
Ac950	Searching for proteins annotated as SAUGI (DUF950)	*In vivo* (conditional lethal) UDG assay and structure prediction	Inability to inhibit Ung – Protein fold different from SAUGI
VMY22	p56 PSI-BLAST	*In vitro* UDG assay and Crystal structure	Proved as UngIn
DK2	p56 PSI-BLAST (limiting the search to *Salasmaviridae*)	*In vivo* (conditional lethal) UDG assay and genomic analysis	Inability to inhibit Ung – presence of dUTPase gene, a potential alternative strategy to replicate efficiently
DK3	p56 PSI-BLAST (limiting the search to *Salasmaviridae*)	*In vivo* (conditional lethal) UDG assay and genomic analysis	Inability to inhibit Ung – presence of dUTPase gene, a potential alternative strategy to replicate efficiently
Goe4	p56 PSI-BLAST (limiting the search to *Salasmaviridae*)	*In vivo* (conditional lethal) UDG assay and genomic analysis	Inability to inhibit Ung – presence of dUTPase gene, a potential alternative strategy to replicate efficiently
HI1450	Ugi structural homology search	*In vitro* and *In vivo* (conditional lethal) UDG assay, and structural analysis	Inability to inhibit Ung – structural differences from Ugi that would prevent Ung inhibition
Phir1-37 g365	Ugi PHI-BLAST (Limiting the search to PhiR1-37 genome)	*In vitro* UDG assay	Inability to inhibit Ung
MarsHill hit	Synthetic Ugi PHI-BLAST (limiting the search to *Myoviridae*)	Sequence alignment with Ugi	UDG assay is suggested for future research
PhiR1-37 g239	Ugi heuristics-match in PhiR1-37 genome	*In vivo* (conditional lethal) UDG assay	Inability to inhibit Ung
PhiR1-37 g280	Ugi heuristics-match in PhiR1-37 genome	*In vivo* (conditional lethal) UDG assay	Inability to inhibit Ung
LPJP1 hit	Ugi heuristics-match in LPJP1 genome	Sequence alignment with Ugi	UDG assay is suggested for future research
S6 hit	Ugi heuristics-match in S6 genome	Sequence alignment with Ugi	UDG assay is suggested for future research
T5.003	Manual alignment with p56	*In vivo* (conditional lethal) UDG assay	Inability to inhibit Ung
T5.015	Previously reported to bind Ung	Structure prediction and analysis	Different fold from Ugi/SAUGI but use very similar motifs to bind Ung

6.1. Engineered synthetic UngIn sequences

Insights from the tolerated mutations at the essential ESI motif of Ugi/SAUGI UngIn fold proved to add power to sequence similarity search tools and uncovered novel putative UngIn sequences (section 5.3). A suggestion for future work is that other libraries of synthetic mutants can be generated targeting other motifs such as surface residues, hydrophobic pocket forming residues, or protein core forming residues. Viable synthetic variants that might be obtained from such libraries would increase our knowledge of UngIn sequence plasticity and could provide greater power in bioinformatics searches to find other naturally occurring UngIn variants.

6.2. UngIn in *Myoviridae* uracil-DNA phages

Genomic analysis of Ugi variants from phages PBS1, AR9, vB_BspM_Internexus, and vB_BpuM-BpSp shows that, unlike what is shown with SAUGI and p56, the genomic context surrounding the Ugi gene is not conserved, hence such an approach could not support searching for Ugi homolog sequences in other uracil-DNA genomes in the *Myoviridae* family.

Multiple approaches were addressed in this thesis to address the enigma of PhiR1-37 and other uracil-DNA phages lacking an identifiable encoding sequence for an UngIn.

6.2.1. Heuristics-driven search within uracil-DNA phages

In this study we addressed the UngIn search via structure and biophysics-informed heuristic sequence interrogations of uracil-DNA genomes. Using this approach, a list of 367 ORFs was shortened to only 2 heuristic matches in the PhiR1-37 genome (section 5.4.1.4).

The heuristics-driven search for an Ugi-type UngIn in the other uracil-DNA genomes returned no candidate heuristic matches. However, dropping the glycine/proline count filter led to identification of 2 heuristic matches in the uracil-DNA *Staphylococcus* phage S6 and *Listeria* phage LPJP1 (section 5.4.1.5). It is suggested that further investigation of these heuristic matches could be considered as a part of future research.

6.2.2. Shotgun library screening

A shotgun cloning library would enable more extensive screening of ORFs in targeted genomes. This could be helpful especially that a completely novel type of UngIn might not satisfy the general UngIn properties exhibited by the known types. For example, the HIV protein Vpr binds different proteins using different interfaces and hence it is the only non-acidic UngIn. A new UngIn that has a high pI could be missed via application of the acidity filter in a heuristics-driven search. A shotgun library screening for PhiR1-37 was performed in this thesis (section 5.5.2), repeating the screening using other restriction enzymes increases the possibility of finding any new UngIn in case a restriction site of the enzyme used in this thesis sits within the UngIn coding gene sequence. Using higher yields of the phage genomic DNA and repeating the experiment multiple times will increase the efficiency of the experiment and possibly the potential of uncovering any novel UngIn sequences.

Empirical screening of other known uracil-DNA bacteriophages via the developed conditional lethal assay is suggested for future UngIn discovery research. Building constructs carrying the specific host Ung would probably add specificity to that assay as the affinity to Ung variants other than the specific host Ung might be low for a yet to be identified UngIn.

6.2.3. Targeting proteins exclusive to uracil-DNA phages

A novel method of identifying uracil-DNA phages was suggested in this thesis (section 5.5.3). Two *Staphylococcus* phages (PALS2, and vB_StaM_SA1) exhibit genomic signatures that imply uracil-DNA genomes according to our results (section 5.5.3). Interestingly, uracil-DNA phages have shown to be the most closely related phages to *Staphylococcus* phage PALS2[155].

Uracil-DNA *Myoviridae* phages are known to encode a non-canonical RNA polymerase that recognises uracil-containing promoters[161]. Performing a PSI-BLAST search using this RNA polymerase as an input (encoded by AR9 uracil-DNA phage; accession: YP_009283131.1) output homolog sequences from known uracil-DNA phages in addition to homolog sequences from PALS2 and vB_StaM_SA1 phages, further implicating them as putative uracil-DNA phages.

It might be the case that a novel UngIn is encoded by at least some uracil-DNA phages, and in that case such a novel UngIn will not be detected easily by performing the suggested method (section 5.5.3).

Ugi, and p56 are known to inhibit Ung from various organisms other than *Bacillus* species, including HSV1 UNG. However, if a novel UngIn is not able to inhibit HSV1 UNG, then the conditional lethal assay applied to test the hits exclusive to uracil-DNA phages may not be a suitable test (section 5.5.3). An optimised version of the assay that involves a bacterial Ung in the construct rather than HSV1 UNG would be more suitable.

Further analyses of the hits exclusively encoded by uracil-DNA phages (section 5.5.3) is an interesting avenue of future work that might uncover novel proteins crucial for uracil-DNA synthesis or protection.

6.2.4. Novel Ung modulation mechanisms in the *Myoviridae* family

It is possible that novel Ung inhibitor folds exist in *Myoviridae* uracil-DNA phages, in which case targeting matches satisfying general UngIn heuristic properties (section 5.5.1) that cannot be compared to known Ung inhibitors could still be scrutinised via *ab initio* structure prediction via the RoseTTAFold and AlphaFold methods. However, the efficiency of these methods depends on the multiple sequence alignment depth that any given protein has in a database[136,137]. Bacteriophages tend to contain many annotations of hypothetical proteins, therefore proteins encoded by a phage may not have any homologous proteins in the database, rendering RoseTTAFold and AlphaFold methods inefficient or even currently incapable of determining accurate structures of possible Ung inhibitor structural homologues in phages such as PhiR1-37. Interestingly, two recent structure prediction methods, ESMFold[162] and OmegaFold[163], have outperformed RoseTTAFold and AlphaFold significantly in predicting protein structure from a single protein sequence without any need for a MSA, these methods could thus be useful for predicting structures of PhiR1-37 orphan proteins.

There is also a possibility that these phages protect their uracil-DNA genomes in alternative ways, such as transcription modulation or directed Ung degradation (Vpr is known to promote both those processes in addition to canonical UngIn-type sequestration)[31].

Some bacteriophages of the *Myoviridae* family form a unique nucleus-like compartment (pseudo-nucleus) within the host cell, which protects phage DNA from restriction modification by separating it from the host cell cytoplasm[164]. Such a pseudo-nucleus is known to be absent during *Bacillus* uracil-DNA phage infections. Indeed, uracil-DNA is suggested to be the reason why these phages don’t need an anti-restriction pseudo-nucleus compartment[164]. Could uracil-

DNA phages infecting organisms other than *Bacilli* utilise a pseudo-nucleus as an alternative strategy to the encoding of an UngIn sequence?

Biological studies of infected cells on whether these phages form a pseudo-nucleus inside the host to protect their genomes would be an interesting line of future research.

UngIn genes encoded by phages are known to be early genes[165]. The progression of *Yersinia* phage PhiR1-37 interaction with its host was previously investigated[166]. The host Ung expression levels showed no changes between early and late times after infection. However, RNA sequencing showed that 92 genes out of 367 genes in the PhiR1-37 genome are early transcribed after phage infection[166]. Similar studies on other uracil-DNA phages infecting *Staphylococcus* or *Listeria* species, and a subsequent analysis of the mutually early transcribed genes of these phages could add focus to the search for any novel type of UngIns or other Ung modulating proteins in these phages.

6.3. Structural homology to UngIns

The protein SAUGI was identified as an UngIn initially by performing a structural homology search[62]. In this thesis, the protein HI1450 was identified as a potential Ugi structural homolog (section 3.6.1) that has the ESI motif and the ESI surrounding acidity motif (Figure 3.6.1b). The UDG assay of the HI1450 protein did not demonstrate Ung inhibition (section 3.6.1). Superimposition of HI1450 with Ugi onto an Ugi-Ung complex showed potential clashes that would prevent HI1450 from inhibiting Ung (Figure 3.6.1d). Interestingly, a previous report showed that one of the proteins bound by HI1450, when expressed in *E.coli* cells, has a molecular weight of ~26 kDa (close to MW of *E.coli* Ung: ~25.7 kDa)[48]. However, a DNA-mimicry analysis of Ugi and HI1450 surfaces shows that HI1450 might dock the Ung-DNA binding cleft using one of its α-helices rather than the Ugi-equivalent β-strand (Figure 6.3), indicating that HI1450 is unlikely to act as an Ugi-like Ung-specific DNA-mimic protein, hence showed no ability to inhibit Ung in our assay (section 3.6.1).

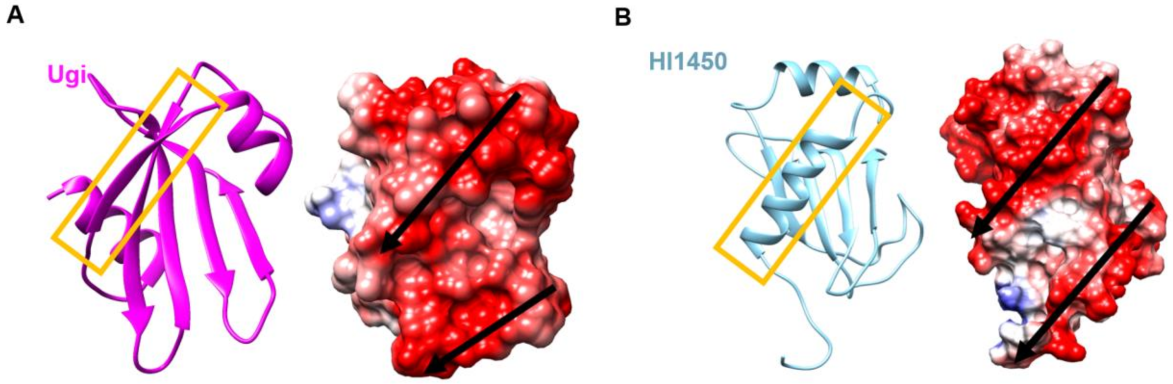


Figure 6.3. DNA-mimicry of HI1450 and Ugi. (A) Ugi structure (PDB: 1UDI): Ribbon representation on the left with orange rectangle surrounding the β-strand that docks onto Ung DNA-binding cleft; and surface view on the right with columbic surface colouring. Arrows indicate the dsDNA-mimicry (B) HI1450 structure (PDB: 1NNV): Ribbon representation on the left with orange rectangle surrounding the α-helix that potentially docks onto Ung DNA-binding cleft; and surface view on the right with columbic surface colouring. Arrows indicate the dsDNA-mimicry.

The structural homology tool that identified SAUGI is DALI, the same one that was applied in this thesis. There are other available tools to perform structural homology searches, each one of them has its own strengths and weaknesses. DALI has outperformed other methods including CE, FatCat, VAST and PDBeFold, particularly in the protein structures of the class of mixed alpha-beta[137], which is the class of Ugi, p56, and SAUGI.

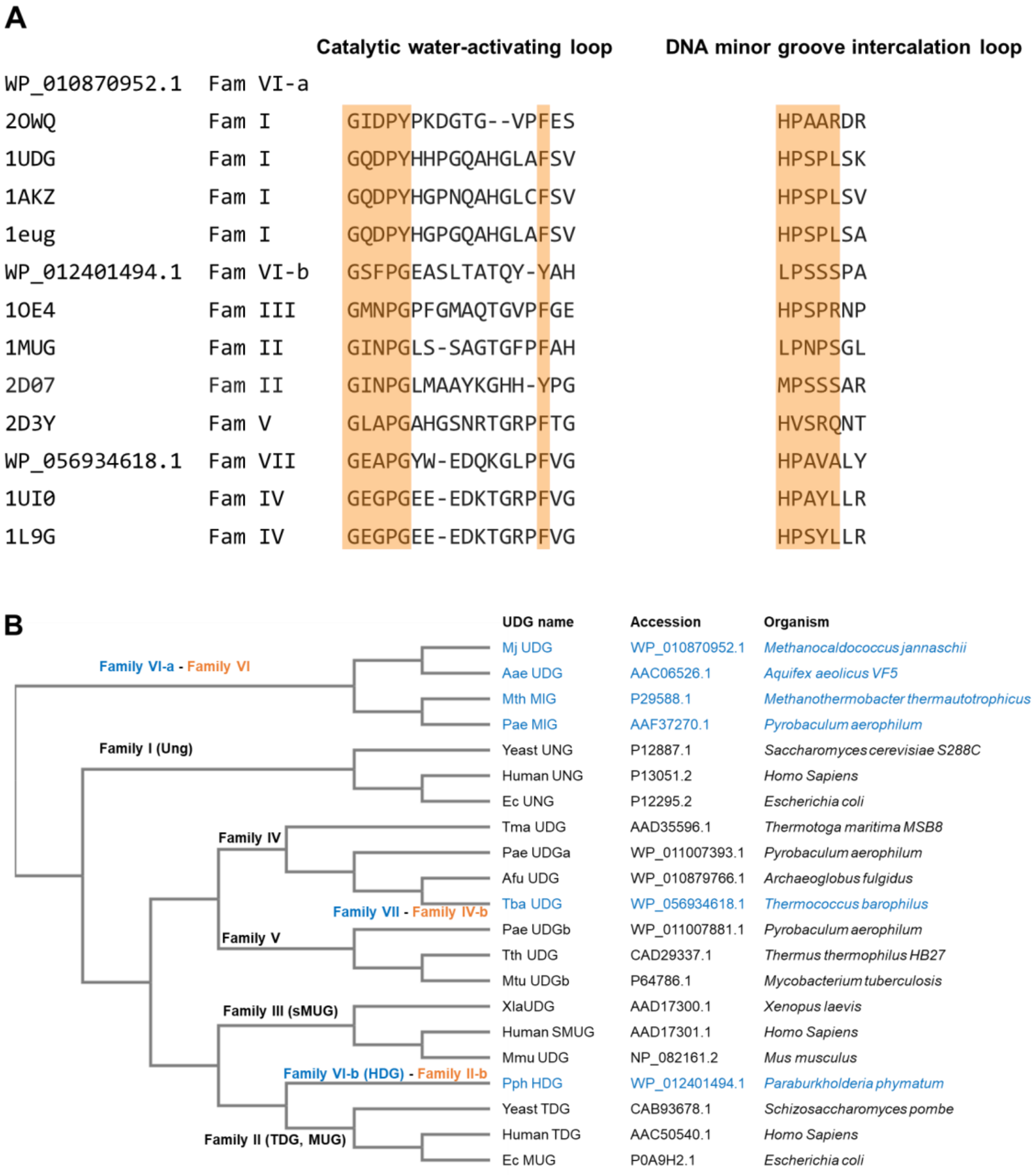
6.4. Inhibitors of other UDG families

There is a dichotomy in the classification of UDG families (section 1.3). There is a consensus view of families I-V classification. Chung *et al.* (2003)[15] suggested MjUDG as family VI UDG (designated VI-a in this section), Lee *et al.* (2011)[16] suggested HDG as family VI UDG (designated family VI-b in this section), and Zhang *et al*. (2021)[18] suggested the first identified bifunctional UDG to be classified in a novel UDG family (section 1.3). A multiple sequence alignment (MSA) of UDG conserved motifs across the families, and a phylogenetic tree analysis of UDG variant sequences from each family provide new insights (Figure 6.4).

Family VI-a UDG has distinct motifs and compose a distinct branch in a UDG superfamily phylogenetic tree, this family is suggested to be the consensus family VI UDG.

Family VI-b UDG, hypoxanthine DNA-glycosylases (HDGs), are not active on uracil-DNA but have the conserved motifs across UDG families I-V (a water-activating loop and a DNA minor groove intercalation loop). The sequences of these motifs in HDG have similarities to equivalent motifs of UDG II family (Figure 6.4, A). A phylogenetic tree analysis of UDG superfamily shows HDG branch very close to family II UDG (Figure 6.4, B). Interestingly, family II UDG has been reported to have more robust activity on xanthine than on uracil[167]. Family VI-b UDG is suggested to be classified as a sub-family of family II UDG.

Family VII UDG or Tba UDG, the first identified bifunctional UDG, shows remarkable sequence similarity to family IV UDG (Figure 6.4, A) and shows as a branch of family IV UDG in a phylogenetic tree analysis (Figure 6.4, B), this family is suggested to be classified as a subfamily of family IV UDG.



		Catalytic water-activating loop	DNA minor groove intercalation loop
WP_010870952.1	Fam VI-a		
2OWQ	Fam I	GIDPYPKDGTG--VPFES	HPAARDR
1UDG	Fam I	GQDPYHHPGQAHGLAFSV	HPSPLSK
1AKZ	Fam I	GQDPYHGPNQAHGLCFSV	HPSPLSV
1eug	Fam I	GQDPYHGPGQAHGLAFSV	HPSPLSA
WP_012401494.1	Fam VI-b	GSFPGEASLTATQY-YAH	LPSSSPA
1OE4	Fam III	GMNPGPFGMAQTGVPFGE	HPSPRNP
1MUG	Fam II	GINPGLS-SAGTGFPFAH	LPNPSGL
2D07	Fam II	GINPGLMAAYKGHH-YPG	MPSSSAR
2D3Y	Fam V	GLAPGAHGSNRTGRPFTG	HVSRQNT
WP_056934618.1	Fam VII	GEAPGYW-EDQKGLPFVG	HPAVALY
1UI0	Fam IV	GEGPGEE-EDKTGRPFVG	HPAYLLR
1L9G	Fam IV	GEGPGEE-EDKTGRPFVG	HPSYLLR

UDG name	Accession	Organism
Mj UDG	WP_010870952.1	*Methanocaldococcus jannaschii*
Aae UDG	AAC06526.1	*Aquifex aeolicus VF5*
Mth MIG	P29588.1	*Methanothermobacter thermautotrophicus*
Pae MIG	AAF37270.1	*Pyrobaculum aerophilum*
Yeast UNG	P12887.1	*Saccharomyces cerevisiae S288C*
Human UNG	P13051.2	*Homo Sapiens*
Ec UNG	P12295.2	*Escherichia coli*
Tma UDG	AAD35596.1	*Thermotoga maritima MSB8*
Pae UDGa	WP_011007393.1	*Pyrobaculum aerophilum*
Afu UDG	WP_010879766.1	*Archaeoglobus fulgidus*
Tba UDG	WP_056934618.1	*Thermococcus barophilus*
Pae UDGb	WP_011007881.1	*Pyrobaculum aerophilum*
Tth UDG	CAD29337.1	*Thermus thermophilus HB27*
Mtu UDGb	P64786.1	*Mycobacterium tuberculosis*
XlaUDG	AAD17300.1	*Xenopus laevis*
Human SMUG	AAD17301.1	*Homo Sapiens*
Mmu UDG	NP_082161.2	*Mus musculus*
Pph HDG	WP_012401494.1	*Paraburkholderia phymatum*
Yeast TDG	CAB93678.1	*Schizosaccharomyces pombe*
Human TDG	AAC50540.1	*Homo Sapiens*
Ec MUG	P0A9H2.1	*Escherichia coli*

Figure 6.4. Comparison of UDG families. A) MSA of UDG variants from each family, displaying the characteristic motifs: a water-activating loop and a DNA minor groove intercalation loop of each sequence (Family VI-a lack those conserved motifs and has instead a distinct HhH motif and a motif for an Iron-Sulfur Cluster). Highlighted orange are active site residues of the catalytic water-activating loop as well as the DNA minor groove intercalation loop. **B)** Phylogenetic tree of multiple sequence alignment for the UDG superfamily. Consensus families (families I–V) are labelled in black font, while other families are labelled with blue font (literature classification) and orange font (new classification suggested in this thesis).

The last residue of the DNA minor groove intercalation loop (Figure 6.4, panel A) is the apical residue that enters the DNA minor groove to displace deoxyuridine and form a pseudo base-pair with the partner base (section 1.4.4). This residue in Ung is sequestered by UngIns when it is a hydrophobic residue. However, some Ung variants (such as human cytomegalovirus UNG and vaccinia virus Ung) have a polar apical residue (arginine) in the DNA minor groove intercalation loop and are not inhibited by UngIns[8,168]. Furthermore, L191V and L191F mutation of this residue in *E. coli* Ung led to mutants as efficient as the wild type variant while L191A and L191G mutants retained only 10% and 1% of Ung activity, respectively[169]. This residue (leucine in most Ungs) is conserved in only family IV UDG across the other UDG families (Figure 6.4, panel A). Interestingly, Ugi was reported to cause partial inhibition of UDG activity in UDG IV encoding hyperthermophilic micro-organisms[170]. Are there any inhibitors of UDG families other than Ung? Recently, uracil-DNA phages whose host encode UDG IV rather than Ung have been identified[100]. These phages need to encode UDG IV inhibitor in order to survive. A UDG inhibitor encoded in these phages could potentially have at least partial Ung inhibition ability. It is suggested that, in future research, a conditional lethal assay model that utilises a UDG IV encoding construct could be utilised to find any UDG IV inhibitors in these phages via shotgun library screening.

It is also possible that other UDG families are inhibited by phages whose genomes use base-modifications. There are already some phages that are known to incorporate 5-hydroxymytheluracil (5-hmU) in their genomes[171]. Family V UDG is reported to have the ability to excise 5-hmU and initiate BER process[8]. Inhibitors of UDG superfamily could be much wider than it's reported in the literature, more studies are needed to explore this space.

Appendix A - Cloning and synthetic gene design

This appendix includes supplementary information for the oligonucleotide and synthetic gene sequences used in this thesis, alongside supplementary information for codon optimisation procedure, and construction of non-functional Ugi constructs (section 3.1.1) and pBpST-CAT construct (section 3.1.2)

A.1. Oligonucleotide sequences

All the oligonucleotide sequences used in this thesis are listed in Table A.1.

Table A.1. Oligonucleotide sequences used in this thesis. Upper case indicates that the sequence is a reverse complement with respect to coding sense.

Primer symbol	Primer name	5'-3' sequence	Uses/notes
P1	LBA2delst2iR	CATATGTATATCTCCTTCTTAAAG	Linearising pRSET-C via iPCR for downstream OE-PCR/ligation cloning
P2	dsblntpRS_iF	taagcttgatccggctgctaac	
P3	PTrc-BE1_5p	cataacggttccggaaatattctgaaatgagctgttg	To amplify HSV1-UNG gene from pTS106.1. BspEI and AgeI restriction sites are written in bold font.
P4	rrnB-Age_3m	CATTTATACCGGTTATTGTCTCATGAGCGGATAC	
P5	ugiBrndDS_F	nnnnnnnnnnccagaagaagtagaggaagtaattg	Ugi Library mutagenesis - Library 1
P6	ugiBrndUS_R	NNNNNNNNNNAATCACTAGTTGTTTTCCTGTTTC	
P7	ugiBrndDS_F2	NTNCTAATGTTACCAGAAGAAGTAGAGGAAGTAATTG	Ugi Library mutagenesis - Library 2
P8	ugiBrndUS_R2	ngnntcttgaatcactagttgttttcctgtttc	
P9	ugiBrndDS_F3	NNNCTAATGTTACCAGAAGAAGTAG	Ugi Library mutagenesis - Library 3
P10	ugiBrndUS_R3	nnnntcttgaatcactagttgttttcc	
P11	Bp8pS259_5p	aataaaaacattgaagatttgaataag	Ugi-2 108 Amplification
P12	Bp8pS259_3m	GTTAGCAGCCGGATCAAGCTTAGATGCTCTTGCTGTACAG	
P13	Bp8pS259_iF	aataaaaacattgaagatttgaataag	constructing pRSUgi-2 89 from pRSUgi-2 108 template
P14	3wdgSaUgi_5p	actctggaactgcaactcaaac	SAUGI Amplification
P15	3wdgSaUgi_3m	GTTAGCAGCCGGATCAAGCTTATTGGCCACCTGTGAGCAAG	

P16	McUgi1_5p	agcaaacaaatcaaagcgcatctc	MCUGI1 Amplification
P17	McUgi1_3m	GTTAGCAGCCGGATCAAGCTTACCCACGTTGACAAATTCCAAATC	
P18	McUgi2_5p	tctatcaagaagaatctgacg	MCUGI2 Amplification
P19	McUgi2_3m	GTTAGCAGCCGGATCAAGCTTACCCACGTCGAAGTTGGC	
P20	McUgi3_5p	agcttgtctgaacagctgtg	MBUGI Amplification
P21	McUgi3_3m	GTTAGCAGCCGGATCAAGCTTACCCACAGTTTCATTTTGAAC	
P22	SLnUgi_5p	catcagaaattgaaacagtatatc	SYUGI Amplification
P23	SLnUgi_3m	GTTAGCAGCCGGATCAAGCTTAAGTTCTTCAACCAGGATTTC	
P24	JMUGI_5p	aacaccaaactgaaaatctatatc	JMUGI Amplification
P25	JMUGI_3m	TTTCCTACGCGAATTCATGATTACAGGTGATTGATCAGAGTTTC	
P26	McUgi1LGI_iF	CATATGagcaaacaaatcaaagcgcatc	MCUGI1 construct correction
P27	McUgi1LGI_iR	GTTTAACTTTAAGAAGGAGATATA	
P28	AcDuf950_5p	ggttctactcgtcttttgacc	Ac950 cloning
P29	AcDuf950_3m_OE	GTTAGCAGCCGGATCAAGCTTAataacgacgaatttcgattttttttc	
P30	DK2p56_5p	attaaaaccattaaaaccaatgac	DK2 hit cloning
P31	DK2p56_3m	GTTAGCAGCCGGATCAAGCTTACAGCACCTCCTTACGATC	
P32	DK3p56_5p	caagttagcgatgttgttattag	DK3 hit cloning
P33	DK3p56_3m	GTTAGCAGCCGGATCAAGCTTACAACACCTCCTTTTCGAG	
P34	Goe4p56_5p	agcaaaaaactgatcaaattggaagag	Goe4 hit cloning
P35	Goe4p56_3m	GTTAGCAGCCGGATCAAGCTTATGCACGGCGATCAGACAG	
P36	DK3_KEE_5p	gaaaaagaagaaaaagaagaaaaagaagaattttggactc	DK3 hit motif insertion
P37	DK3_KEE_3m	gttattagcctcctggaaggtctgaaagaagaaaaagaa	
P38	E.c.HI1450_5p	gatatggatctaaacaatcgc	E.c.HI1450 cloning
P39	E.c.HI1450_3m	GTTAGCAGCCGGATCAAGCttattcccgccagatgatatg	
P40	UgiChk-F	CGCGGCCTTTTTACGGTTC	Colony PCR for Ugi library mutagenesis to exclude wild-type sequence colonies
P41	UgiChk-R	GGTAACATTAGAATTGATTCTTG	
P42	T7_to_Trc_iF	ATGTGTGGAATTGTGAGCGGATAACAAtttaactttaagaaggagatatac	Mutating T7 promoter into Trc promoter to convert pRSET-C vector into pBpST-CAT
P43	T7_to_Trc_iR	TATACGAGCCGGATGATTAATTGTCAAatttcgcgggatcgagatc	
P44	fR137365_5p	aaagctttatctattatcgatgatg	amplify g365 gene
P45	fR137365_3m	GTTAGCAGCCGGATCAAGCttacataaaGcGAataaagctctcc	
P46	T5S1_02to03_5p	gctattaaaattaatcttcccag	amplify T5.003 gene
P47	T5S1_02to03_3m	GTTTCCTACGCGAATTCATGATttaaaggccatacgctagcg	
P48	1FR1-32to38_5p	aaaataacagttcttggccctg	amplify PhiR1-37_g32
P49	g032_3m	GTTTCCTACGCGAATTCATGAttaactagaatcctttttagctatg	
P50	g034_5p	aaatttgttaaatttcgtgagaatg	amplify PhiR1-37_g34
P51	g034_3m	GTTTCCTACGCGAATTCATGATcacattttgatattttgctatcc	

P52	g038_5p	aaatctattttacgttataacgaag	amplify PhiR1-37_g38
P53	1FR1-32to38_3m	GTTTCCTACGCGAATTCATGAtcatatataaaatttcttctccg	
P54	g050c_5p	gttagaataaaatacccattaaaag	amplify PhiR1-37_g50
P55	g050c_3m	GTTTCCTACGCGAATTCATGATaactatctaattgttactgcttc	
P56	fR1-37.51_5p	gtaatcttcaaactgacggaag	amplify PhiR1-37_g51
P57	fR1-37.51_3m	GTTAGCAGCCGGATCAAGCttattccagccatttgccacg	
P58	2FR1c-52to50_5p	aagatatacaaaataaatggtaaaatg	amplify PhiR1-37_g52
P59	g052c_3m	GTTTCCTACGCGAATTCATGAtcatgcgattttagagattcttc	
P60	3FR1-65to73_5p	ctttttgcagtaaatgtatactc	amplify PhiR1-37_g65/66
P61	g065/6_3m	GTTTCCTACGCGAATTCATGAtcaataatcttttcctttacagttac	
P62	g071_5p	tctaagaaatatagtacttactcc	amplify PhiR1-37_g71
P63	g071_3m	GTTTCCTACGCGAATTCATGAttaaatcgattggtattcgttaattg	
P64	g073_5p	atgttttttgaagactatggatg	amplify PhiR1-37_g73
P65	3FR1-65to73_3m	GTTTCCTACGCGAATTCATGAttactcatcttcagtatcctc	
P66	4FR1-g088_5p	gaaagtactgaaaggctatcg	amplify PhiR1-37_g88
P67	4FR1-g088_3m	GTTTCCTACGCGAATTCATGAttactcataaatagtaaaacctc	
P68	5FR1-g098_5p	tctagcgttaagattacacctg	amplify PhiR1-37_g98
P69	5FR1-g098_3m	GTTTCCTACGCGAATTCATGAttacactacggtaaatccaatg	
P70	6FR1-110to119_5p	tttattataggttatgtagtacc	amplify PhiR1-37_g110/111
P71	g110/1_3m	GTTTCCTACGCGAATTCATGAttataatcgagatcttaaaatctc	
P72	g112/3_5p	atcgctgaaaatatatctttacttg	amplify PhiR1-37_g112/113
P73	g112/3_3m	GTTTCCTACGCGAATTCATGATtacccctccaccttaaaatattc	
P74	g117/8_5p	atgtctgctagtaaagaacttg	amplify PhiR1-37_g117/118
P75	g117/8_3m	GTTTCCTACGCGAATTCATGATacgttacgctaataggttacg	
P76	g119_5p	atgaatcaatacagcgaatcttac	amplify PhiR1-37_g119
P77	6FR1-110to119_3m	GTTTCCTACGCGAATTCATGAttagttaatcaaactcaaagaacc	
P78	7FR1-127to132_5p	tcaactgtaaatgctcaagaag	amplify PhiR1-37_g127
P79	g127_3m	GTTTCCTACGCGAATTCATGAttaaccatgaagaacagcagc	
P80	g132_5p	aatgtaataaagatagccttaataac	amplify PhiR1-37_g132
P81	7FR1-127to132_3m	GTTTCCTACGCGAATTCATGAttaaagaagattaatagacctgc	
P82	8FR1-137to142_5p	gttattataagaaattgtctagac	amplify PhiR1-37_g137/142
P83	8FR1-137to142_3m	GTTTCCTACGCGAATTCATGAttaacctgcaaaaacagttgaag	
P84	9FR1-g172_5p	acttattttgacattactgatg	amplify PhiR1-37_g172
P85	9FR1-g172_3m	GTTTCCTACGCGAATTCATGAttaattgtcagaagttgatatac	
P86	10FR1-182to189_5p	aatattatacaaattaaagagaacg	amplify PhiR1-37_g182/183
P87	g182/3_3m	GTTTCCTACGCGAATTCATGAtcatatagctgcacctataaaac	

P88	g186_5p	aaaatctttaagaatccttatgatc	amplify PhiR1-37_g186
P89	g186_3m	GTTTCCTACGCGAATTCATGAtcaataagttcactcattatttttc	
P90	g189_5p	tctgataaagaatctaaaccatc	amplify PhiR1-37_g189
P91	10FR1-182to189_3m	GTTTCCTACGCGAATTCATGATCACAAagaaccgagacgcatattac	
P92	11FR1-214to219_5p	atggcatcttttgtaactcaac	amplify PhiR1-37_g214
P93	g214_3m	GTTTCCTACGCGAATTCATGATcaaaactcgtcgtagttatcaaag	
P94	g218/9_5p	gatgatttacaatattattcatcg	amplify PhiR1-37_g218/219
P95	11FR1-214to219_3m	GTTTCCTACGCGAATTCATGAttagatattatatgtccgtttag	
P96	12FR1c-g228_5p	accgacgtaaaagttaccaatg	amplify PhiR1-37_g228
P97	12FR1c-g228_3m	GTTTCCTACGCGAATTCATGAttatgaaagttgtcttactctac	
P98	13FR1c-g236_5p	gacagtatactaaataattttgtg	amplify PhiR1-37_g236
P99	13FR1c-g236_3m	GTTTCCTACGCGAATTCATGAttacagatcctcaagattatcaac	
P100	14FR1-g239_5p:	ccatttttcttaaaagaaaatcttg	amplify PhiR1-37_g239
P101	14FR1-g239_3m:	GTTTCCTACGCGAATTCATGAttagttgtgatcaaggtcaattac	
P102	15FR1-g268_5p	gctaaactcagtcaacagctc	amplify PhiR1-37_g268
P103	15FR1-g268_3m	GTTTCCTACGCGAATTCATGAttagagtttattacttaattttaaatc	
P104	g277c_5p	ttctggagacgtaaagttgtac	amplify PhiR1-37_g277
P105	16FR1c-277to280_3m	GTTTCCTACGCGAATTCATGAttattcttctacttcatcaatag	
P106	16FR1c-277to280_5p	gaGgactttataaaaagtctatttatg	amplify PhiR1-37_g280
P107	g280c_3m	GTTTCCTACGCGAATTCATGAttatttcaagaattttagtctatag	
P108	17FR1-g284_5p	attgtaccaattatcattcatcc	amplify PhiR1-37_g284
P109	17FR1-g284_3m	GTTTCCTACGCGAATTCATGAttaactaatatctaccgaaacc	
P110	18FR1c-g291_5p	acagatattgacaaattacctg	amplify PhiR1-37_g291
P111	18FR1c-g291_3m	GTTTCCTACGCGAATTCATGAttattcttttttcttactgagttc	
P112	19FR1-318to323_5p	ctaaaagttaaaatggatgcag	amplify PhiR1-37_g318/323
P113	19FR1-318to323_3m	GTTTCCTACGCGAATTCATGAtTatagaatacttttatccataaatttc	
P114	20FR1-334to361_5p	aaGaagattaaatacgaaactaataatag	amplify PhiR1-37_g334/361
P115	g343/4_3m	GTTTCCTACGCGAATTCATGATctatttataatcaaactccactatac	
P116	g346_5p	ataactgttccagcagcaatag	amplify PhiR1-37_g346
P117	g346_3m	GTTTCCTACGCGAATTCATGATgttaataatctacttctcctaaag	
P118	g351_5p	tatatcacttcagtccaactcatg	amplify PhiR1-37_g351
P119	g351_3m	GTTTCCTACGCGAATTCATGATctaccactctatatgaaatacc	
P120	g353_5p	aaggtatattatgccatttactatc	amplify PhiR1-37_g353
P121	g353_3m	GTTTCCTACGCGAATTCATGAttacctcttaaatatagatttctc	
P122	g361_5p	atgaaatctataaatatttatgatg	amplify PhiR1-37_g361
P123	20FR1-334to361_3m	GTTTCCTACGCGAATTCATGAttatgtagatacatggttagaatc	

P124	fR137365_5p	aaagctttatctattatcgatgatg	amplify PhiR1-37_g365
P125	fR137365_3m	GTTAGCAGCCGGATCAAGCttacataaaGcGAataaagctctcc	
P126	SDM4_EcoRI_MfeI_iR	cgtaGAATTCCAATTGcttataacattttaattttattttctc	Introducing EcoRI site into pSDM4_U12_Ung
P127	SDM4_BclI_iF	cagtTGATCAcacacaggaaacagaccatg	
P128	PhiR1-37_g278_5p	CTTTAAGAAGGAGATATACATatgggtactataggaatacgag	amplify PhiR1-37_g278
P129	PhiR1-37_g278_3m	GTTTCCTACGCGAATTCATGAttattcatcatcctcaagctgttc	
P130	PhiR1-37_g282_5p	CTTTAAGAAGGAGATATACATatggctaaaaagaagcaaaaagaag	amplify PhiR1-37_g282
P131	PhiR1-37_g282_3m	GTTTCCTACGCGAATTCATGAttaagtgttatctttaggagattc	
P132	PhiR1-37_g207_5p	CTTTAAGAAGGAGATATACAT atggcaacttccactacaactag	amplify PhiR1-37_g207
P133	PhiR1-37_g207_3m	GTTTCCTACGCGAATTCATGAttaaatataagtgatatcaatacc	
P134	PhiR1-37_g234_5p	CTTTAAGAAGGAGATATACAT atgaaaataaatgataataccсttg	amplify PhiR1-37_g234
P135	PhiR1-37_g234__3m	GTTTCCTACGCGAATTCATGAttaatcttcaccagtaaccgttattac	
P136	PhiR1-37_g244_5p	CTTTAAGAAGGAGATATACATatggccactttagatgctatg	amplify PhiR1-37_g244
P137	PhiR1-37_g244_3m	GTTTCCTACGCGAATTCATGA tcaagatttaactttattgtcaac	
P138	PhiR1-37_g196_5p	CTTTAAGAAGGAGATATACAT atggcaaataatgacttacaagaaatg	amplify PhiR1-37_g196
P139	PhiR1-37_g196__3m	GTTTCCTACGCGAATTCATGAttaaaagtattttctaaattcagttattttaag	

A.2. *In silico* design of synthetic Ugi-2 gene sequence

Bacillus phage vB_BpuM-BpSp Gene “Bp8pS_259” product [Ugi-2] (Protein ID: ALN97938.1)

Protein sequence:

```
  1 MKFNISIISF IFTMIHKNNK RKHNLKRKDN SLMYKNIEDL NKFASKILET EISFEESITF
 61 TPDEVEENIG EKPNRDKICH STSLEDGRVI MlLTELEPNY TPWKLLELEE DGFKELYSKS
121 I
```

Nucleotide sequence:

```
 202377                                                              ttga
 202381 aattcaatat ctctattatt tcttttattt ttactatgat ccacaaaaac aacaagagaa
 202441 aacataatct aaagagaaag gataattcat taatgtataa aaacattgaa gatttgaata
 202501 agtttgcttc taaaatccta gaaactgaaa tatcatttga agaaagtatt acatttactc
 202561 ctgatgaggt agaagaaaat attggagaga aacctaatag agataagatc tgtcatagta
 202621 catcattaga agacggtaga gtaattatgt tattaacaga attagaacca aactatactc
 202681 cttggaagtt attagaatta gaagaagatg gatttaaaga actgtatagt aagagtatct
 202741 ag
```

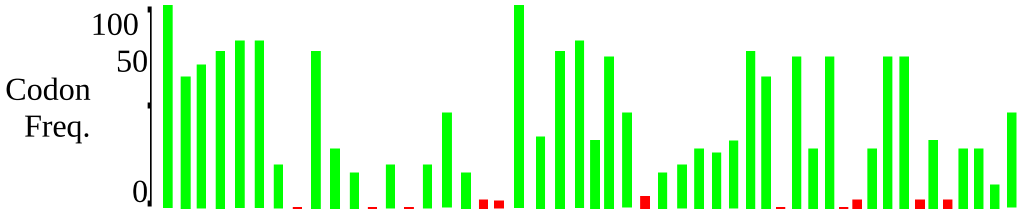
The 1st methionine and its corresponding translation codon are coloured red. The 2nd and the 3rd methionines and their related translation codons are coloured blue.

Sequence starting at the second methionine in the annotated sequence was analysed for codon usage. *E. coli* Codon Usage Analyzer 2.1 tool was used to optimise codons accordingly[101].

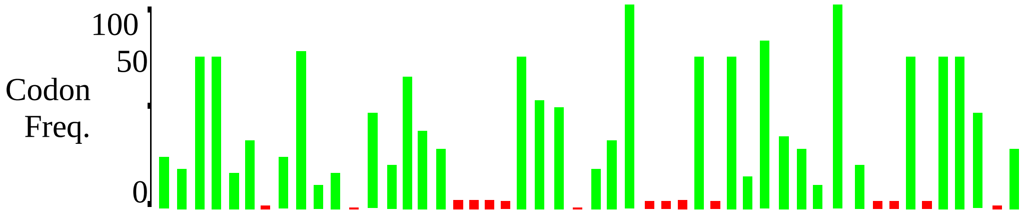
The generated report is shown in the next page:

The report generated by

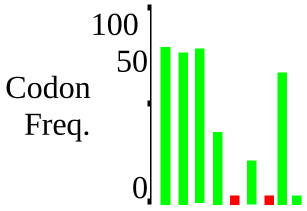
E. coli **Codon Usage Analysis 2.1** by Morris Maduro



```
             a a c a a a a a a c a c a a a g a t t a t a a a g g t a a t g t a a c g a g a t t g g a a a t a c g
       Codon t t a a a a a g a a a t a g a a a c t t a a a t a a t a a t c c a t t a c a t c t a a g t c t c c a
             g c c a c c g a a t t a g a g t t a a g t a c t a t g t g t t t a c a a t a a a t a a t t a t t t t
  Amino Acid MIHKNNKRKHNLKRKDNSLMYKNIEDLNKFASKILETEISFEESITFTPD
    Position                  10                  20                  30                  40                  50
```



```
             g g g g a a g g a c a a g a a t c a a t t g g g a g a a t t a g t g c a t a c t a t t g t g g g g t
       Codon a t a a a t g a a c a g a a t g a g c c t a a g g t t t t t c a t a c a a c c g a t t a t a a a g t
             g a a a t t a g a t t a t g c t t t a a a a c t a a t g a a a a a a a c t t t g g a a a a a a t a t
  Amino Acid EVEENIGEKPNRDKICHSTSLEDGRVIMLLTELEPNYTPWKLLELEEDGF
    Position                  60                  70                  80                  90                 100
```



```
             a g c t a a a a t
       Codon a a t a g a g t a
             a a g t t g t c g
  Amino Acid KELYSKSIZ
    Position
```

Colours: 🟥 = less than 10% of codons for same amino acid; 🟩 = at least 10%

Fraction of sense codons below threshold (=10.00): **28/108 (25%)**

-- End of report –

Considering minimal manipulation of genomic sequence, 21 out of 28 codons below the threshold were optimised. Restriction sites and mRNA secondary structures were checked and necessary amendments were made (see section 2.1.1.4 for more details).

Genomic nucleotide sequence:

```
 202377                                                              ttga
 202381 aattcaatat ctctattatt tcttttattt ttactatgat ccacaaaaac aacaagagaa
 202441 aacataatct aaagagaaag gataattcat taatgtataa aaacattgaa gatttgaata
 202501 agtttgcttc taaaatccta gaaactgaaa tatcatttga agaaagtatt acatttactc
 202561 ctgatgaggt agaagaaaat attggagaga aacctaatag agataagatc tgtcatagta
 202621 catcattaga agacggtaga gtaattatgt tattaacaga attagaacca aactatactc
 202681 cttggaagtt attagaatta gaagaagatg gatttaaaga actgtatagt aagagtatct
 202741 ag
```

Synthetic codon optimised sequence (uppercase indicates base changes from genomic sequence):

```
 202416                                       atgat ccacaaaaac aacaagCgTa
 202441 aacataatct GaagCgCaag gataattcCt taatgtataa aaacattgaa gatttgaata
 202501 agtttgcttc taaaatcctG gaaactgaaa tCtcCtttga agaaagCatt acCtttactc
 202561 ctgatgaggt agaagaaaat attggTgaga aacctaatCg Tgataagatc tgtcatagCa
 202621 cGtcCttaga agacggtCgT gtaattatgt taCtGacTga attagaacca aactatactc
 202681 cttggaagCt Gttagaatta gaagaagatg gCttCaaaga actgtaCagC aagagCatct
 202741 aA
```

A.3. Protein accessions and gene synthetic DNA sequences

SAUGI homologs

Staphylococcus aureus SAUGI protein sequence (PDB: 3WDG):

```
 1 MTLELQLKHY ITNLFNLPKD EKWECESIEE IADDILPDQY VRLGALSNKI LQTYTYYSDT
61 LHESNIYPFI LYYQKQLIAI GYIDENHDMD FLYLHNTIMP LLDQRYLLTG GQ
```

Codon-optimised DNA Sequence for SAUGI:

```
  1 atgactctgg aactgcaact caaacactat atcaccaatc tgttcaacct gccaaaggat
 61 gaaaagtggg aatgtgaatc tatcgaagaa atcgctgatg atatcctgcc tgaccaatat
121 gtacgtctcg gtgcactcag caataaaatc ctgcaaacct atacctacta ctctgatact
181 ctgcacgaaa gcaatatcta ccctttcatt ctctactatc agaaacagct catcgccatc
241 ggctatatcg atgaaaatca cgatatggat ttcctgtacc tccacaacac catcatgcca
301 ctcttggatc aacgttactt gctcacaggt ggccaataa
```

Macrococcus caseolyticus MCUGI1 protein (Accession: WP_101156358.1)

```
  1 MKQIKAHLTH YVEEILNLSS QEYLTEFIQL GIEELNWGER KIPEKLKGAI IDTYTFYNHS
 61 LIKDYIYSFI GTYQGKIILL GYTKGEYEHF FYINDTDKTL HSELHLLNLT EEDLEFVNVG
```

Codon-optimised DNA Sequence for MCUGI1:

```
  1 atgaaacaaa tcaaagcgca tctcactcgc tacctcgaag aaattctgaa actctcttct
 61 caagaatacc tgactgaatt cgtacaactc ggcatcgagg aattggcatg ggtagagcgt
121 aaaattccag agaagctcaa aggtgcaatc atcgacactt acacctttta caaccactcc
181 cttatcaaag attacatcta ctctttcatc ggtacctatc aaggcaagat catcttagtg
241 ggttacacta acggtgaata cgaacatttc ttctacatca atgatacggt caagactctg
301 cacagcgagc tgcatttgct gaatcttacg gaggaggatt tggaatttgt caacgtgggt
361 taa
```

Macrococcus caseolyticus MCUGI2 protein sequence (Accession: WP_101143899.1)

```
 1 MSIKKNLTDF VERIHRLPHY HYSVEHVQLG VEEFIIEPKV ISPSLEGKVL DTYTYYSDEL
61 EDIYSFIAYY KDTVVSIGYV KGDECYSIYL NNLEETLHDE LYLINLKVED LFYANFDVG
```

Codon-optimised DNA Sequence for MCUGI2:

```
  1 atgtctatca agaagaatct gacggatttc gtagaacgta ttcaccgtct gccacattat
 61 cattattctg tcgaacatgt tcagttaggc gtcgaagaat tcatcattga accaaaggtc
121 attagccctt ccctcgaagg taaagtactg gacacctata cctactatag cgatgaactg
181 gaggatatct actcctttat agcctattac aaggataccg ttgtcagcat cggttatgtc
241 aaaggtgacg agtgctatag catctacctg aacaacctgg aagagaccct gcacgatgag
301 ctctacctga tcaacctgaa agtggaggac ctgttctatg ccaacttcga cgtgggttaa
```

Macrococcus bohemicus MBUGI protein sequence (Accession: WP_165958605.1)

```
 1 MSLSEQLCKF VERRFKYLND IWYFEHVETT LGEIFDSKDL SGDLSADKEV DTFTYFSMTL
61 DDEHVYPFIV QDDDQIIAMG YVEEEEVKLI YLTDGKSIFI DELHLLDTNK ESVQNETVG
```

Codon-optimised DNA Sequence for MBUGI:

```
  1 atgagcttgt ctgaacagct gtgtaagttt gtagaacgtc gctttaagta tctgaatgat
 61 atctggtatt tcgaacatgt agaaaccact ctgggcgaaa tctttgatag caaggatctg
121 tctggtgatc ttagcgccga caaggaagtt gataccttta cctatttttc tatgactctg
181 gatgatgaac atgtttatcc atttatcgta caggatgacg atcagattat cgcaatgggt
241 tatgtcgaag aagaagaagt gaaactgatc tatctcacag atggtaaaag cattttcatc
301 gatgagctgc atcttctcga tactaacaag gagagcgttc aaaatgaaac tgtgggttaa
```

Salinicoccus sp. YB14-2 SYUGI protein sequence (Accession: WP_052256111.1)

```
 1 MHQKLKQYIT RHLKKSEDEY LSESFVLPST ETFQSPQFQR LFDDQSLSHQ LYYSTTDDEP
61 FFPFEVYQDD TLIALGYMEE DKQHILYLKH DDEILVEEL
```

Codon-optimised DNA Sequence for SYUGI:

```
  1 atgcatcaga aattgaaaca gtatatcact cgccacctga agaaatcgga ggacgagtat
 61 ctgtcggagt ccttcgtact gccgagcacg gaaacgttcc agtctccaca gttccaacgt
121 ctctttgacg atcagtccct ctctcatcag ctgtactatt ccaccacgga tgatgagcca
181 ttcttcccgt tcgaagttta tcaagatgac actctgattg cgctcggcta tatggaagaa
241 gataaacagc atatcttgta cctgaaacat gacgatgaaa tcctggttga agaactttaa
```

Jeotgalicoccus meleagridis JMUGI protein sequence (Accession: WP_185124884.1):

```
 1 MNTKLKIYIK KYFPELSTLT WSDEAVSMSG DELFEDTKLK SLYENESLDT RLYYPIEINS
61 AILPFEIYKE ETLVALGYTN DESQKIIYFK HGAETLINHL
```

Codon-optimised DNA Sequence for JMUGI

```
  1 atgaacacca aactgaaaat ctatatcaaa aaatactttc cggagctgtc cactctcacc
 61 tggtctgatg aagcagttag catgagcgga gacgagctgt ttgaggacac gaagctgaaa
121 tctctgtacg aaaacgagag cctggacacg cgcctctatt accctattga aatcaattct
181 gcaattctgc cgtttgagat ctacaaagag gaaactttgg ttgccctggg ctataccaat
241 gacgaatccc aaaaaattat ttattttaag catggcgcgg aaactctgat caatcacctg
301 taa
```

Acinetobacter pittii Ac950 cloned protein sequence (Accession: KQD32686.1).

```
  1 MGSTRLLTNI IQRKVMLPEE MSPSMQRDNF EVTLTDFEKH PIIKCLFKAD NQRSTECWSV
 61 QEIANFIEDC TEDQNINLCI LYWKDIHSNI YIIDGAHRLS CIYAWINRYF ADEQVPQAPN
121 FNDQQKQDIR YLRNYLGDLA DFQKICTDAE FAEKKIEIRR Y*
```

Codon-optimised DNA Sequence for Ac950

```
  1 atgggttcta ctcgtctttt gaccaatatt attcagcgta aagtcatgct gccagaagaa
 61 atgtcgcctt ctatgcagcg cgacaatttt gaagtgactc tgactgactt tgaaaagcat
121 ccaatcatca aatgtttgtt taaggcggat aaccaacgtt ctacggagtg ctggagcgtg
181 caagaaattg caaactttat tgaggattgc actgaagatc aaaacatcaa tctgtgcatt
241 ttgtattgga aagatatcca cagcaatatt tacattattg atggtgcaca tcgtctgtcc
301 tgtatctatg cgtggatcaa tcgttatttt gcggatgagc aagtccctca agcacctaat
361 ttcaatgatc aacaaaaaca agatattcgt tacctccgca attatctggg tgatctggct
421 gatttccaga aaatttgtac tgatgccgaa tttgcagaaa aaaaaatcga aattcgtcgt
481 tattaa
```

---

p56 homologs

Bacillus phage Goe4 gene "Goe4_c00070" product (Accession: AYD87716.1). Blue coloured methionine indicates the start codon of the cloned truncated sequence

```
  1 MRKYETILIN DFMSKEIVTT VKEEDYLKVV EEKEVLENTI KMYKLEHKKI EKELKEKDEE
 61 IEKLKGNNEK WEEISLKTKQ NFTKEINKKD EEIKRKNKTI DNLMKKLIKL EEKEEQLYTL
121 KYIYDVDGVV KEYEQNGMLK EDAEELIGMD SDNWNHWSLT KEERPDKDKI VEGLLIRCES
181 NEELIKELES RNENLEIENR QLLNDRRLKI GLSDRRA
```

Codon-optimised DNA Sequence for the cloned truncated Goe4

```
  1 atgaaaaaac tgatcaaatt ggaagagaaa gaagaacaac tgtacactct caaatacatc
 61 tatgatgtcg acggtgttgt taaagaatat gaacaaaacg gtatgttgaa agaagatgct
121 gaagaactga ttggtatgga ttctgataat tggaatcact ggagcctgac taaagaagaa
181 cgtcctgata aagataaaat tgttgagggt ctcctgattc gttgtgaatc taacgaggaa
241 ctgatcaaag agttggaaag ccgcaatgaa aacctcgaaa tcgaaaatcg tcaactgttg
301 aatgatcgtc gtctcaaaat tggtctgtct gatcgccgtg cataa
```

Bacillus phage DK2 gene "DK2_00007" product (Accession: AZU99760.1). Blue coloured methionine indicates the start codon of the cloned truncated sequence

```
  1 MRKLTCNLGM KWVGEDDYLK VVEEKELLKI GQDNGIKEVC KLNRELLEQD NLMKEKDEVI
 61 ERLDKENKFH NNEFKRLSQY ILNNNYQNGK LLIVDSIIAQ CEIFDKENQG LSIRCESLEE
121 EVEGLRKENI EMIKTIKTND KKDTYTLSYS YLGSDGVTIK NYRQSGLLKE EYEEMYGMDS
181 DNWLSHSLVK DRKEVL
```

Codon-optimised DNA Sequence for the cloned truncated DK2

```
  1 atgattaaaa ccattaaaac caatgacaaa aaagatactt ataccctgag ctattcttat
 61 ctcggttctg atggtgtaac cattaaaaat tatcgtcaat ctggtctgtt gaaagaagaa
```

```
121 tatgaggaaa tgtatggtat ggattctgat aattggctgt ctcatagcct cgtaaaagat
181 cgtaaggagg tgctgtaa
```

Bacillus phage DK3 gene "DK3_00008" product (Accession: AZU99806.1). Blue coloured methionine indicates the start codon of the cloned truncated sequence

```
  1 MKEKDEEIEI LKKQWEDSPF YNWYREDNVK RMLKEKDEEI EMLKEKLDKV MNDDTYLKEI
 61 ENKNKDIDNL IEKVNKLDKE NQGLSIRCES LEEEVEGLRN QIHFCKIDEL TRYMSKNYPM
121 FAGMQVSDVV ISLLEGLKEE KEEKEEKEEK EEFWTLRYNL VVNNKEKEVV QYHMIKEDAE
181 ELIGMDSDNW NKYSLEKEVL
```

Codon-optimised DNA Sequence for the cloned truncated DK3

```
  1 atgcaagtta gcgatgttgt tattagcctc ctggaaggtc taaagaagaa aaagaagaaa
 61 aagaagaaaa aggaagaaaa agaagaattt tggactctgc gttataatct ggtcgttaac
121 aataaagaaa aagaagttgt tcaatatcac Atgatcaaag aagacgctga agaactgatt
181 ggtatggatt ccgataattg gaataaatat tctctcgaaa aggaggtgtt gtaa
```

A.4. Generating non-functional Ugi mutant construct

The work reported in this section was performed by Mr Daniele Mestriner, formerly an undergraduate project student in the Savva research lab.

In order to generate a construct that carries a non-functional Ugi mutant, a library mutagenesis was performed for the construct pBUgi.8 randomly mutating the residues comprising the Ung-binding β-strand of Ugi. Primers P5 and P6 were used for this library mutagenesis. A verification colony PCR was performed for Colonies returned from transformation into NEB 5α strain of *E. coli*, using primers P40 and P41 (Table A.1). The reverse primer used for this colony PCR binds specifically to the wildtype Ugi sequence. Out of 12 tested colonies, six did not produce a PCR product and their constructs were sequenced (Table A.4).

Table A.4. Amino acid sequence of wild-type Ugi and 6 Ugi mutants obtained through library mutagenesis. Red colour indicates the residues targeted for library mutagenesis. Asterisks indicate stop codons.

Protein	Amino acid sequence
Ugi	MTNLSDIIEKETGKQLVIQESILMLPEEVEEVIGNKPESDILVHTAYDESTDENVMLLTSDAPEYKPWALVIQDSNGENKIKML
U10	MTNLSDIIEKETGKQLVIPTRSIVKPEEVEEVIGNKPESDILVHTAYDESTDENVMLLTSDAPEYKPWALVIQDSNGENKIKML
U11	MTNLSDIIEKETGKQLVIPPDQHQKK*
U12	MTNLSDIIEKETGKQLVIKSNKSLPPEEVEEVIGNKPESDILVHTAYDESTDENVMLLTSDAPEYKPWALVIQDSNGENKIKML
U13	MTNLSDIIEKETGKQLVIRFLT*
U14	MTNLSDIIEKETGKQLVIPVIIGEDQKK*
U15	MTNLSDIIEKETGKQLVIHARM*

The constructs carrying mutants U10 and U12 (designated pRSET-B-U10 and pRSET-B-U12, respectively) were expressed at small-scale and assayed for Ung inhibition using a protein-based analysis assay (section 2.1.2.7.2.); both mutants showed inability to inhibit Ung. SDS-PAGE analysis showed the expression of soluble mutants.

A.5. pBpST-CAT vector construction

A high-copy-number vector, designated pBpST-CAT was created by Dr James Horton, formerly an MRes student in the Savva research lab. This vector was created by site directed mutagenesis (SDM) of the T7 promoter sequence of pRSET-C into a Trc promoter sequence using pre-phosphorylated primers P42 and P43 (Table A.1). iPCR/ligation strategy was used to obtain the target construct (section 2.1.1.11).

Appendix B - The SAUGI HHMER and HHblits search results

This appendix includes the results of SAUGI HMMER and HHblits searches (section 3.4.1).

B.1. SAUGI HMMER results

The SAUGI HMMER search output 31 hits (Table B.1), all of which have an e-value < 0.0007. Among these hits, MBUGI, MCUGI1, MCUGI2, JMUGI, and SYUGI were identified (hits 9, 10, 12, 21, 24; respectively).

Table B.1. SAUGI HMMER search outputs ranked by E-value in ascending order

Nr	ID	Title	E-value	Ind. E-value	Bitscore	Aligned Positions
1	PTH40128.1	hypothetical protein BU619_06310 [Staphylococcus capitis]	1.40E-49	1.90E-49	178.4	111
2	WP_069812756.1	SAUGI family uracil-DNA glycosylase inhibitor [Staphylococcus equorum]	2.30E-49	2.50E-49	178	113
3	WP_193621224.1	SAUGI family uracil-DNA glycosylase inhibitor [Staphylococcus epidermidis]	3.90E-48	4.50E-48	174	110
4	2KCD_A	Chain A, Uncharacterized protein SSP0047 [Staphylococcus saprophyticus]	9.10E-48	1.00E-47	172.9	112
5	WP_002464574.1	SAUGI family uracil-DNA glycosylase inhibitor [Staphylococcus simiae]	2.80E-46	3.10E-46	168.1	112
6	WP_083280712.1	SAUGI family uracil-DNA glycosylase inhibitor, partial [Staphylococcus sp. HMSC14C08]	1.40E-34	1.50E-34	130.5	80

7	WP_031906462.1	SAUGI family uracil-DNA glycosylase inhibitor, partial [Staphylococcus aureus]	2.10E-32	2.40E-32	123.4	81
8	WP_204179878.1	SAUGI family uracil-DNA glycosylase inhibitor, partial [Staphylococcus sp. GDY8P64P]	6.60E-32	7.30E-32	121.8	78
9	WP_188017758.1	SAUGI family uracil-DNA glycosylase inhibitor [Macrococcus bohemicus]	2.70E-31	3.10E-31	119.8	109
10	WP_101143899.1	SAUGI family uracil-DNA glycosylase inhibitor [Macrococcus caseolyticus]	1.10E-30	1.30E-30	117.8	106
11	WP_241556620.1	SAUGI family uracil-DNA glycosylase inhibitor, partial [Staphylococcus aureus]	7.60E-28	8.30E-28	108.8	71
12	BAI83361.1	conserved hypothetical protein [Macrococcus caseolyticus]	5.90E-27	7.70E-27	105.6	108
13	TDM48726.1	hypothetical protein ETI06_09635 [Macrococcus goetzii]	1.60E-25	1.60E-25	101.4	108
14	WP_242443966.1	SAUGI family uracil-DNA glycosylase inhibitor, partial [Staphylococcus pseudintermed	7.20E-23	7.60E-23	92.8	58
15	PIH33210.1	hypothetical protein CTJ09_13155, partial [Staphylococcus epidermidis]	1.10E-18	1.10E-18	79.3	60
16	WP_242440208.1	SAUGI family uracil-DNA glycosylase inhibitor, partial [Staphylococcus aureus]	1.00E-16	1.10E-16	72.9	44
17	WP_162637890.1	SAUGI family uracil-DNA glycosylase inhibitor [Staphylococcus aureus]	8.80E-15	1.00E-14	66.6	95
18	MCE3367552.1	hypothetical protein [Staphylococcus aureus]	5.30E-12	5.60E-12	57.8	47
19	WP_234449352.1	SAUGI family uracil-DNA glycosylase inhibitor [Staphylococcus haemolyticus]	1.40E-11	1.40E-11	56.5	47
20	EGQ3727876.1	hypothetical protein [Staphylococcus pseudintermedius]	3.30E-10	3.50E-10	52	38
21	WP_185124884.1	hypothetical protein [Jeotgalicoccus meleagridis]	1.10E-08	1.20E-08	47.1	99
22	MBO9296959.1	hypothetical protein [Staphylococcus hominis]	2.70E-08	2.90E-08	45.8	40
23	MCC5310505.1	hypothetical protein [Staphylococcus aureus]	2.30E-07	2.50E-07	42.8	35
24	WP_052256111.1	hypothetical protein [Salinicoccus sp. YB14-2]	2.50E-07	2.90E-07	42.6	98
25	WP_240698117.1	SAUGI family uracil-DNA glycosylase inhibitor, partial [Staphylococcus epidermidis]	4.80E-07	5.20E-07	41.8	37
26	KAA2271381.1	hypothetical protein F1592_13420, partial [Staphylococcus sp. GDX7P312P]	0.0000012	0.0000013	40.5	29

27	HBK6003673.1	hypothetical protein [Staphylococcus pseudintermedius]	0.0000014	0.0000015	40.3	31
28	HAR5902093.1	hypothetical protein [Staphylococcus pseudintermedius]	0.0000036	0.0000037	39	31
29	WP_145447412.1	hypothetical protein [Staphylococcus epidermidis]	0.0000048	0.0000052	38.6	55
30	WP_186306019.1	SAUGI family uracil-DNA glycosylase inhibitor [Staphylococcus epidermidis]	0.0005	0.00055	32.1	53
31	WP_158256177.1	MULTISPECIES: SAUGI family uracil-DNA glycosylase inhibitor [Staphylococcus]	0.00067	0.00067	31.8	28

B.2. SAUGI HHblits search results

The SAUGI HHblits search output 89 hits (Table B.2), of which 14 have an E-value ≤ 1. Among these hits, SAUGI homologues MCUGI1, MBUGI, and JMUGI were identified (hits 7, 10, and 14; respectively).

Table B.2. SAUGI HHblits search outputs ranked by E-value in ascending order

	Hit	Name	Probability	E-value	Aligned cols	Target Length
1	UniRef100_A0A023UEA0	Uncharacterized protein n=1 Tax=Staphylococcus haemolyticus TaxID=1283 RepID=A0A023UEA0_STAHA	100	3.70E-54	111	185
2	UniRef100_UPI000D72C75D	SAUGI family uracil-DNA glycosylase inhibitor n=1 Tax=Staphylococcus pseudintermedius TaxID=283734 RepID=UPI000D72C75D	100	2.90E-47	86	89
3	UniRef100_A0A8I1X9L2	Uncharacterized protein n=1 Tax=Staphylococcus hominis TaxID=1290 RepID=A0A8I1X9L2_STAHO	100	1.10E-39	76	79
4	UniRef100_UPI001AEC0E13	SAUGI family uracil-DNA glycosylase inhibitor n=1 Tax=unclassified Staphylococcus	99.93	6.60E-30	61	81

		TaxID=91994 RepID=UPI001AEC0E13				
5	UniRef100_UPI00044CB1E6	SAUGI family uracil-DNA glycosylase inhibitor n=1 Tax=Staphylococcus TaxID=1279 RepID=UPI00044CB1E6	99.58	3.20E-18	53	55
6	UniRef100_A0A5B2YY29	Uncharacterized protein (Fragment) n=1 Tax=Staphylococcus sp. GDX7P459A TaxID=2608390 RepID=A0A5B2YY29_9STAP	99.03	1.50E-12	77	81
7	UniRef100_A0A0D6DR66	Uncharacterized protein n=1 Tax=Macrococcus canis TaxID=1855823 RepID=A0A0D6DR66_9STAP	98.76	5.20E-11	107	138
8	UniRef100_W1UFY1	Uncharacterized protein (Fragment) n=2 Tax=Firmicutes TaxID=1239 RepID=W1UFY1_9FIRM	97.98	1.20E-07	50	51
9	UniRef100_A0A5B2YVP5	Uncharacterized protein (Fragment) n=1 Tax=Staphylococcus sp. GDX7P459A TaxID=2608390 RepID=A0A5B2YVP5_9STAP	97.92	1.90E-07	44	47
10	UniRef100_A0A4R5Y3X6	Uncharacterized protein n=2 Tax=Macrococcus bohemicus TaxID=1903056 RepID=A0A4R5Y3X6_9STAP	96.31	0.00083	108	159
11	UniRef100_UPI00139AD538	SAUGI family uracil-DNA glycosylase inhibitor n=1 Tax=Staphylococcus aureus TaxID=1280 RepID=UPI00139AD538	96.26	0.00098	81	103
12	UniRef100_UPI0011A62BF5	hypothetical protein n=1 Tax=Staphylococcus epidermidis TaxID=1282 RepID=UPI0011A62BF5	94.35	0.048	53	56
13	UniRef100_A0A5B2YX67	Uncharacterized protein (Fragment) n=2 Tax=Staphylococcaceae TaxID=90964 RepID=A0A5B2YX67_9STAP	90.95	0.57	29	33
14	UniRef100_A0A6V7R2S6	Uncharacterized protein n=1 Tax=Jeotgalicoccus meleagridis TaxID=2759181 RepID=A0A6V7R2S6_9STAP	89.09	1	96	100

15	UniRef100_A0A815HPF6	Hypothetical protein (Fragment) n=1 Tax=Rotaria sordida TaxID=392033 RepID=A0A815HPF6_9BILA	85.95	2.6	27	41
16	UniRef100_A0A3D8YIT2	Lipopolysaccharide heptosyltransferase family protein (Fragment) n=2 Tax=Bacteria TaxID=2 RepID=A0A3D8YIT2_STAPS	84.67	3.4	24	90
17	UniRef100_A0A6C0CHD2	Uncharacterized protein n=1 Tax=viral metagenome TaxID=1070528 RepID=A0A6C0CHD2_9ZZZZ	82.54	4.9	56	330
18	UniRef100_A0A4S3J8I0	ATP-grasp domain-containing protein n=1 Tax=Aspergillus tanneri TaxID=1220188 RepID=A0A4S3J8I0_9EURO	81.76	5.6	52	243
19	UniRef100_A0A2D8UNY8	Uncharacterized protein n=1 Tax=Mesonia sp. TaxID=1960830 RepID=A0A2D8UNY8_9FLAO	80.58	6.2	39	120
20	UniRef100_A0A0C2FNL9	Uncharacterized protein n=1 Tax=Ancylostoma duodenale TaxID=51022 RepID=A0A0C2FNL9_9BILA	76.57	11	23	72
21	UniRef100_O27799	Uncharacterized protein n=1 Tax=Methanothermobacter thermautotrophicus (strain ATCC 29096 / DSM 1053 / JCM 10044 / NBRC 100330 / Delta H) TaxID=187420 RepID=O27799_METTH	76.41	11	45	53
22	UniRef100_A0A8F3HC51	Uncharacterized protein n=1 Tax=Aeromonas sp. FDAARGOS 1414 TaxID=2778063 RepID=A0A8F3HC51_9GAMM	74.81	13	27	130
23	UniRef100_A0A3C0TG72	NMT1 domain-containing protein n=1 Tax=Lachnospiraceae bacterium TaxID=1898203 RepID=A0A3C0TG72_9FIRM	72.97	15	68	172
24	UniRef100_A0A368FJY9	G_PROTEIN_RECEP_F1_2 domain-containing protein n=1 Tax=Ancylostoma caninum TaxID=29170	72.28	17	22	75

		RepID=A0A368FJY9_ANCCA				
25	UniRef100_R6D7S1	Spore coat protein YsxE n=1 Tax=Clostridium sp. CAG:594 TaxID=1262826 RepID=R6D7S1_9CLOT	71.6	18	25	226
26	UniRef100_A0A7D9MBV4	Uncharacterized protein (Fragment) n=1 Tax=Paramuricea clavata TaxID=317549 RepID=A0A7D9MBV4_PARCT	71.48	18	42	197
27	UniRef100_R6R6L9	Uncharacterized protein n=1 Tax=Firmicutes bacterium CAG:449 TaxID=1263023 RepID=R6R6L9_9FIRM	69.83	21	36	90
28	UniRef100_A0A7C6SCS9	DnaB_2 domain-containing protein n=1 Tax=Acholeplasmataceae bacterium TaxID=1898209 RepID=A0A7C6SCS9_9MOLU	69.76	21	28	143
29	UniRef100_UPI0015D4F26C	hypothetical protein n=1 Tax=Pseudomonas TaxID=286 RepID=UPI0015D4F26C	68.45	23	43	68
30	UniRef100_A0A2H9PZV8	Uncharacterized protein (Fragment) n=1 Tax=Candidatus Pacearchaeota archaeon CG_4_10_14_0_2_um_filter_31_10 TaxID=1974426 RepID=A0A2H9PZV8_9ARCH	67.32	26	23	214
31	UniRef100_A0A1E4TVY1	Uncharacterized protein n=1 Tax=Pachysolen tannophilus NRRL Y-2460 TaxID=669874 RepID=A0A1E4TVY1_PACTA	67.28	26	39	165
32	UniRef100_A0A0L7L611	Putative SET and MYND domain-containing protein 3-like protein (Fragment) n=1 Tax=Operophtera brumata TaxID=104452 RepID=A0A0L7L611_9NEOP	65.68	29	45	123
33	UniRef100_A0A2I3RAC9	Arylformamidase n=2 Tax=Pan troglodytes TaxID=9598 RepID=A0A2I3RAC9_PANTR	64.7	27	38	101
34	UniRef100_A0A554LH33	Uncharacterized protein n=1 Tax=Parcubacteria group bacterium Licking1014_17 TaxID=2017171 RepID=A0A554LH33_9BACT	64.55	32	28	137

35	UniRef100_A0A6J4XEP8	Uncharacterized protein n=1 Tax=Olavius sp. associated proteobacterium Delta 1 TaxID=698986 RepID=A0A6J4XEP8_9DELT	63.14	35	46	401
36	UniRef100_A0A3R8R4I5	Uncharacterized protein n=1 Tax=Maribacter algicola TaxID=2498892 RepID=A0A3R8R4I5_9FLAO	62.31	34	33	154
37	UniRef100_X1ASW9	Uncharacterized protein (Fragment) n=1 Tax=marine sediment metagenome TaxID=412755 RepID=X1ASW9_9ZZZZ	61.42	40	25	180
38	UniRef100_UPI001E1DDE1E	uncharacterized protein LOC123558787 n=1 Tax=Mercenaria mercenaria TaxID=6596 RepID=UPI001E1DDE1E	60.21	43	41	502
39	UniRef100_A0A0G1GVB3	Uncharacterized protein n=1 Tax=Candidatus Peregrinibacteria bacterium GW2011_GWF2_43_17 TaxID=1619068 RepID=A0A0G1GVB3_9BACT	59.8	43	41	440
40	UniRef100_A0A7R8CBG7	(salmon louse) hypothetical protein n=1 Tax=Lepeophtheirus salmonis TaxID=72036 RepID=A0A7R8CBG7_LEPSM	59.46	45	55	120
41	UniRef100_A0A8I5N495	Arylformamidase n=2 Tax=Cercopithecinae TaxID=9528 RepID=A0A8I5N495_PAPAN	59.38	46	38	96
42	UniRef100_A0A044TNL4	Sodium/hydrogen exchanger n=1 Tax=Onchocerca volvulus TaxID=6282 RepID=A0A044TNL4_ONCVO	57.97	48	64	841
43	UniRef100_A0A356AJ05	Nitrate ABC transporter substrate-binding protein (Fragment) n=1 Tax=Oscillospiraceae bacterium TaxID=2485925 RepID=A0A356AJ05_9FIRM	57.43	49	72	254
44	UniRef100_A0A2I3HR30	Uncharacterized protein n=1 Tax=Nomascus leucogenys TaxID=61853 RepID=A0A2I3HR30_NOMLE	56.42	48	39	59

45	UniRef100_E4Y1U8	Uncharacterized protein n=2 Tax=Oikopleura dioica TaxID=34765 RepID=E4Y1U8_OIKDI	56.12	57	40	158
46	UniRef100_A0A1Y1KWS9	Uncharacterized protein n=1 Tax=Photinus pyralis TaxID=7054 RepID=A0A1Y1KWS9_PHOPY	55.33	53	41	145
47	UniRef100_A0A2P8YPQ7	Uncharacterized protein n=1 Tax=Blattella germanica TaxID=6973 RepID=A0A2P8YPQ7_BLAGE	55.32	60	37	120
48	UniRef100_A0A166NNS4	Heterochromatin protein one n=1 Tax=Moelleriella libera RCEF 2490 TaxID=1081109 RepID=A0A166NNS4_9HYPO	54.85	61	30	344
49	UniRef100_A0A170VKW3	Protein fam73b (Fragment) n=1 Tax=Triatoma infestans TaxID=30076 RepID=A0A170VKW3_TRIIF	54.85	61	38	147
50	UniRef100_A0A194RAA1	SET and MYND domain-containing protein 3 n=1 Tax=Papilio machaon TaxID=76193 RepID=A0A194RAA1_PAPMA	54.67	62	41	145
51	UniRef100_A0A2J6Q811	Uncharacterized protein n=1 Tax=Hyaloscypha hepaticicola TaxID=2082293 RepID=A0A2J6Q811_9HELO	54.38	63	34	158
52	UniRef100_A0A2K9YE14	Putative lucine-rich repeat protein n=1 Tax=Cladonia uncialis subsp. uncialis TaxID=180999 RepID=A0A2K9YE14_CLAUC	53.93	65	35	135
53	UniRef100_A0A2I3LEQ4	Arylformamidase n=1 Tax=Papio anubis TaxID=9555 RepID=A0A2I3LEQ4_PAPAN	53.24	66	53	131
54	UniRef100_UPI0012FB742F	T9SS type A sorting domain-containing protein n=1 Tax=Spirosoma spitsbergense TaxID=431554 RepID=UPI0012FB742F	53.22	68	25	587
55	UniRef100_A0A453M7G7	Protein kinase domain-containing protein n=1 Tax=Aegilops tauschii subsp. strangulata TaxID=200361	49.08	88	26	159

		RepID=A0A453M7G7_AEGTS				
56	UniRef100_A0A7Y5I250	Uncharacterized protein n=1 Tax=Pseudanabaena biceps TaxID=927669 RepID=A0A7Y5I250_9CYAN	48.04	82	28	331
57	UniRef100_UPI00117C660E	hypothetical protein n=1 Tax=Paenibacillus sp. VTT E-133280 TaxID=1986222 RepID=UPI00117C660E	47.93	94	37	84
58	UniRef100_A0A7T8E0H1	Uncharacterized protein n=2 Tax=unclassified Actinomyces TaxID=2609248 RepID=A0A7T8E0H1_9ACTO	47.29	98	28	649
59	UniRef100_A0A073K662	MerR family transcriptional regulator n=1 Tax=Bacillus gaemokensis TaxID=574375 RepID=A0A073K662_9BACI	46.3	81	45	269
60	UniRef100_A0A2K9VD80	Uncharacterized protein n=1 Tax=Lactobacillus phage Semele TaxID=2079433 RepID=A0A2K9VD80_9CAUD	45.75	100	44	103
61	UniRef100_A0A8J1TY57	Ofus.G071847 protein n=1 Tax=Owenia fusiformis TaxID=6347 RepID=A0A8J1TY57_OWEFU	44.81	110	33	127
62	UniRef100_W8Y1P4	Uncharacterized protein n=1 Tax=Bacillus thuringiensis DB27 TaxID=1431339 RepID=W8Y1P4_BACTU	44.26	120	41	166
63	UniRef100_A0A0R3RMM6	Sodium/hydrogen exchanger n=1 Tax=Elaeophora elaphi TaxID=1147741 RepID=A0A0R3RMM6_9BILA	41.3	120	61	625
64	UniRef100_A0A6S7FP12	DDE Tnp4 domain-containing protein n=1 Tax=Paramuricea clavata TaxID=317549 RepID=A0A6S7FP12_PARCT	40.85	140	61	551
65	UniRef100_A0A7S2L994	Hypothetical protein n=1 Tax=Leptocylindrus danicus TaxID=163516 RepID=A0A7S2L994_9STRA	40.5	150	71	274

66	UniRef100_UPI001C0F79B1	ABC transporter permease subunit n=1 Tax=Paenibacillus phytorum TaxID=2654977 RepID=UPI001C0F79B1	39.45	150	50	131
67	UniRef100_UPI0018F733CD	hypothetical protein n=1 Tax=Antrihabitans sp. YC3-6 TaxID=2799499 RepID=UPI0018F733CD	39.21	160	53	149
68	UniRef100_A0A0J9VT11	Uncharacterized protein n=2 Tax=Fusarium oxysporum f. sp. lycopersici (strain 4287 / CBS 123668 / FGSC 9935 / NRRL 34936) TaxID=426428 RepID=A0A0J9VT11_FUSO4	35.7	200	25	542
69	UniRef100_K4Q1E9	ATG13 domain-containing protein n=1 Tax=Beta vulgaris TaxID=161934 RepID=K4Q1E9_BETVU	35.62	200	71	891
70	UniRef100_A0A067BYT3	Uncharacterized protein n=1 Tax=Saprolegnia parasitica (strain CBS 223.65) TaxID=695850 RepID=A0A067BYT3_SAPPC	34.71	210	64	111
71	UniRef100_A0A5J4TCY2	Uncharacterized protein (Fragment) n=1 Tax=Streblomastix strix TaxID=222440 RepID=A0A5J4TCY2_9EUKA	34.22	210	57	387
72	UniRef100_A0A7C3N471	ROK family protein n=1 Tax=Armatimonadetes bacterium TaxID=2033014 RepID=A0A7C3N471_9BACT	33.97	220	29	359
73	UniRef100_A0A7S2CEE1	Hypothetical protein n=1 Tax=Florenciella parvula TaxID=236787 RepID=A0A7S2CEE1_9STRA	32.49	240	54	289
74	UniRef100_A0A448WS79	Hypothetical protein n=1 Tax=Protopolystoma xenopodis TaxID=117903 RepID=A0A448WS79_9PLAT	32.32	240	41	290
75	UniRef100_A0A1Z4R7I5	Serine/threonine protein kinase with two-component sensor domain n=1 Tax=Calothrix sp. NIES-4101 TaxID=2005461 RepID=A0A1Z4R7I5_9CYAN	32.21	210	76	410
76	UniRef100_A0A1J9SQE8	Choloylglycine hydrolase n=1 Tax=Bacillus albus TaxID=2026189	31.67	200	37	138

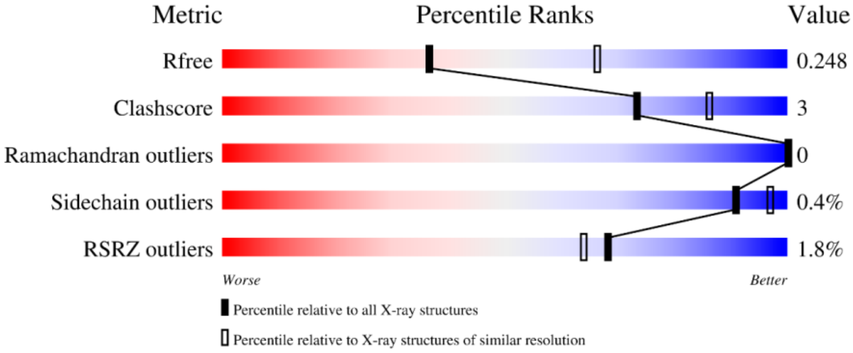
		RepID=A0A1J9SQE8_9BACI				
77	UniRef100_A0A7C4RIP1	ATP-binding protein (Fragment) n=1 Tax=Candidatus Bathyarchaeota archaeon TaxID=2026714 RepID=A0A7C4RIP1_9ARCH	31.54	250	34	254
78	UniRef100_A0A7M3Q9K0	Hypothetical protein n=1 Tax=Spirometra erinaceieuropaei TaxID=99802 RepID=A0A7M3Q9K0_SPIER	31.18	260	52	158
79	UniRef100_A0A1V5Q1H5	CDP-Glycerol:Poly(Glycerophosphate) glycerophosphotransferase n=1 Tax=Firmicutes bacterium ADurb.Bin354 TaxID=1852889 RepID=A0A1V5Q1H5_9FIRM	30.59	190	45	513
80	UniRef100_A0A662GZZ0	Uncharacterized protein (Fragment) n=1 Tax=Thermoprotei archaeon TaxID=2250277 RepID=A0A662GZZ0_9CREN	28.32	310	46	720
81	UniRef100_A0A1M7YQY6	Cyclic di-GMP phosphodiesterase Gmr n=1 Tax=Vibrio quintilis TaxID=1117707 RepID=A0A1M7YQY6_9VIBR	26.94	310	60	652
82	UniRef100_A0A815NC13	Hypothetical protein n=1 Tax=Rotaria sordida TaxID=392033 RepID=A0A815NC13_9BILA	26.69	350	37	165
83	UniRef100_A0A2N5ZT46	Uncharacterized protein n=1 Tax=Candidatus Parcubacteria bacterium TaxID=2762014 RepID=A0A2N5ZT46_9BACT	23.28	440	63	252
84	UniRef100_A0A815TY15	Hypothetical protein n=1 Tax=Adineta steineri TaxID=433720 RepID=A0A815TY15_9BILA	22.32	480	35	1178
85	UniRef100_A0A6S6XQN8	NMT1 domain-containing protein n=1 Tax=Ruminococcaceae bacterium BL-6 TaxID=2799561 RepID=A0A6S6XQN8_9FIRM	22.04	490	70	257

86	UniRef100_A0A7S0L0P4	Hypothetical protein n=1 Tax=Coccolithus braarudii TaxID=221442 RepID=A0A7S0L0P4_9EUKA	21.75	500	47	327
87	UniRef100_A0A8A1BWC0	Uncharacterized protein n=1 Tax=Marinobacterium georgiense TaxID=48076 RepID=A0A8A1BWC0_9GAMM	21.32	510	45	305
88	UniRef100_UPI001CBEF2D6	SUMF1/EgtB/PvdO family nonheme iron enzyme n=1 Tax=Synechocystis sp. PCC 7339 TaxID=2782213 RepID=UPI001CBEF2D6	21.07	530	35	623
89	UniRef100_A0A267FVY7	Uncharacterized protein n=1 Tax=Macrostomum lignano TaxID=282301 RepID=A0A267FVY7_9PLAT	21.03	530	56	412

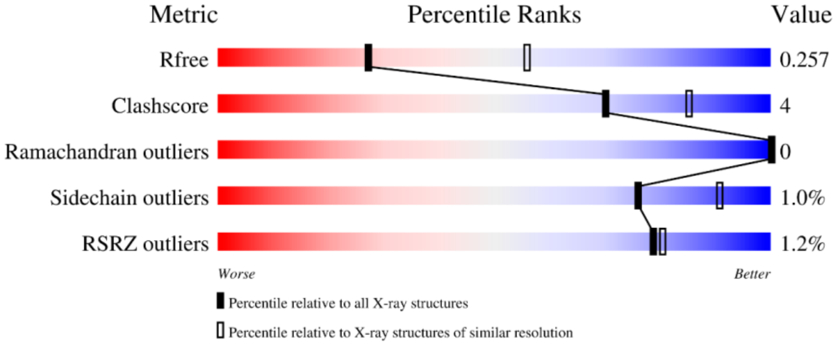
Appendix C - Overall quality of the deposited structures

This appendix articulates the quality of the X-ray crystallography structures presented in chapter 4. Percentile scores for validation metrics of are shown in the following graphics

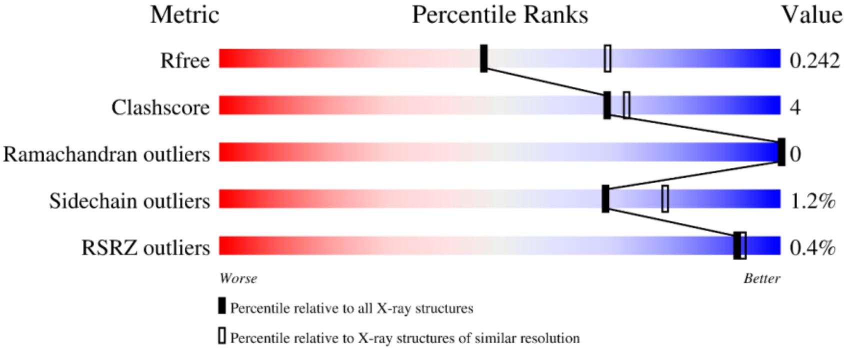
1. Ugi-2 structure in complex with SAUNG (PDB ID: 8AIM; section 4.2.3)



2. MCUGI1 structure in complex with SAUNG (PDB ID: 8AIN; section 4.3.4)



3. VMY22 p56 structure in complex with BwUng (PDB ID: 8AIL; section 4.4.1)



Appendix D - Heuristics-driven approach generated matches

All the information in this appendix is relevant to section 5.4.

D.1. Ugi-heuristic matches from *Myoviridae*

All the matches that were generated from applying heuristics-driven search for Ugi-type UngIn in the *Myoviridae* family are listed in Table D.1.

Table D.1. Matches generated by applying Ugi parameters set on *Myoviridae* genomes. Matches annotated in their respective genomes as Ugi are listed in bold font.

Phage name	accession number	Heuristic match sequence
Clostridium phage JD032	MK473382.1	LVELKKEYIDKTIKLLDEVRELKVRKEVLNDEIEFLKKNEKLSSINFNELGFPVRSGNYGIDDMIINSQEKIYIKETEIEIIDSRIEMINIYTKRLSEEEQEIISLRHFDPKINSYGEISELLMISKTVVQRKYTDALRKITLMKYGEEAKEDRE
***Bacillus* phage vB_BpuM-BpSp**	**KT895374.1**	**VIIIEILKFNISIISFIFTMIHKNNKRKHNLKRKDNSLMYKNIEDLNKFASKILETEISFEESITFTPDEVEENIGEKPNRDKICHSTSLEDGRVIMLLTELEPNYTPWKLLELEEDGFKELYSKSI**
***Bacillus* phage AR9**	**KU878088.1**	**VNYIKIGIIERENLNWEYTSNTKIRRNFNMTNLSDIIEKETGKQLVIQESILMLPEEVEEVIGNKPESDILVHTAYDESTDENVMLLTSDAPEYKPWALVIQDSNGENKIKML**
Clostridium phage phiC2	DQ466086.1	VHAEHIRLCSRSHLYFTSERLQNTTFFLCNCLKNKLKEETMTEYIEVGRRIFFDEEGEIIFYEGQSKGNVPERKNIKKIEYIDLEYDYVDYDKYKIIGIDIRTKQPILEEIPVYMSEEEK
Myoviridae sp. isolate ctbc_4	MH622943.1	VNGVENMTDVYDDVEKYVKEAVSNNTEMSSSAKYCFDMAIKTRTFTNTTHSMTWLQKAMSYSVGVHHSEYKRLFENEEVIVPELDEGDVITVVPKVWWKLGFSREKGFFSRYLVDCNKKI
Salicola phage SCTP-2	MF360958.1	LSKVYTMEEEFTSEQIPDELYDELNSSFGSVRTDLVVEYDENDPNKPQVNQNWRSFKSITTSSGDNQLVTQKAVNPETFEVEEMEGYNTFSFKVRVEFSENEDPILRQVLGFFTFTPVNK

***Bacillus* phage vB_BspM_Internexus**	**MW749003.1**	**VNYIKIGIIERENLNWEYTSNTKIRRNFNMTNLSDIIKKETGKELVIQESILMTPEEIEEVIGNKPESDILVHTAYDESTDENVMLLTSDAPEYKPWALVIQNNNGENKIKML**
Bacillus phage vB_BspM_Internexus	MW749003.1	VDVANLLEDTVKSGTLLNIAIWICKYNREIVTKDKTKKPCVLYVTQENSIRETLQRIFVHSTGSNIENYTEEEALRIIREEVIGIDENGEAPIDLFIKYRPNKSISTSDLDTMCDDLELEGYEVVCLIQDYTKRIRSSSYNPSDLRLELGSVVDDFTVA
Pelagibacter phage HTVC008M	KC465899.1	IKSINVQSKVTVKTRSVMTIKTKEDIIETIEYAIKDIKSGMEESGIAELEDLVKDIKDPKMMTVDLKKQYDWRTDYDWTDLNDHPVELPDGWVKV
Vibrio phage ValB1MD	MK387337.1	VKMSKEIEEAQIEEMKEVVNEESVESEGGEDTNVLPGDSQIVNIFGRKYLADSVSPRAHAIVKDLGTVDEEIKRLETSMRVAQLARSALVINFQEESKNFTEIEVQVLKK
Lake Baikal phage Baikal-20-5m-C28	MG198570.1	INVSTVISRDIRQIKLINGEELLTEVIGEDSLELFIRMPLKVVKEKVTMGEMNREANMFTNWMSFSDSEDFIISRVNVLVESAVNVSVARHYLEMTENIDQDHVTRVNSNKDVPNPEKLREIARAIASQLLDDEEPTYH
Yersinia phage fHe-Yen9-01	KY593455.1	IQRSKMQEGKFYSFNEAFRADFIEDNSTNENMLRLIEEGGGSFEVLEMVTEDTGKYVRRVRMKNGEVYDADIPGDEYFELSNWEFKYFVEVVGLTTEGAQSMSLVVTRENAEEMIELIKKAFNK
Proteus phage SJ_PmiM	MW367898.1	LFLILIGMCAFNYRKNLMFLMNREEALEAMKDGHKIMHEYFSPEEHFYMVNGNIIDESGYDWNFLFFKRDMYEKGWAIKE
Providencia phage PSTRCR_127	MW358927.1	INNITSLVNRLFDNGKTPMEMFGVNWNQIEELEKTMETIQFECVEKPDDVDWYTKGKIYDAVNSRVNSICEDHYVKTDDGVLGLISNVNGVYMSLLIHDIKFRRV
Providencia phage PSTRCR_121	MW385300.1	INNITSLVNRLFDNGKTPMEMFGVNWNQIEELEKMMETIQFECVEKPDDVDWYTKGKIYDVVTSRVNSICEDHYVKTDDGVLGLISNVNGVYMSLLIHDIKFRRV
Cronobacter phage vB_CsaM_GAP32	JN882285.1	IMRYEAFMGTDEFLYSGESGEISTKMLKFMEAITDENEAGTFEIKEEVVGGIPLYYASVCFPDWAEALFLIQFYEYSWELLEES
Cronobacter phage vB_CsaM_GAP32	JN882285.1	VFVCQFTTTKIRINYMSKVLKHVSLVTENCEVFTLNGSDLIWTDYQKQDDIRNPFERSMSEVLGLCFSKDSQILLEKSFISSLQFNKRTDITWVEFVFDNDETEIVTICWPEGEENRYLTEHPGQRWFVTPEGNFMFQSWYATDKERMINLDESIYDLRAMDKKA
Vibrio phage 1.193.O._10	MG592562.1	MEYFHFKKVNDLRAFLAKRHEDINLNASISQPSFYFVTTDKVVHRATGTKMLAQVLNDLFGNYYSVEDSVTVTRQRGTLAYALQEKQEKVETPVQEEKEEVITEPEVEEVVEDSPSLISLEVEEDVTEDSKVPDWAWIESLENTPEDKLELDKYAESEFSVKLSRTMKLGNMVKKFKEELAKR
Vibrio phage 1.187.O._10	MG592553.1	MEYFHFKKVNDLRAFLAKRHEDINLNASISQPSFYFVTTDKVVHRATGTKMLAQVLNDLLGNYYSVEDSVTVTRQRGTLAYALQEKQEKVETPVQEEKEEEEVITEPEVEEVVEDSPSLISLDVEEDVVAEGSKEPDWAWIESLENTKDDKNEFDKYCESEFSVKLSRTMKLSNMIKKFKEELAKR
Acinetobacter phage Ab_121	MT623546.1	VDFDSINQQIENFIESRVNEFKGCDPSALMDIFENPIYAEKLEEFMEKCGPDIDDDIFAKRINQHFK

		EIGFETEVNVGKIFDEALCIQLVMQPFFVNALKEFVGYQK
Synechococcus phage B3	MN695334.1	IKTLNKTLKNPTQMKISRKSIIDDRSAIGHFILSAITEVRPEVFETIDRGDDEFEVCMTFNGVEIPIDVVINKWVERNRQGINKAAVNLIAEKLDVINDKIRQAEEEIQEKLSDLKSDIAESFNLQYNSWDDYFYENDDHA
Vibrio phage RYC	AP014858.1	MAQRLFFRNKALLVNALKEYFPTLDVEKSKSRIRSRFSLVFGEEVIYSGTVSEMVSNLNKVAGKDIFVKRAYLKGKSGYVIFLNEDPKELVVEDAVQEQEEKSEVEKTEIDKEVVAPDAVDWEWVEGLGNTKEDKLALDQYAEKFDVKLSRTMKIENMVAKFKEALEAK
Marine virus AG-345-E15 Ga0172270_11	MH319741.1	IRDLESEIQKLTEQLADRNTEHEKLTTFKDKLSSTYEELSSRKDTISYYDFMYSLLRDGGVKTKIIKKYLPLINQQVNRYLQKMDFYINFTLDEEFNETVQSPIHEDFSYASFSEGEKMRIDLALLFTWREVARMKNSVNTNLLIMDEVFDSSLDGMGTDEFLKIIRYVIKDTNIFVISHKPEMHEKFESMIRFEKVKGFSRMVEQ
Aeromonas phage Aszh-1	MN871442.1	MKMFSTVTSLLTVRNIKFFEFYIKDISSGELIWFTYDGFAYLFKKDTNEFIDCEIDYDDPEEPQRVVDKFINSPCDLPHRFSLVDQIDQLQEELKDRLYQDFRFNRDMTDKK
Bacillus phage Shbh1	KU640380.1	IRRMMKVHEISLVKYAGEYWVTLADFSHTRDVSNYADHASVKSAIRTFVVRNNPDKYISFRGEQQIKNLITENKTNNLFILPHFQGHTRTALIHWSILEELTDKFPTLEEYEEDFENFIEEAKEFMDQPKPVQDPESSVIGERAAVIHSLRTQIQRLDEEIKVRQSNREKILQAINALENLDVTEV
Pseudoalteromonas phage HM1	KF302034.1	IMSDNTLQKIRNLEQDIVDLQQLLRVESSFRSKYELEVDAIYKTLIKMGMLETFKEEHFSGGREDDVEEFTHFMGESPLLGEDISVNMSGYTLEEHKQRVAALREFNEGE
Escherichia phage vB_EcoM_G8	MK373787.1	IMDFFTPEANQKNINKFFSIASTITRQLETALLCMETVENIHTYPFKNICGWEGYKIVISLREVKCAYSPTDKEIYQQKCDEIVNTPKEETTLEELMECLDDSPEPVEIRPEVIALEKAYKEVLEISNKAQKEYEQAKKIWEESVNRLDRLEQALQLIK
Synechococcus phage B23	MN695335.1	IKTLNKTLKNPTQMKISRKSIIDDRSAIGHFILSAITEVRPEVYETIDRGDDEFEVCMTFNGVEIPIDVVINKWVERNRQGINKAAVNLIAEKLDELHVKILQAEQEIQEKLSDLKSDIAESFNLQYNSWDDYFYENDDHA
Aeromonas phage AsSzw2	MN871441.1	MKMFSTVTSLLTVRNIKFFEFYIKDISSGELSWFTYDGFAYLFKKDTNEFIDCEIDYDDPEEPQRVVDKFINSPCDLPHRFSLVDQIDQLQEELKDRLYQDFRFNRDMTDKK
Acinetobacter phage TAC1	MK170160.1	VDFDFINQQIENFIESRVNEFKGCDPSALMDIFENPIYAEKLEEFMEKCGPDIDDDIFAKRINQHFKEIGFETEVNVGKIFDEALCIQLVMQPFFVNALKEFVGYQK
Escherichia phage T2	LC348380.1	MNLIKIKQLFVNYEFFTPETNQKNINKFFSIASTITRQLETALLCMETVENIHTYPFKNICGWEGYKIVVSLREVKCAYSPTDKEIYQQKCDEIVNTPKEETTLEELMECLDDSPEPVEIRPEVIALEKAYKEVLEISNKAQKEYEQAKRIWEESVNRLDRLEQALQLIK
Aeromonas phage CC2	JX123262.1	MKMFSTVTSLLTVRNINFFEFYIKEISSGELSWFTYDGFGYLFKKDTNEFVDCEIDYEDPEEPQQV

		VDKFINSPCDLPHRFSLVDQINQLQEELKDRLYQDFRFNRDMTDKK
Aeromonas phage AS-yj	MF498774.1	IMTDVMRIRFLSEKDKEHFVSRSVNANTHIANHMGMVWNRVSFDGRRWYLVDENNNEVVVDDGDVDSFIHPSEYQFFEWDILPIEKKKSIKELWDIAQQKKAEYDDAMTEYNKAVMEKLDESAI
Vibrio phage 1.161.O._10	MG592529.1	MEHFHFKKINDLRAFLSKRHEEINLNASISQPSFYFVTKDGKVHRATGTKMLAQVLNELFGNFYSVEDSITVTRQRGTLAYTLQEQEKVVKAEVVVPVEEVKDVVSEPEVLEELAVDESVAEVAVEEVVEDAKEPDWAWIESLENTKEDRIELDRYAEEEFSVKLSRTMKLANMVKKFKEELAKR
Aeromonas phage AS-szw	MF498773.1	IMTDIMRIRFLSEKDKEHFVSRSVNANTHIANHMGMDWNRVSFDGRRWYLVDENNNEVVVDDGDVDSFIHPSEYQFFEWDILPVEKKKPIKELWDIAQQKKAEYDDAMTEYNKAVMEKLDESAI
Myoviridae sp. isolate 131	MN856013.1	MKYPEYSGTYYDKFKVTEIDYGVVTVKFFSELHEDWIEAWAQTEVEEITLHVGDSLRPEHDHDEDFYIGQLKDNVAWELEASSIDQEFGSKFIFNHHIQQINQVLQDDFADKL
Campylobacter phage C2	MG065655.1	VSDMAKIRKLSSNQVVKDFESGEILYVRDADDEGEDLVMLVGHAEDGCIQAVFLETAMVHWIELDLKARKPKAEILIYED
Synechococcus phage ACG-2014d isolate Syn7803C45	KJ019028.1	MKHHIPDEIRANCFSCFTSLNAAERACVLLGDEAYRESLDLENDDAPCWQIPSGEHSTFAGWNPQCVPTIEYIVWKLKNREGIIKGEIY
Escherichia phage APCEc02	KR698074.1	VSDMAKIRKLSNNQVVKDFESGEILYVRDADDEGEDLVMLVGHAEDGCIQAVFLETAMVHWIELDLKARKPKAEILIYED
Salmonella phage STML-13-1	JX181828.1	IMSVKMKGVEFNAYYNDDEYWVKNAWHDDHCVKVNGEYREELDENIPDDADVVIESGTVYIPVEGENGAEEKDISLVNHFKTWRKQKNFSFIVVTVKKDKLAEVREAMRCIPGVIEVKGN
Escherichia coli O157 typing phage 14	KP869112.1	VSNMAKIRKLSNNQVVKDFESGEILYVRDADDEGEDLVMLVGHAEDGCIQAVFLETAMVHWIELDLKARKPKAEILIYED
Escherichia phage naswa	MN850595.1	VSNMAKIRKLSNNQVVKDFEPGEILYVRDADDEGEDLVMLLGHAAEDGCIQAVFLETAMVHWIELDMKARKPKAEILIYED
Escherichia phage V18	KY683736.1	VSNMAKIRKLSNKQVVKDFEPGEILYVRDADDEGEDLVMLLGHAAEDGCIQAVFLETAMVHWIELDMKARKPKAEILIYED
Serratia phage Muldoon	MN095771.1	MKNQNSVRVFTPNTYVLCMEFFYGDDDFQHNHRCTFDPKKYSLDMLREFVTDAKEVTDEQPEELPEWFTKKWDHLVQLLKSCEVWWTLAYVDVWYVDEVGAPWSLEKL
Enterobacteria phage vB_EcoM-FV3	JQ031132.1	VSNMARIRKLSNNQTVKDFEPGEILYVHDADDEGEDLVMLLGHAAEDGCIQAVFLETAMVHWIELDLKARKPKAEILIYED
Escherichia phage naam	MN850630.1	VSDMAKIRKLSNNQIVKDFEPGEILYVRDADDEGEDLVMLLGHAAEDGCIQAVFLETAMVHWIELDMKARKPKAEILIYED
Campylobacter phage D#	MG065647.1	VSDMAKIRKLSSNQVVKDFESGEILYVRDADDEGEDLVMLVGHAEDGCIQAVFLETTMVHWIELDLKARKPKAEILIYED
Serratia phage PS2	KJ025957.1	MITANAVKVFTPNTYVLNLDFFWGDGDICTGYRKILDPKCFDINQIREFLIEAKEVTEEQPEDLP

		EWFTDKWPHIYTLLKSDEDIWWTLERADIWYVDEVGTPWSLEKI
Vibrio phage 1.031.O._10	MG592415.1	IMATTGAKLLEICEAILTNPDYEGERLSDESAIQVTLGKEDVKSFSKLAQEAGLTTKTVGDDTVRVYVDADTADDSMKAINAIDRAQSYKLEARPVNLEGWHAEPGE
Aeromonas phage asfd_1h	MN871507.1	MEHVKYRFKEYSHISDFVHKDNVNLAIYRDLHDKEFYLKQVEVDKYVAVDAIGDRMYDENIFDFNKTEVDEFLEIVDELAAKPVEKPAEEPVEKPLTISQLHDWSIRHALTSLSVEKVVEMYKIYIK
Escherichia phage nomo	MN850578.1	VSDMAKIRKLSNNQVVKDFEPGEILYVRDADDEGEDLVMLLGHAAEDGCIQAVFLETAMVHWIELDMKARKPKAEILIYED
Escherichia phage PDX	MG963916.1	VSDMAKIRKLSSNQVVKDFEPGEILYVSYADDEGEDLVMLVGHAEDGCIQAVFLETAMVHWIELDLKARKPKAELIYED
Salmonella phage ISTP3	MT974436.1	IMSVKMKGVEFNAYYNDDEYWEKGAWHDDHCVKVNGEYREELDENIPDDADVVIESGTVYIPVAGESGAEEKDIQLVTHFKNWRKKKNFSFIVVIVKKDKASEVREALKNISGVMEVKGN
Pectobacterium phage DU_PP_I	MF979560.1	LMSKNTRYSHLLFHIMSPELREQFFTDDEENFDAGGDLFETLHPEKAEVLVSLLEPHLEYVIKELKFQRDHNILLGKGDELGAARLAICHRADKLDW
Halorubrum coriense virus Hardycor2	MN901520.1	MARSGRGRTVTNGRIEVVTEPEDSDFEFPPLFQEQSFHYQETDFYRTVDGQLFHYITLRNDEDFNWPPTVDVGINIIEVNDDTHHS
Cronobacter phage CR3	JQ691612.1	LMSKNTRYSHLLFHIMSPELREQFFTDDEENFDAGWDLFETLHPEKAEVLVSLLEPHLEYVIKELKFQRDHNILLGKGDELGAARLAICHRADKLDW
Vibrio phage vB_ValM-yong1	MN563793.1	MSELHQKAEEFVLCALADDTYGDSEQGNEVGLFLLQLILRHKKGQLRAAQMDELGVKTMPIVVSGDLGCQEITPL
Bacteriophage P27	AJ298298.1	LYSRTYKSAEEAMKKFENITVLHVDDFDYTNPELLPEVVKAIDVADIVIRGKRIVKNRLACTSGAMTETTSQQDDYEGICLEPDSFAVNVYHLLHATQVLHMSSNHETKTLGSEILNFACEYAKSAAEKELAQ
Klebsiella phage vB_KaeM_KaOmega	MN013077.1	LMGKYTRYSNLLVHVMSPELREEFFGDEEEDTDGGWDLFETLHPEKAEMLMEVLQPMLDDTIKELQFRRDYNTLLGQGKQTEATRLMICHRASKINWDED
Rhizobium phage RHEph04	JX483876.1	MAKFNKFRKGASTFVCECCGHHTRETGQALGAKICYACFELAGLENMLSDDGEEQFAKVGADEVKSWMNEIRKRSEAEFERAKASFSSLAPYFPSDEDFTTEAPLLSF
Escherichia phage ECML-4	JX128257.1	IMSVKMKGVEFNAYYNDDEYWVKNAWHDDHCVKVNGEYREELDENIPDDADVVIESGTVYIPVEGENGAEEKDISLVNHFKTWRKQKNFSFIVVTVKKDKLAEVREAMRCIPGVIEVKGN
Halorubrum coriense virus Serpecor1	MN901521.1	LSLRVLEEHKTMTELTNFTDPTPDLTDHERRLLRWVGADERLIEVCSFDVTSMERKRGDGKAVTRNSALVEVKKYTREWESLTEENAEDFDHYGGHFFSALWDGDLYEAYTRADYNNKAIMLEVFDVRRINSTRPAHAAEVTV
Salmonella phage GEC_vB_MG	MW006477.1	VKMTKDLWEVFQDDDEIKVIVSGSLEEGCGWRSYSDVCSEINTLQDAKLIAAAPELLDAVLDLKHKLYGNGPANPKIEELLNRLKGE

D.2. p56-heuristic matches from *Myoviridae*

All the matches that generated from applying heuristics-driven search for p56-type UngIn in the *Myoviridae* family are listed in Table D.2.

Table D.2. Matches generated by applying p56 parameters set on *Myoviridae* genomes.

Phage name	accession number	Heuristic match sequence
Acinetobacter phage BS46	QEP53246.1	VGKFFKNDWLYDIEVYPNICTFAFKEAETNRKLLFEISKRRNDLDDFLEFCRESVRNEYRWVGFNNLGFDYPVIHWILLKASSAKSSREKLKLTANQIYKYAMKVIDSKRDGQFGINIKSEDCLIIQLDLFKINHYDNKAKMTSLKLLEFNMRLDNISDLPFPVGKTLTSEEMDTLIEYNFSDVDATHIFYELNYDAIKFRSELTEKYGFDCTNLNDSNIGEQFFIRKIEAENPNAFYDYDIVSGKKIKKQTKRPFVRIGDCLFDYLKFQTKEMQALHAWFKKQVITETKGVFSDIEEHDLGELANYCELVTKEVKFKTEPTEQEKEEFLKAHPMGWFEVRELKAMETLKDENGNPVKESYVDDKGKTKERVVKVHKKAFYGCYRLADTLNVVLGGMRIDYGVGGLHGAVQGHHKKTADKRIKSWDVASMYPNIAIANRVYPEHLDETFCDSYEDFYNERKKFPKGTGENLAIKLGLNCVYGKSNSEFSCFYDTAYTMKITVNGQLTLSMLLERLIIDCNVKPLMANTDGFEVLIDNDQIEKADSIVTRWEKYVGLQMEAVEYSDMFIRDVNNYTAVYYVNGDIKQKGAYEYKPFLSKDLGMMHKNHSAIIVPMAVEHELMGKGSAEDFIRSHKDKFDFMLRAKVPRNSKLVLEIDGEDIEQQNICRYYVSEEGGYLTKIMPPLHEGGEDRRMSIESGIKVKTCNDINDFKWDINYSYYIEQADKLLEFFR
Morganella phage vB_MmoM_MP1	YP_009280040	MKKNLPEFIEFMNQDKTEPTAEQKRTHNHILDKNEVICVVEDYKNQLIILGCKEEDVESCFDILKNSEYFDKIRVKIDLLNVLPSNETTPYYNFKDSYVETSLIKRYPEQVQSVEDFITKNCTKENEVEPCYWEYEELFRKVLYGILQTEVDNIKEGFRDSGIRFYETWEDTDGKYGLVQYVIMYGGEIIGQLTQSGRWLGDTNVTIYSEPDFFTGLFMKYFITDTSWLAVRKLDNMINEFKTPEFHDKEYK
Providencia phage PSTRCR_127	QQK88256.1	MKKNLPEFIEFMNQDKTEPTAEQKRTHNHILDKNEVICVVEDNKNQLIIFGCKEEDVATCVDVLKNHEYYDKIRDKIDFLNILGPKAESVYHNFKDSYVETSLIKRYPEQVQSVEDFITKNCTKENEVESCYWEYEELFRKVLYGILQTEVDNIKDGFRDSGIKFYQTWEDTDDKYGLVQYVIMYDGKIIGQLSQSGRWLGDNNVTIYSEPDFFTGLFMKYFITDTSWLSVRTLDNMINEFKTPEFHDKEY
Providencia phage PSTRCR_121		MKKNLPEFIEFMNQDKTEPTAEQKRTHNHILDKNEVICVVEGNKNQLVIFGCKEEDVATCFDVLKNHEYYDKIRDKIDFLNVLPSNETSPYYNFKDSYVETSLIKRYPEQVQSVEDFITKNCTKENEVKSDNWEYEDLFRKVLYGILQTEVDNIKEGFRDSGIRFYETWEDTDGKYGLVQYAIMYDGKIIGQLSQSGRWLGDNNVTIYSEPDFFTGLFMKYFITDTSWLSVRTLDNMINEFKTPEFHDKEY

Synechococcus phage S-CAM1	YP_007672945.1	MSIYSLDLDTPPDLIKWIEESVSTIPKEEAQVHGIKDKSTQFTNKEVRSCITQVCNMNECWIPGFFDSYIRRFNQVYYNYDIAHLRDSVQYITYGVDDHYDWHIDETTYLKPQFKGDRNVVRKISFSFLCNDDYEGGDLEFWNKEDGSTYTVPKKRSRLIVFPSVTRHRVKPVTSGVRKSIVGWMVGPPWK
Synechococcus phage S-CAM1	AOV57570.1	MSIYSLDLDTPPDLIKWIEESVSTIPKEEAQVHGIKDKSTQFTNKEVRSCITQVCNMNECWIPGFFDSYIRRFNQVYYNYDIAHLRDSVQYITYGVDDHYDWHIDETTYLKPQFKGDRNVVRKISFSFLCNDDYEGGDLEFWNKEDGSTYTVPKKRSRLIVFPSVTRHRVKPVTSGVRRSIVGWMVGPPWK
Phage NCTB	SBV38431.1	MEKSSKKRKMPAIEDARSGGRKRAAKVATGSKKLNRSSPRFKPLDKVPNEKTAGVDFENYQWRKLVDPKFVYNKTRSKGFELTKNEKFGLRFHKQARGGFIVMQDGQYWPLPTSVYDDLVARSEVLPLNRWLKGKLSAEEIESFQARKIANQKQRDADARAEERERSRALKEELKRKERMKNLAEREAKREAAQKALNKIPDAPKVFDDVKQIGAIDALKEQLKEDSPIIKDIDDMGADDEDFDLQIDDDIDTLDNEVDDGDDEIMEDEVDVVDADQDAIDDALDAELDKAKDLARASQVIYEDDEDDNLGLDDLDEEDEDDEDYDDEDDDLTMEEENEEADIDAALDSADYDDDEEYEDDSEPEEPDEEEDEDSDSGEDVVDEDESDGEDESSDSEEETFEEGNVITFHKDESEKREWVILDIYPLKNNDAITVYKLYDVNADDGEYRTVRVKTGSKSTIEKMAKLVRKLNPKEFSKYFSVMDSYDKNPEPITS
Cyanophage P-RSM1	YP_007877718.1	MINPEVKGTLAKLLATENLTVEHRKVTTAYFDVQKRVLCLPIWKTASNTVYDLLVGHEVGHALYTPNTGLDGVNKGFVNVLEDVRIEKMMKDTYPGLRKSFFQGYKELWNDDFFGVNDEDISKLPFIDRINLFYKGNPEIEFTEEEQVYVDRAANTKTFEDVLKLAEDLFGRAEDIEDKKMDIDVPAAEPTPGVGDGEGEVTPQSSDSETENTDDGESEQQTASQPAPPVDGDRIGNPDAQISVTTGGNNSFGDEEYDETESITQEAFNQALETLIDDNAKEWVYLTLPKVDLEEIVIGHKEIQDDLHKHFITGERNMPSHYYYEDDAEKYAMYLEAQVSMMKTRYESYKKDAQKSVNYLVKQFEMKKSADDYKRQSTSRTGVIDTNSLYKYKLTDDIFKKITVVPDGKNHGLVMHIDWSGSMSHILLDTLKQTYNLIWFCRKAGIPFRVLAFQDSYSSSREENHGKEGDLNIHESFKLLEFFSSKQNKQSLDKSMFLVWSQAYSMNGCNVQAACKYGLGGTPLAEAVLCTRQIVDQMKKEENIQKVNVVCLTDGEANPMAFNEWYDPDYEYYKPYMKRSSLCHQSGKIFFLRDPKTGFTKKISSSPYETTKQIVGFHREITDYNWIGIRICSKSELGRAVRNNMDIVPADMDRKWKKEKFFSISKEAGFSESFFIPDKRLGDGTEDLQVSQKGEVATKAELQRAFKKHMGSKMGNKTILNKFIEQIA

Bibliography

1. De Bont, R. & Van Larebeke, N. Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis* **19**, 169–185 (2004).
2. Lindahl, T. Instability and decay of the primary structure of DNA. *nature* **362**, 709–715 (1993).
3. Myles, G. M. & Sancar, A. DNA repair. *Chem. Res. Toxicol.* **2**, 197–226 (1989).
4. KROKAN, H. E., STANDAL, R. & SLUPPHAUG, G. DNA glycosylases in the base excision repair of DNA. *Biochem. J.* **325**, 1–16 (1997).
5. Metz, A. H., Hollis, T. & Eichman, B. F. DNA damage recognition and repair by 3-methyladenine DNA glycosylase I (TAG). *EMBO J.* **26**, 2411–2420 (2007).
6. Krokan, H. E. & Bjørås, M. Base excision repair. *Cold Spring Harb. Perspect. Biol.* **5**, a012583 (2013).
7. Parsons, J. L. & Edmonds, M. J. The Base Excision Repair Pathway. in *Encyclopedia of Cell Biology* 442–450 (Elsevier, 2016). doi:10.1016/B978-0-12-394447-4.10046-X.
8. Schormann, N., Ricciardi, R. & Chattopadhyay, D. Uracil-DNA glycosylases—structural and functional perspectives on an essential family of DNA repair enzymes. *Protein Sci.* **23**, 1667–1685 (2014).
9. Robertson, A. B., Klungland, A., Rognes, T. & Leiros, I. DNA Repair in Mammalian Cells: Base excision repair: the long and short of it. *Cell. Mol. Life Sci.* **66**, 981–993 (2009).
10. Lindahl, T. An *N* -Glycosidase from *Escherichia coli* That Releases Free Uracil from DNA Containing Deaminated Cytosine Residues. *Proc. Natl. Acad. Sci.* **71**, 3649–3653 (1974).

11. Liu, P., Burdzy, A. & Sowers, L. C. Substrate recognition by a family of uracil-DNA glycosylases: UNG, MUG, and TDG. *Chem. Res. Toxicol.* **15**, 1001–1009 (2002).

12. Barrett, T. E. *et al.* Crystal structure of a G: T/U mismatch-specific DNA glycosylase: mismatch recognition by complementary-strand interactions. *Cell* **92**, 117–129 (1998).

13. Wibley, J. E., Waters, T. R., Haushalter, K., Verdine, G. L. & Pearl, L. H. Structure and specificity of the vertebrate anti-mutator uracil-DNA glycosylase SMUG1. *Mol. Cell* **11**, 1647–1659 (2003).

14. Schärer, O. D. & Jiricny, J. Recent progress in the biology, chemistry and structural biology of DNA glycosylases. *Bioessays* **23**, 270–281 (2001).

15. Chung, J. H. *et al.* A novel uracil-DNA glycosylase family related to the helix–hairpin–helix DNA glycosylase superfamily. *Nucleic Acids Res.* **31**, 2045–2055 (2003).

16. Lee, H.-W., Dominy, B. N. & Cao, W. New family of deamination repair enzymes in uracil-DNA glycosylase superfamily. *J. Biol. Chem.* **286**, 31282–31287 (2011).

17. Shi, H. *et al.* Biochemical characterization and mutational studies of a thermostable uracil DNA glycosylase from the hyperthermophilic euryarchaeon Thermococcus barophilus Ch5. *Int. J. Biol. Macromol.* **134**, 846–855 (2019).

18. Zhang, L. *et al.* Identification of a novel bifunctional uracil DNA glycosylase from Thermococcus barophilus Ch5. *Appl. Microbiol. Biotechnol.* **105**, 5449–5460 (2021).

19. Zharkov, D. O., Mechetin, G. V. & Nevinsky, G. A. Uracil-DNA glycosylase: Structural, thermodynamic and kinetic aspects of lesion search and recognition. *Mutat. Res. Mol. Mech. Mutagen.* **685**, 11–20 (2010).

20. Nilsen, H. *et al.* Nuclear and mitochondrial uracil-DNA glycosylases are generated by alternative splicing and transcription from different positions in the UNG gene. *Nucleic Acids Res.* **25**, 750–755 (1997).

21. Aravind, L. & Koonin, E. V. The α/β fold uracil DNA glycosylases: a common origin with diverse fates. *Genome Biol.* **1**, 1–8 (2000).

22. Savva, R. & Pearl, L. H. Nucleotide mimicry in the crystal structure of the uracil-DNA glycosylase–uracil glycosylase inhibitor protein complex. *Nat. Struct. Biol.* **2**, 752–757 (1995).

23. Savva, R., McAuley-Hecht, K., Brown, T. & Pearl, L. The structural basis of specific base-excision repair by uracil-DNA glycosylase. *Nature* **373**, 487–493 (1995).

24. Kavli, B. *et al.* Excision of cytosine and thymine from DNA by mutants of human uracil-DNA glycosylase. *EMBO J.* **15**, 3442–3447 (1996).

25. Kimber, S. T., Brown, T. & Fox, K. R. A mutant of uracil DNA glycosylase that distinguishes between cytosine and 5-methylcytosine. *PloS One* **9**, e95394 (2014).

26. Pearl, L. H. Structure and function in the uracil-DNA glycosylase superfamily. *Mutat. Res. Repair* **460**, 165–181 (2000).

27. Sudina, A. E., Volkov, E. M. & Kubareva, E. A. The repair enzyme uracil-DNA-glycosylase: study of the mechanism of functioning using modified analogues of DNA. *Biocatalysis* **41**, 121–123 (2000).

28. Kubareva, E. *et al.* Role of DNA definite structural elements in interaction with repair enzyme uracil-DNA glycosylase. *IUBMB Life* **46**, 597–606 (1998).

29. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).

30. Friedman, J. I., Majumdar, A. & Stivers, J. T. Nontarget DNA binding shapes the dynamic landscape for enzymatic recognition of DNA damage. *Nucleic Acids Res.* **37**, 3493–3500 (2009).

31. Savva, R. Targeting uracil-DNA glycosylases for therapeutic outcomes using insights from virus evolution. *Future Med. Chem.* **11**, 1323–1344 (2019).

32. Clauson, C. L., Oestreich, K. J., Austin, J. W. & Doetsch, P. W. Abasic sites and strand breaks in DNA cause transcriptional mutagenesis in *Escherichia coli*. *Proc. Natl. Acad. Sci.* **107**, 3657–3662 (2010).

33. Kunkel, T. A. Mutational specificity of depurination. *Proc. Natl. Acad. Sci.* **81**, 1494–1498 (1984).

34. Slupphaug, G. *et al.* A nucleotide-flipping mechanism from the structure of human uracil–DNA glycosylase bound to DNA. *Nature* **384**, 87–92 (1996).

35. Tye, B. K., Nyman, P. O., Lehman, I. R., Hochhauser, S. & Weiss, B. Transient accumulation of Okazaki fragments as a result of uracil incorporation into nascent DNA. *Proc. Natl. Acad. Sci.* **74**, 154–157 (1977).

36. Dianov, G. L. *et al.* Repair of uracil residues closely spaced on the opposite strands of plasmid DNA results in double-strand break and deletion formation. *Mol. Gen. Genet. MGG* **225**, 448–452 (1991).

37. Kiljunen, S. *et al.* Yersiniophage ϕR1-37 is a tailed bacteriophage having a 270 kb DNA genome with thymidine replaced by deoxyuridine. *Microbiology* **151**, 4093–4102 (2005).

38. Takahashi, I. & Marmur, J. Replacement of thymidylic acid by deoxyuridylic acid in the deoxyribonucleic acid of a transducing phage for Bacillus subtilis. *Nature* **197**, 794–795 (1963).

39. Chen, R., Le Rouzic, E., Kearney, J. A., Mansky, L. M. & Benichou, S. Vpr-mediated Incorporation of UNG2 into HIV-1 Particles Is Required to Modulate the Virus Mutation Rate and for Replication in Macrophages. *J. Biol. Chem.* **279**, 28419–28425 (2004).

40. Savva, R. The Essential Co-Option of Uracil-DNA Glycosylases by Herpesviruses Invites Novel Antiviral Design. *Microorganisms* **8**, 461 (2020).

41. Stavnezer, J., Guikema, J. E. J. & Schrader, C. E. Mechanism and regulation of class switch recombination. *Annu. Rev. Immunol.* **26**, 261–292 (2008).

42. Yousif, A. S., Stanlie, A., Mondal, S., Honjo, T. & Begum, N. A. Differential regulation of S-region hypermutation and class-switch recombination by noncanonical functions of uracil DNA glycosylase. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E1016-1024 (2014).

43. Dryden, D. T. F. & Tock, M. R. DNA mimicry by proteins. *Biochem. Soc. Trans.* **34**, 317 (2006).

44. Putnam, C. D. & Tainer, J. A. Protein mimicry of DNA and pathway regulation. *DNA Repair* **4**, 1410–1420 (2005).

45. Wang, H.-C., Ho, C.-H., Hsu, K.-C., Yang, J.-M. & Wang, A. H.-J. DNA mimic proteins: functions, structures, and bioinformatic analysis. *Biochemistry* **53**, 2865–2874 (2014).

46. Wang, H.-C., Chou, C.-C., Hsu, K.-C., Lee, C.-H. & Wang, A. H.-J. New paradigm of functional regulation by DNA mimic proteins: recent updates. *IUBMB Life* **71**, 539–548 (2019).

47. Parsons, L. M., Yeh, D. C. & Orban, J. Solution structure of the highly acidic protein HI1450 from Haemophilus influenzae, a putative double-stranded DNA mimic. *Proteins Struct. Funct. Bioinforma.* **54**, 375–383 (2004).

48. Parsons, L. M., Liu, F. & Orban, J. HU-α binds to the putative double-stranded DNA mimic HI1450 from Haemophilus influenzae. *Protein Sci.* **14**, 1684–1687 (2009).

49. Wang, H.-C., Wu, M.-L., Ko, T.-P. & Wang, A. H.-J. Neisseria conserved hypothetical protein DMP12 is a DNA mimic that binds to histone-like HU protein. *Nucleic Acids Res.* **41**, 5127–5138 (2013).

50. Liu, D. *et al.* Solution Structure of a TBP–TAFII230 Complex. *Cell* **94**, 573–583 (1998).

51. León, E. *et al.* A bacterial antirepressor with SH3 domain topology mimics operator DNA in sequestering the repressor DNA recognition helix. *Nucleic Acids Res.* **38**, 5226–5241 (2010).

52. Wang, H.-C. *et al.* Neisseria conserved protein DMP19 is a DNA mimic protein that prevents DNA binding to a hypothetical nitrogen-response transcription factor. *Nucleic Acids Res.* **40**, 5718–5730 (2012).

53. Wang, Z. *et al.* A Bacteriophage DNA Mimic Protein Employs a Non-specific Strategy to Inhibit the Bacterial RNA Polymerase. *Front. Microbiol.* **12**, 692512 (2021).

54. Walkinshaw, M. D. *et al.* Structure of Ocr from Bacteriophage T7, a Protein that Mimics B-Form DNA. *Mol. Cell* **9**, 187–194 (2002).

55. McMahon, S. A. *et al.* Extensive DNA mimicry by the ArdA anti-restriction protein and its role in the spread of antibiotic resistance. *Nucleic Acids Res.* **37**, 4887–4897 (2009).

56. Court, R., Cook, N., Saikrishnan, K. & Wigley, D. The Crystal Structure of λ-Gam Protein Suggests a Model for RecBCD Inhibition. *J. Mol. Biol.* **371**, 25–33 (2007).

57. Ramirez, B. E., Bax, A., Voloshin, O. N. & Camerini-otero, R. D. Solution structure of DinI provides insight into its mode of RecA inactivation. *Protein Sci.* **9**, 2161–2169 (2000).

58. Cole, A. R. *et al.* Architecturally diverse proteins converge on an analogous mechanism to inactivate Uracil-DNA glycosylase. *Nucleic Acids Res.* **41**, 8760–8775 (2013).

59. Serrano-Heras, G., Salas, M. & Bravo, A. A uracil-DNA glycosylase inhibitor encoded by a non-uracil containing viral DNA. *J. Biol. Chem.* **281**, 7068–7074 (2006).

60. Serrano-Heras, G., Bravo, A. & Salas, M. Phage φ29 protein p56 prevents viral DNA replication impairment caused by uracil excision activity of uracil-DNA glycosylase. *Proc. Natl. Acad. Sci.* **105**, 19044–19049 (2008).

61. Serrano-Heras, G. *et al.* Protein p56 from the Bacillus subtilis phage ϕ29 inhibits DNA-binding ability of uracil-DNA glycosylase. *Nucleic Acids Res.* **35**, 5393–5401 (2007).

62. Wang, H.-C. *et al.* Staphylococcus aureus protein SAUGI acts as a uracil-DNA glycosylase inhibitor. *Nucleic Acids Res.* **42**, 1354–1364 (2013).

63. Wu, Y. *et al.* The DDB1–DCAF1–Vpr–UNG2 crystal structure reveals how HIV-1 Vpr steers human UNG2 toward destruction. *Nat. Struct. Mol. Biol.* **23**, 933–940 (2016).

64. Bochkareva, E. *et al.* Single-stranded DNA mimicry in the p53 transactivation domain interaction with replication protein A. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15412–15417 (2005).

65. Hegde, S. S. *et al.* A Fluoroquinolone Resistance Protein from *Mycobacterium tuberculosis* That Mimics DNA. *Science* **308**, 1480–1483 (2005).

66. Ziach, K. *et al.* Single helically folded aromatic oligoamides that mimic the charge surface of double-stranded B-DNA. *Nat. Chem.* **10**, 511–518 (2018).

67. Corvaglia, V. *et al.* Carboxylate-functionalized foldamer inhibitors of HIV-1 integrase and Topoisomerase 1: artificial analogues of DNA mimic proteins. *Nucleic Acids Res.* **47**, 5511–5521 (2019).

68. Chowdhury, S. *et al.* Structure Reveals Mechanisms of Viral Suppressors that Intercept a CRISPR RNA-Guided Surveillance Complex. *Cell* **169**, 47-57.e11 (2017).

69. Guo, T. W. *et al.* Cryo-EM Structures Reveal Mechanism and Inhibition of DNA Targeting by a CRISPR-Cas Surveillance Complex. *Cell* **171**, 414-426.e12 (2017).

70. Dong, D. *et al.* Structural basis of CRISPR-SpyCas9 inhibition by an anti-CRISPR protein. *Nature* **546**, 436–439 (2017).

71. Shin, J. *et al.* Disabling Cas9 by an anti-CRISPR DNA mimic. *Sci. Adv.* **3**, e1701620 (2017).

72. Huang, M.-F. *et al.* The monomeric form of Neisseria DNA mimic protein DMP19 prevents DNA from binding to the histone-like HU protein. *PloS One* **12**, e0189461 (2017).

73. Wang, H.-C. *et al.* White spot syndrome virus protein ICP11: A histone-binding DNA mimic that disrupts nucleosome assembly. *Proc. Natl. Acad. Sci.* **105**, 20758–20763 (2008).

74. Ghosh, M., Meiss, G., Pingoud, A. M., London, R. E. & Pedersen, L. C. The nuclease a-inhibitor complex is characterized by a novel metal ion bridge. *J. Biol. Chem.* **282**, 5682–5690 (2007).

75. Tucker, A. T. *et al.* A DNA mimic: the structure and mechanism of action for the anti-repressor protein AbbA. *J. Mol. Biol.* **426**, 1911–1924 (2014).

76. Lee, C.-H., Shih, Y.-P., Ho, M.-R. & Wang, A. H.-J. The C-terminal D/E-rich domain of MBD3 is a putative Z-DNA mimic that competes for Zα DNA-binding activity. *Nucleic Acids Res.* **46**, 11806–11821 (2018).

77. Asensio, J. L. *et al.* Novel dimeric structure of phage ϕ29-encoded protein p56: insights into uracil-DNA glycosylase inhibition. *Nucleic Acids Res.* **39**, 9779–9788 (2011).

78. Lavysh, D. *et al.* The genome of AR9, a giant transducing Bacillus phage encoding two multisubunit RNA polymerases. *Virology* **495**, 185–196 (2016).

79. Huang, L.-H., Farnet, C. M., Ehrlich, K. C. & Ehrlich, M. Digestion of highly modified bacteriophage DNA by restriction endonucleases. *Nucleic Acids Res.* **10**, 1579–1591 (1982).

80. Wang, Z. & Mosbaugh, D. W. Uracil-DNA glycosylase inhibitor of bacteriophage PBS2: cloning and effects of expression of the inhibitor gene in Escherichia coli. *J. Bacteriol.* **170**, 1082–1091 (1988).

81. Putnam, C. D. *et al.* Protein mimicry of DNA from crystal structures of the uracil-DNA glycosylase inhibitor protein and its complex with Escherichia coli uracil-DNA glycosylase. *J. Mol. Biol.* **287**, 331–346 (1999).

82. Lundquist, A. J., Beger, R. D., Bennett, S. E., Bolton, P. H. & Mosbaugh, D. W. Site-directed mutagenesis and characterization of uracil-DNA glycosylase inhibitor protein: role of specific carboxylic amino acids in complex formation with Escherichia coli uracil-DNA glycosylase. *J. Biol. Chem.* **272**, 21408–21419 (1997).

83. Mol, C. D. *et al.* Crystal structure of human uracil-DNA glycosylase in complex with a protein inhibitor: protein mimicry of DNA. *Cell* **82**, 701–708 (1995).

84. Pérez-Lago, L. *et al.* Characterization of Bacillus subtilis uracil-DNA glycosylase and its inhibition by phage φ29 protein p56. *Mol. Microbiol.* **80**, 1657–1666 (2011).

85. Banos-Sanz, J. I. *et al.* Crystal structure and functional insights into uracil-DNA glycosylase inhibition by phage ϕ29 DNA mimic protein p56. *Nucleic Acids Res.* **41**, 6761–6773 (2013).

86. Mir-Sanchis, I. *et al.* Staphylococcal SCC mec elements encode an active MCM-like helicase and thus may be replicative. *Nat. Struct. Mol. Biol.* **23**, 891–898 (2016).

87. Wang, H.-C. *et al.* Using structural-based protein engineering to modulate the differential inhibition effects of SAUGI on human and HSV uracil DNA glycosylase. *Nucleic Acids Res.* **44**, 4440–4449 (2016).

88. Papp-Kádár, V., Balázs, Z., Vékey, K., Ozohanics, O. & Vértessy, B. G. Mass spectrometry-based analysis of macromolecular complexes of Staphylococcus aureus uracil-DNA glycosylase and its inhibitor reveals specific variations due to naturally occurring mutations. *FEBS Open Bio* **9**, 420–427 (2019).

89. Guenzel, C. A., Hérate, C. & Benichou, S. HIV-1 Vpr—a still "enigmatic multitasker". *Front. Microbiol.* **5**, 127 (2014).

90. WONG-STAAL, F., CHANDA, P. K. & GHRAYEB, J. Human immunodeficiency virus: the eighth gene. *AIDS Res. Hum. Retroviruses* **3**, 33–39 (1987).

91. Morellet, N., Bouaziz, S., Petitjean, P. & Roques, B. P. NMR structure of the HIV-1 regulatory protein VPR. *J. Mol. Biol.* **327**, 215–227 (2003).

92. Ahn, J. *et al.* HIV-1 Vpr loads uracil DNA glycosylase-2 onto DCAF1, a substrate recognition subunit of a cullin 4A-ring E3 ubiquitin ligase for proteasome-dependent degradation. *J. Biol. Chem.* **285**, 37333–37341 (2010).

93. Korn, A. M., Hillhouse, A. E., Sun, L. & Gill, J. J. Comparative genomics of three novel jumbo bacteriophages infecting Staphylococcus aureus. *J. Virol.* **95**, e02391-20 (2021).

94. Uchiyama, J. *et al.* Intragenus generalized transduction in Staphylococcus spp. by a novel giant phage. *ISME J.* **8**, 1949–1952 (2014).

95. Erickson, S. *et al.* Isolation and engineering of a Listeria grayi bacteriophage. *Sci. Rep.* **11**, 1–12 (2021).

96. Schilling, T., Hoppert, M. & Hertel, R. Genomic analysis of the recent viral isolate vB_BthP-Goe4 reveals increased diversity of Φ29-like phages. *Viruses* **10**, 624 (2018).

97. Kong, L. *et al.* Genome sequencing and characterization of three Bacillus cereus-specific phages, DK1, DK2, and DK3. *Arch. Virol.* **164**, 1927–1929 (2019).

98. Warner, H. R., Johnson, L. K. & Snustad, D. P. Early events after infection of Escherichia coli by bacteriophage T5. III. Inhibition of uracil-DNA glycosylase activity. *J. Virol.* **33**, 535–538 (1980).

99. Illergård, K., Ardell, D. H. & Elofsson, A. Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins Struct. Funct. Bioinforma.* **77**, 499–508 (2009).

100. Rihtman, B. *et al.* A new family of globally distributed lytic roseophages with unusual deoxythymidine to deoxyuridine substitution. *Curr. Biol.* (2021).

101. *Maduro, M. E. coli Codon Usage Analyzer 2.1. (2003). http://faculty.ucr.edu/~mmaduro/codonusage/usage.htm.*

102. Eyre-Walker, A. & Bulmer, M. Reduced synonymous substitution rate at the start of enterobacterial genes. *Nucleic Acids Res.* **21**, 4599–4603 (1993).

103. Vincze, T., Posfai, J. & Roberts, R. J. NEBcutter: a program to cleave DNA with restriction enzymes. *Nucleic Acids Res.* **31**, 3688–3691 (2003).

104. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415 (2003).

105. Henrici, R. C., Pecen, T. J., Johnston, J. L. & Tan, S. The pPSU plasmids for generating DNA molecular weight markers. *Sci. Rep.* **7**, 1–9 (2017).

106. Dagert, M. & Ehrlich, S. D. Prolonged incubation in calcium chloride improves the competence of Escherichia coli cells. *Gene* **6**, 23–28 (1979).

107. Rosano, G. L. & Ceccarelli, E. A. Recombinant protein expression in Escherichia coli: advances and challenges. *Front. Microbiol.* **5**, 172 (2014).

108. Tabor, S. Expression using the T7 RNA polymerase/promoter system. *Curr. Protoc. Mol. Biol.* **11**, 16.2. 1-16.2. 11 (1990).

109. Dubendorf, J. W. & Studier, F. W. Controlling basal expression in an inducible T7 expression system by blocking the target T7 promoter with lac repressor. *J. Mol. Biol.* **219**, 45–59 (1991).

110. Gasteiger, E. *et al.* Protein identification and analysis tools on the ExPASy server. *Proteomics Protoc. Handb.* 571–607 (2005).

111. Savva, R. & Pearl, L. H. Cloning and expression of the uracil-DNA glycosylase inhibitor (UGI) from bacteriophage PBS-1 and crystallization of a uracil-DNA glycosylase-UGI complex. *Proteins Struct. Funct. Genet.* **22**, 287–289 (1995).

112. Rupp, B. *Biomolecular crystallography: principles, practice, and application to structural biology*. (Garland Science, 2009).

113. Rhodes, G. *Crystallography made crystal clear: a guide for users of macromolecular models*. (Elsevier, 2010).

114. Bijelic, A. & Rompel, A. Polyoxometalates: more than a phasing tool in protein crystallography. *ChemTexts* **4**, 1–27 (2018).

115. Vagin, A. & Teplyakov, A. Molecular replacement with MOLREP. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 22–25 (2010).

116. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).

117. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 235–242 (2011).

118. Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 355–367 (2011).

119. Broennimann, C. *et al.* The PILATUS 1M detector. *J. Synchrotron Radiat.* **13**, 120–130 (2006).

120. Winter, G. xia2: an expert system for macromolecular crystallography data reduction. *J. Appl. Crystallogr.* **43**, 186–190 (2010).

121. Kabsch, W. Xds. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010).

122. Clabbers, M. T. B., Gruene, T., Parkhurst, J. M., Abrahams, J. P. & Waterman, D. G. Electron diffraction data processing with *DIALS*. *Acta Crystallogr. Sect. Struct. Biol.* **74**, 506–518 (2018).

123. Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr. D Biol. Crystallogr.* **69**, 1204–1214 (2013).

124. Madeira, F. *et al.* The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* **47**, W636–W641 (2019).

125. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

126. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547 (2018).

127. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

128. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).

129. Del Sole, A. Introducing Visual Studio Code. in *Visual Studio Code Distilled* 1–15 (Apress, 2021). doi:10.1007/978-1-4842-6901-5_1.

130. Cock, P. J. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

131. Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **49**, D10 (2021).

132. Ohno, H., Sakai, H., Washio, T. & Tomita, M. Preferential usage of some minor codons in bacteria. *Gene* **276**, 107–115 (2001).

133. Hecht, A. *et al.* Measurements of translation initiation from all 64 codons in E. coli. *Nucleic Acids Res.* **45**, 3615–3626 (2017).

134. Holm, L. Benchmarking fold detection by DaliLite v. 5. *Bioinformatics* **35**, 5326–5327 (2019).

135. Krissinel, E. & Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2256–2268 (2004).

136. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

137. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

138. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv* (2021).

139. Mirdita, M. *et al.* ColabFold-Making protein folding accessible to all. (2021).

140. Kunkel, T. A. Rapid and efficient site-specific mutagenesis without phenotypic selection. *Proc. Natl. Acad. Sci.* **82**, 488–492 (1985).

141. Savva, R. & Pearl, L. H. Crystallization and preliminary X-ray analysis of the uracil-DNA glycosylase DNA repair enzyme from herpes simplex virus type 1. *J. Mol. Biol.* **234**, 910–912 (1993).

142. Savva, R. & Pearl, L. H. Cloning and expression of the uracil–DNA glycosylase inhibitor (UGI) from bacteriophage PBS-1 and crystallization of a uracil–DNA glycosylase–UGI complex. *Proteins Struct. Funct. Bioinforma.* **22**, 287–289 (1995).

143. Henaut, A. & Danchin, A. Analysis and predictions from Escherichia coli sequences, or E. coli in silico. *Escherichia Coli Salmonella Cell. Mol. Biol. ASM Press Wash. DC* 2047–2066 (1996).

144. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

145. Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **49**, D344–D354 (2021).

146. Robert, X. & Gouet, P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **42**, W320–W324 (2014).

147. Stanton, C. R., Rice, D. T., Beer, M., Batinovic, S. & Petrovski, S. Isolation and Characterisation of the Bundooravirus Genus and Phylogenetic Investigation of the Salasmaviridae Bacteriophages. *Viruses* **13**, 1557 (2021).

148. Li, C. *et al.* Isolation and characterization of Bacillus cereus phage vB_BceP-DLc1 reveals the largest member of the Φ29-Like phages. *Microorganisms* **8**, 1750 (2020).

149. Sleigh, R., Sharkey, M., Newman, M. A., Hahn, B. & Stevenson, M. Differential Association of Uracil DNA Glycosylase with SIVSMVpr and Vpx Proteins. *Virology* **245**, 338–343 (1998).

150. Wu, Y. *et al.* Structural basis of clade-specific engagement of SAMHD1 (Sterile α Motif and Histidine/Aspartate-containing Protein 1) restriction factors by lentiviral viral protein X (Vpx) virulence factors. *J. Biol. Chem.* **290**, 17935–17945 (2015).

151. Selig, L. *et al.* Uracil DNA glycosylase specifically interacts with Vpr of both human immunodeficiency virus type 1 and simian immunodeficiency virus of sooty mangabeys, but binding does not correlate with cell cycle arrest. *J. Virol.* **71**, 4842–4846 (1997).

152. Newman, J. *et al.* Towards rationalization of crystallization screening for small-to medium-sized academic laboratories: the PACT/JCSG+ strategy. *Acta Crystallogr. D Biol. Crystallogr.* **61**, 1426–1431 (2005).

153. Vieira, G., de Lencastre, H. & Archer, L. Restriction analysis of PBS1-related phages. *Arch. Virol.* **106**, 121–126 (1989).

154. Berkner, K. L. & Folk, W. R. The effects of substituted pyrimidines in DNAs on cleavage by sequence-specific endonucleases. *J. Biol. Chem.* **254**, 2551–2560 (1979).

155. Lee, Y., Son, B., Cha, Y. & Ryu, S. Characterization and Genomic Analysis of PALS2, a Novel Staphylococcus Jumbo Bacteriophage. *Front. Microbiol.* **12**, 395 (2021).

156. Everett, R. D. DNA replication of bacteriophage T5. 3. Studies on the structure of concatemeric T5 DNA. *J. Gen. Virol.* **52**, 25–38 (1981).

157. Warner, H. R., Thompson, R. B., Mozer, T. J. & Duncan, B. K. The properties of a bacteriophage T5 mutant unable to induce deoxyuridine 5'-triphosphate nucleotidohydrolase. Synthesis of uracil-containing T5 deoxyribonucleic acid. *J. Biol. Chem.* **254**, 7534–7539 (1979).

158. Mahata, T. *et al.* A phage mechanism for selective nicking of dUMP-containing DNA. *Proc. Natl. Acad. Sci.* **118**, e2026354118 (2021).

159. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv* (2021).

160. Acharya, N., Roy, S. & Varshney, U. Mutational Analysis of the Uracil DNA Glycosylase Inhibitor Protein and Its Interaction with Escherichia coli Uracil DNA Glycosylase. *J. Mol. Biol.* **321**, 579–590 (2002).

161. Sokolova, M. *et al.* A non-canonical multisubunit RNA polymerase encoded by the AR9 phage recognizes the template strand of its uracil-containing promoters. *Nucleic Acids Res.* **45**, 5958–5967 (2017).

162. Lin, Z. *et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction*. http://biorxiv.org/lookup/doi/10.1101/2022.07.20.500902 (2022) doi:10.1101/2022.07.20.500902.

163. Wu, R. *et al. High-resolution de novo structure prediction from primary sequence*. http://biorxiv.org/lookup/doi/10.1101/2022.07.21.500999 (2022) doi:10.1101/2022.07.21.500999.

164. Sokolova, M. L., Misovetc, I. & V. Severinov, K. Multisubunit RNA Polymerases of Jumbo Bacteriophages. *Viruses* **12**, 1064 (2020).

165. Lavysh, D., Sokolova, M., Slashcheva, M., Förstner, K. U. & Severinov, K. Transcription Profiling of Bacillus subtilis Cells Infected with AR9, a Giant Phage Encoding Two Multisubunit RNA Polymerases. *mBio* **8**, e02041-16 (2017).

166. Leskinen, K., Blasdel, B., Lavigne, R. & Skurnik, M. RNA-Sequencing Reveals the Progression of Phage-Host Interactions between φR1-37 and Yersinia enterocolitica. *Viruses* **8**, 111 (2016).

167. Lee, H.-W., Brice, A. R., Wright, C. B., Dominy, B. N. & Cao, W. Identification of Escherichia coli Mismatch-specific Uracil DNA Glycosylase as a Robust Xanthine DNA Glycosylase. *J. Biol. Chem.* **285**, 41483–41490 (2010).

168. Liao, Y.-T. *et al.* Structural insight into the differential interactions between the DNA mimic protein SAUGI and two gamma herpesvirus uracil-DNA glycosylases. *Int. J. Biol. Macromol.* **160**, 903–914 (2020).

169. Handa, P., Roy, S. & Varshney, U. The Role of Leucine 191 of Escherichia coliUracil DNA Glycosylase in the Formation of a Highly Stable Complex with the Substrate Mimic, Ugi, and in Uracil Excision from the Synthetic Substrates. *J. Biol. Chem.* **276**, 17324–17331 (2001).

170. Koulis, A., Cowan, D. A., Pearl, L. H. & Savva, R. Uracil-DNA glycosylase activities in hyperthermophilic micro-organisms. *FEMS Microbiol. Lett.* **143**, 267–271 (1996).

171. Neubort, S. & Marmur, J. Synthesis of the Unusual DNA of Bacillus subtilis Bacteriophage SP-15. *J. Virol.* **12**, 1078–1084 (1973).